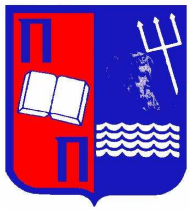
ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ



ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΜΗ ΠΑΡΑΜΕΤΡΙΚΗ ΣΥΜΠΕΡΑΣΜΑΤΟΛΟΓΙΑ ΓΙΑ ΣΤΑΘΜΙΣΜΕΝΕΣ ΚΑΤΑΝΟΜΕΣ

Ιωάννης Γ. Μπαντούνας

Διπλωματική Εργασία

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

Πειραιάς
Ιανουάριος 2012

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

Τα μέλη της Επιτροπής ήταν:

- Ηλιόπουλος Γεώργιος (Επιβλέπων)
- Κούτρας Μάρκος
- Στέγγος Δημήτριος

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

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

NONPARAMETRIC INFERENCE FOR WEIGHTED DISTRIBUTIONS

by
Ioannis G. Badounas

MSc Dissertation

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

Piraeus, Greece
January 2012

Στους γονείς μου,
Γεώργιο και Μαρία

Ευχαριστίες

Κατ' αρχάς θα ήθελα να ευχαριστήσω ιδιαίτερα τον επιβλέποντα της παρούσας διπλωματικής εργασίας Αναπληρωτή Καθηγητή Γεώργιο Ηλιόπουλο για την αμέριστη συμπαράστασή του και την πολύτιμη βοήθεια που μου προσέφερε κατά τη διάρκεια εκπόνησης της εργασίας. Επίσης θα ήθελα να ευχαριστήσω και τα άλλα δύο μέλη της Τριμελούς Εξεταστικής Επιτροπής, Καθηγητή Μάρκο Κούτρα και Επίκουρο Καθηγητή Δημήτριο Στέγγο για την επίβλεψή τους. Ακόμη θα ήθελα να εκφράσω τις ευχαριστίες μου προς όλους τους διδάσκοντες στο Π.Μ.Σ. στην Εφαρμοσμένη Στατιστική του Πανεπιστημίου Πειραιώς, αλλά και στους συμφοιτητές μου για τη εποικοδομητική συνεργασία μας κατά τη διάρκεια των μεταπτυχιακών μου σπουδών.

Τέλος, θα ήθελα να ευχαριστήσω ιδιαίτερα τον κύριο Γεώργιο Πιτσέλη, για την οικονομική βοήθεια μέσω της απασχόλησής μου στο πρόγραμμα εξετάσεων Διαμεσολαβούντων και ειδικών κατηγοριών κατά τη διάρκεια των μεταπτυχιακών μου σπουδών, συμβάλλοντας έτσι καθοριστικά στην επιτυχή ολοκλήρωσή τους.

Περίληψη

Η παρούσα διπλωματική εργασία αποτελεί μία ανασκόπηση της βιβλιογραφίας για τη μη παραμετρική και ημιπαραμετρική εκτίμηση μίας άγνωστης συνάρτησης κατανομής βάσει δεδομένων τα οποία προέρχονται από μεροληπτικές εκδοχές της. Στο πρώτο μέρος μελετάται η μη παραμετρική εκτίμηση μιας συνάρτησης κατανομής όταν τα δεδομένα προέρχονται από σταθμισμένες εκδοχές της με γνωστές αντίστοιχες συναρτήσεις στάθμισης και παρουσιάζονται αλγόριθμοι για τον υπολογισμό του μη παραμετρικού εκτιμητή μεγίστης πιθανοφάνειας της εν λόγω κατανομής. Στο δεύτερο μέρος θεωρείται η περίπτωση που οι κατανομές από τις οποίες προέρχονται τα δεδομένα ικανοποιούν το λεγόμενο μοντέλου λόγου πυκνοτήτων και εξετάζεται η ημιπαραμετρική εκτίμηση των αντίστοιχων συναρτήσεων κατανομής και παρουσιάζονται έλεγχοι υποθέσεων σχετικά με την ισότητα αυτών των κατανομών.

Abstract

The present thesis consists of a review of the literature on nonparametric and semi-parameteric estimation of an unknown distribution function based on several biased samples. In the first part the nonparametric estimation of a distribution function is studied when the data arise from weighted versions of it with known weighting functions and algorithms for the calculation of its nonparametric maximum likelihood estimators are presented. In the second part the case where the underlying distributions satisfy the so-called density ratio model is considered and the semi-parametric estimation of the corresponding distribution functions is investigated as well as tests of hypotheses about the equality of these distributions are presented.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγή

1.1 Η έννοια της σταθμισμένης κατανομής

Σε πολλά πρακτικά προβλήματα, για διάφορους λόγους, οι πιθανότητες των ατόμων (αντικειμένων) που λαμβάνονται ως δείγμα από τον πληθυσμό, μπορεί να διαφέρουν από άτομο σε άτομο και να εξαρτώνται από το προς μελέτη φαινόμενο. Αυτό έχει ως αποτέλεσμα να προκαλείται αυτό που αναφέρεται ως **μεροληψία** (bias) του δείγματος. Αν δεν λάβει κάποιος υπ' όψιν του αυτή τη δειγματοληπτική μεροληψία μπορεί να προκληθούν σοβαρά προβλήματα και να οδηγηθεί σε παραπλανητικά συμπεράσματα.

Έστω λοιπόν ότι μας ενδιαφέρει η μελέτη της κατανομής ενός χαρακτηριστικού Y από τον πληθυσμό. Για να γίνει συμπερασματολογία, λαμβάνουμε ένα δείγμα $\left(Y_1, Y_2, \ldots, Y_n\right)$ από τον πληθυσμό, τέτοιο ώστε τα Y_i να είναι ανεξάρτητα και ισόνομα κατανεμημένα. Αν κάτω από τον μηχανισμό της δειγματοληψίας, κάθε άτομο έχει πιθανότητα να επιλεγεί ανεξάρτητη από την αντίστοιχη τιμή Y, τότε η κατανομή των Y_i είναι ίδια με αυτή της Y. Σε αντίθετη περίπτωση όπου η πιθανότητα εξαρτάται από την τιμή Y, η κατανομή των Y_i θα διαφέρει από της Y. Για να γίνει διάκριση μεταξύ της κατανομής του Y και της κατανομής του Y_i που παρατηρούμε, η κατανομή Y θα αναφέρεται ως "βασική κατανομή".

Έστω $f(y)$ η συνάρτηση πυκνότητας πιθανότητας της βασικής κατανομής. Υπο-

θέτουμε ότι ένα άτομο με χαρακτηριστικό $Y = y$ παρατηρείται με πιθανότητα ανάλογη του $w(y) \geq 0$. Τότε τα παρατηρούμενα Y_i έχουν κατανομή με συνάρτηση κατανομής που δίνεται από τον τύπο

$$f^w(y) = \frac{\int_{-\infty}^{x} w(y)\, dF(y)}{\mu^w},$$

όπου $\mu^w = \mathsf{E}[w(Y)] = \int w(y)\, dF(y)$.
Η κατανομή των Y_i θα αναφέρεται ως **σταθμισμένη κατανομή** (weighted distribution). Η συνάρτηση $w(y)$ θα αναφέρεται ως η **συνάρτηση στάθμισης** ή **συνάρτηση βάρους** (weight function). Ανάλογα με το πρόβλημα, η συνάρτηση βάρους παίρνει και διαφορετική μορφή. Παρακάτω παρουσιάζονται οι πιο συνηθισμένες (δες Liang, 2005):

- $w(y) = y^\alpha$, $\alpha > 0$
- $w(y) = y(y-1)\cdots(y-r)$ όπου r ακέραιος
- $w(y) = e^{\alpha+\beta y}$
- $w(y) = \alpha + \beta y$
- $w(y) = 1 - (1-\beta)^y, \quad 0 < \beta < 1$
- $w(y) = \dfrac{\alpha + \beta y}{\gamma + \delta y}$
- $w(y) = G(y) = \mathsf{P}(Z \leq y)$ για κάποια τ.μ. Z
- $w(y) = r(y)$ όπου r είναι η συνάρτηση επιβίωσης της βασικής κατανομής.

Αν $r(y) = 1$ όταν $y \in T$ και $r(y) = 0$ όταν $y \notin T$, τότε δημιουργούνται οι περικεκομμένες κατανομές με τιμές στο σύνολο T.

Συχνά, οι παράμετροι που συμπεριλαμβάνονται σε κάθε συνάρτηση βάρους δεν εξαρτώνται από την βασική κατανομή και είναι συνήθως άγνωστες. Στην ειδική περίπτωση που χρησιμοποιηθεί η συνάρτηση βάρους $w(y) = y^\alpha$ με $\alpha = 1$, τότε η κατανομή ονομάζεται **μεροληπτική λόγω μεγέθους** (length biased distribution).

Ακολούθως παρουσιάζονται κάποια παραδείγματα και εφαρμογές σταθμισμένων κατανομών που βρίσκονται διάσπαρτα στην βιβλιογραφία.

Παράδειγμα 1. Ας υποθέσουμε ότι μας ενδιαφέρει η μελέτη της κατανομής του αριθμού των παιδιών που έχουν κάποια σπάνια ασθένεια σε οικογένειες που έχουν την τάση να γεννούν τέτοια παιδιά. Δεν είναι τότε εύκολο να βρεθούν τέτοιες οικογένειες με χρήση της απλής τυχαίας δειγματοληψίας. Μία πιο βολική δειγματοληπτική μέθοδος είναι να βρεθεί αρχικά ένα παιδί με αυτή την ασθένεια (από επισκέψεις σε νοσοκομεία για παράδειγμα) και κατόπιν να μετρηθεί ο αριθμός των αδερφών του με την ίδια ασθένεια. Ας θεωρήσουμε ότι η πιθανότητα να είναι θετικό το τεστ ως προς την ασθένεια για ένα παιδί είναι ίση με β. Έστω επίσης ότι κάθε παιδί παρουσιάζει την ασθένεια ανεξάρτητα από κάθε άλλο παιδί. Τότε μία οικογένεια με y παιδιά έχει πιθανότητα να έχει τουλάχιστον ένα ασθενές παιδί ίση με $w(y) = 1 - (1 - \beta)^y$. Άρα ο αριθμός των αδερφών που έχουν την ασθένεια δοθέντος του αριθμού των αδερφών, ακολουθεί μία σταθμισμένη διωνυμική κατανομή με συνάρτηση βάρους $w(y)$.

Παράδειγμα 2. Για την μελέτη της πυκνότητας του πληθυσμού των αγρίων ζώων, μία μέθοδος που χρησιμοποιείται ευρέως είναι αυτή που ονομάζεται **τετραγωνική δειγματοληψία** (quadrat sampling, δες Buckland, Anderson, Burnham, and Laake, 1993). Όταν ένας επιστήμονας θέλει να μάθει πόσα ζώα υπάρχουν σε έναν βιότοπο, δεν είναι εφικτό να τα μετρήσει όλα! Αντ' αυτού είναι αναγκασμένος να μετρήσει ένα μικρότερο υποσύνολο, αντιπροσωπευτικό του πληθυσμού. Η δειγματοληψία φυτών αλλά και ζώων που δεν κινούνται πολύ (όπως τα σαλιγκάρια) μπορεί να γίνει χρησιμοποιώντας ένα **δειγματοληπτικό τετράγωνο** (quadrat). Το μέγεθος αυτού του τετραγώνου εξαρτάται από το μέγεθος των οργανισμών που μελετούνται. Για παράδειγμα, για να μετρηθούν τα φυτά που μεγαλώνουν σε ένα σχολικό χώρο, θα μπορούσε κανείς να χρησιμοποιήσει ένα τετράγωνο με πλευρά 0.5 ή 1 μέτρο. Είναι σημαντικό επίσης η δειγματοληψία σε μία περιοχή να γίνεται τυχαία ούτως ώστε να αποφευχθεί η μεροληψία. Ένας τρόπος για να πάρει κάποιος καλό δείγμα είναι να τοποθετήσει τα τετράγωνα στις συντεταγμένες ενός αριθμημένου πλέγματος.

Έτσι λοιπόν η δειγματοληψία ξεκινά επιλέγοντας τυχαία τον αριθμό των τετραγώνων σταθερού εμβαδού. Ακολούθως χρησιμοποιώντας συχνά εναέριες μέθοδους, προκύπτει ο αριθμός των αντικειμένων που μας ενδιαφέρει (ζώα, φυτά κ.τ.λ.) και τα οποία εμφανίζονται σε ομάδες. Αν κάθε αντικείμενο έχει πιθανότητα β να παρατηρηθεί, τότε μία ομάδα από y ανεξάρτητα αντικείμενα θα παρατηρηθεί με πιθανότητα $w(y) = 1 - (1-\beta)^y$. Ας υποθέσουμε επίσης ότι η πραγματική κατανομή του αριθμού των αντικειμένων σε μία ομάδα έχει συνάρτηση πυκνότητας $f(y)$. Τότε ο παρατηρούμενος αριθμός των αντικειμένων στην ομάδα ακολουθεί την σταθμισμένη κατανομή με συνάρτηση πυκνότητας $w(y)f(y)/\int w(y)f(y)dy$.

Παράδειγμα 3. Ένα άλλο σχέδιο δειγματοληψίας, είναι το λεγόμενο **line-transect sampling** (δες Safranyik and Linton, 2002). Αυτό έχει χρησιμοποιηθεί κυρίως για την εκτίμηση του αριθμού των φυτών ή ζώων ενός συγκεκριμένου είδους σε μία περιοχή. Η μέθοδος βασίζεται στην κατάρτιση μίας βάσης σε όλη την περιοχή που πρέπει να ερευνηθεί και εν συνεχεία να χαραχθεί μία **γραμμή** (line) σε ένα τυχαία επιλεγμένο σημείο της βάσης. Ο ερευνητής περπατά κατά μήκος αυτής της γραμμής, καταγράφει τα αντικείμενα που τον ενδιαφέρουν και τα αντίστοιχα χαρακτηριστικά αυτών. Με την διαδικασία αυτή δημιουργεί το δείγμα. Να παρατηρήσουμε ότι συνήθως τα μεμονωμένα αντικείμενα χωρίζονται σε ομάδες οι οποίες χρησιμοποιούνται ως δειγματοληπτικές μονάδες. Έτσι λοιπόν οι εκτιμήσεις που προκύπτουν για κάποιο χαρακτηριστικό των ομάδων μπορούν να προσαρμοστούν και να ερμηνεύσουν το αντίστοιχο χαρακτηριστικο του πληθυσμού. Είναι όμως προφανές ότι όσο πιο κοντά στον παρατηρητή (δηλαδή στην γραμμή) είναι η ομάδα και όσο μεγαλύτερο μέγεθος έχει, τόσο πιο πιθανό είναι να παρατηρηθεί. Αντίθετα, όσο πιο μακριά είναι από τον παρατηρητή και όσο μικρότερο μέγεθος έχει, τόσο μειώνεται και η πιθανότητα να παρατηρηθεί. Με άλλα λόγια, η πιθανότητα να παρατηρηθεί μία ομάδα είναι ανάλογη του μεγέθους της. Έτσι οδηγούμαστε σε μία μεροληπτική λόγω μεγέθους κατανομή.

Παράδειγμα 4. Ας υποθέσουμε ότι εκτελούμε ένα πείραμα όπου βασικός στόχος είναι να εκτιμηθεί ο χρόνος που απαιτείται μέχρι να συμβεί ένα γεγονός. Απαραίτητη

προϋπόθεση για την διεξαγωγή του πειράματος είναι η συνεχής παροχή ηλεκτρικής ενέργειας. Αν η ηλεκτρική ενέργεια διακοπεί πριν παρατηρηθεί το γεγονός, τότε το πείραμα θα πρέπει να σταματήσει. Κάποια παραδείγματα είναι τα ακόλουθα:

- Μας ενδιαφέρει να εκτιμήσουμε τον χρόνο που απαιτείται ώστε να τερματίσει κάποιος ένα ηλεκτρονικό παιχνίδι. Είναι σαφές ότι αν διακοπεί το ρεύμα πριν τερματίσει το παιχνίδι, τότε δεν μπορούμε να ξέρουμε ποιός θα είναι ο χρόνος τερματισμού. Το ερώτημα επομένως είναι, ποιά είναι η κατανομή που προκύπτει **μόνο** από τις παρατηρήσεις που έχουν καταγραφεί?
- Μας ενδιαφέρει ο μέσος χρόνος που απαιτείται ώστε n αβγά που τοποθετούνται σε μία εκκολαπτική μηχανή να ανοίξουν και να προκύψει το αντίστοιχο είδος ζώου. Όμως η εκκόλαψη των αβγών απαιτεί άριστες συνθήκες θερμοκρασίας και υγρασίας οι οποίες ρυθμίζονται από την μηχανή. Σε περίπτωση που διακοπεί το ρεύμα οι συνθήκες αλλάζουν.

Έστω $f(x)$ η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής X που δηλώνει τον χρόνο που απαιτείται ώσπου να συμβεί ένα γεγονός. Έστω επίσης $g(t)$ η συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής T που δηλώνει τον χρόνο που θα σταματήσει η παροχή ηλεκτρικής ενέργειας. Θεωρούμε ότι οι τ.μ. X και T είναι ανεξάρτητες. Μία παρατήρηση θα συμπεριληφθεί στο δείγμα αν και μόνον αν για το ζευγάρι (x,t) ισχύει $x \leq t$. Η από κοινού κατανομή των τυχαίων μεταβλητών (X,T) δοθέντος $X \leq T$ έχει πυκνότητα

$$\frac{f(x)g(t)}{\mathsf{P}[X \leq T]}, \quad x \leq t.$$

Τότε η καταγεγραμμένη (recorded) τιμή της X θα έχει συνάρτηση πυκνότητας πιθανότητας

$$f^{(r)}(x) = \int_x^\infty \frac{f(x)g(t)}{\mathsf{P}[X \leq T]} dt = \frac{f(x)[1-G(x)]}{\mathsf{P}[X \leq T]},$$

όπου η $G(t)$ είναι η συνάρτηση κατανομής της T. Επίσης ισχύει

$$\mathsf{P}[X \leq T] = \int_0^\infty f(x)\big[1-G(x)\big]dx.$$

Είναι προφανές ότι η $f^{(r)}$ είναι μία σταθμισμένη εκδοχή της f.

Παρ' όλ' αυτά υπάρχουν περιπτώσεις όπου είναι σημαντικό να καταγραφεί ο χρόνος ακόμα και όταν $X > T$ (π.χ. ανάλυση επιβίωσης). Συγκεκριμένα, η τυχαία μεταβλητή $Z = \min(X, T)$ παρατηρείται σε κάθε πείραμα. Για την συνάρτηση κατανομής της Z ισχύει

$$\begin{aligned} H_Z(z) &= \mathsf{P}[Z \leq z] \\ &= 1 - \mathsf{P}[Z > z] \\ &= 1 - \mathsf{P}[\min(X, T) > z] \\ &= 1 - \mathsf{P}[X > z, T > z] \\ &= 1 - \mathsf{P}[X > z]\mathsf{P}[T > z] \\ &= 1 - [1 - F(z)][1 - G(z)]. \end{aligned}$$

Άρα λοιπόν η συνάρτηση πυκνότητας πιθανότητας της Z είναι

$$\begin{aligned} h_Z(z) &= \frac{dH_Z(z)}{dz} \\ &= \frac{d}{dz}\{1 - [1 - F(z)][1 - G(z)]\} \\ &= \frac{d}{dz}\{1 - [1 - G(z) - F(z) + F(z)G(z)]\} \\ &= [1 - F(z)]g(z) + [1 - G(z)]f(z). \end{aligned}$$

Έστω $\overline{F}(z)$ και $\overline{G}(z)$ οι συναρτήσεις επιβίωσης των X και T αντίστοιχα. Τότε παρατηρούμε ότι η τελευταία σχέση μπορεί να γραφεί και ως εξής:

$$\begin{aligned} h_Z(z) &= \overline{F}(z)g(z) + \overline{G}(z)f(z) \\ &= \int_0^\infty \overline{F}(z)g(z)dz \frac{\overline{F}(z)g(z)}{\int_0^\infty \overline{F}(z)g(z)dz} + \int_0^\infty \overline{G}(z)f(z)dz \frac{\overline{G}(z)f(z)}{\int_0^\infty \overline{G}(z)f(z)dz}, \end{aligned}$$

το οποίο σημαίνει ότι η συνάρτηση πυκνότητας πιθανότητας της Z μπορεί να γραφεί ως μείξη δύο σταθμισμένων κατανομών.

Παράδειγμα 5. Πηγαίνουμε μία τυχαία μέρα σε ένα εστιατόριο που τρώνε φοιτητές και κάνουμε σε όλους την ερώτηση "Πόσα άτομα (μαζί με εσένα) κάθονται στο τραπέζι

σου?" Έστω ότι τα αποτελέσματα της έρευνας είναι τα ακόλουθα (δες Arratia and Goldstein, 2009):

20% Είπαν ότι έφαγαν μόνοι τους

30% Είπαν ότι έφαγαν με ακόμα ένα άτομο

30% Είπαν ότι έφαγαν με ακόμα δύο άτομα και

20% Είπαν ότι έφαγαν με ακόμα τρία άτομα

Από τα δεδομένα αυτά μπορεί, λανθασμένα, να εξάγουμε το συμπέρασμα ότι

20% των τραπεζιών είχαν μόνο ένα άτομο

30% των τραπεζιών είχαν δύο άτομα

30% των τραπεζιών είχαν τρία άτομα και

20% των τραπεζιών είχαν τέσσερα άτομα.

Ο πιο απλός τρόπος να σκεφτούμε την κατάσταση αυτή είναι να φανταστούμε εκατό μαθητές που πήγαν για φαγητό και αποτέλεσαν το δείγμα μας. Σύμφωνα με τα παραπάνω ποσοστά θα έπρεπε να είχαμε τα ακόλουθα αποτελέσματα:

20 άτομα έφαγαν μόνα τους

30 άτομα έφαγαν ανά δύο

30 άτομα έφαγαν ανά τρία και

20 άτομα έφαγαν τέσσερα

Ας υπολογίσουμε τον συνολικό αριθμό τραπεζιών που χρησιμοποιήθηκαν. Συγκεκριμένα θα πάρουμε τα εξής:

20 τραπέζια για τα άτομα που έφαγαν μόνα τους

15 τραπέζια για τα άτομα που έφαγαν δύο μαζί

10 τραπέζια για τα άτομα που έφαγαν τρία μαζί

5 τραπέζια για τα άτομα που έφαγαν σε τετράδες

Επομένως αυτό που προκύπτει τελικά είναι ότι για να καθίσουν τα εκατό άτομα χρειάστηκαν συνολικά 20+15+10+5 = 50 τραπέζια. Τότε λοιπόν προκύπτουν τα ακόλουθα ποσοστά:

$\frac{20}{50} = 40\%$ των τραπεζιών είχαν μόνο ένα άτομο

$\frac{15}{50} = 30\%$ των τραπεζιών είχαν δύο άτομα

$\frac{15}{50} = 20\%$ των τραπεζιών είχαν τρία άτομα

$\frac{5}{50} = 10\%$ των τραπεζιών είχαν τρία άτομα.

Όπως φαίνεται τα ποσοστά είναι διαφορετικά. Η πιθανοθεωρητική πλευρά του παραδείγματος ξεκινάει θεωρώντας ένα πείραμα όπου ένα κατειλημμένο τραπέζι επιλέγεται τυχαία και καταγράφεται ο αριθμός X των ατόμων που κάθονται εκεί. Τότε είναι σαφές ότι $\mathsf{P}[X = 1] = 0.4$ και ούτω καθεξής. Ένα διαφορετικό πείραμα, που σχετίζεται με το πρώτο αλλά δεν συγχέεται με αυτό, θα ήταν να επιλέξουμε ένα άτομο από τον πληθυσμό στην τύχη, και να καταγραφεί ο αριθμός X^* των ατόμων που κάθονται στο τραπέζι του. Τότε θα ισχύει ότι $\mathsf{P}[X^* = 1] = 0.2$ και ούτω καθεξής.

Παρακάτω δίνεται ο πίνακας με τις δύο συναρτήσεις πιθανότητας:

k	$\mathsf{P}[X = k]$	$\mathsf{P}[X^* = k]$
1	0.4	0.2
2	0.3	0.3
3	0.2	0.3
4	0.1	0.2

Οι κατανομές των τυχαίων μεταβλητών X και X^* όμως έχουν κάποια σχέση. Στο πείραμα με την τυχαία μεταβλητή X, κάθε τραπέζι έχει την ίδια πιθανότητα να επιλεγεί,

ενώ στην δεύτερη αυτό δεν ισχύει. Συγκεκριμένα, η πιθανότητα να επιλεγεί ένα τραπέζι είναι ανάλογη του αριθμού των ατόμων που κάθισαν εκεί. Έτσι λοιπόν η πιθανότητα $\mathsf{P}[X^* = k]$ είναι ανάλογη του $k\mathsf{P}[X = k]$. Επομένως

$$\mathsf{P}[X^* = k] = ck\mathsf{P}[X = k],$$

όπου c είναι μία σταθερά. Επειδή όμως

$$\begin{aligned} 1 &= \sum_k \mathsf{P}[X^* = k] \\ &= \sum_k ck\mathsf{P}[X = k] \\ &= c\sum_k k\mathsf{P}[X = k] \\ &= c\,\mathsf{E}[X]. \end{aligned}$$

Επομένως, από την τελευταία σχέση θα πάρουμε $c = 1/\mathsf{E}[X]$. Τελικά θα προκύψει ότι οι κατανομές των δύο τυχαίων μεταβλητών έχουν μία σχέση της μορφής

$$\mathsf{P}[X^* = k] = \frac{k\mathsf{P}[X = k]}{\mathsf{E}[X]}.$$

Προφανώς λοιπόν η X^* είναι η μεροληπτική λόγω μεγέθους εκδοχή της X.

1.2 Βιβλιογραφική ανασκόπηση

Η έννοια της σταθμισμένης κατανομής μπορεί να αναχθεί στον Fisher (1934), ο οποίος μελέτησε τα αποτελέσματα των μεθόδων για την εκτίμηση των συχνοτήτων εμφάνισης μίας σπάνιας ασθένειας. Παρ' όλ' αυτά η αρχική ιδέα της μεροληπτικής δειγματοληψίας λόγω μεγέθους εμφανίστηκε στον Cox (1962). Η έννοια όμως των σταθμισμένων κατανομών διαμορφώθηκε και αναπτύχθηκε από τον Rao (1965). Ο Rao συγκεκριμένα προσδιόρισε τις ακόλουθες τρεις βασικές πηγές που οδηγούν σε σταθμισμένες κατανομές:

Μη παρατηρησιμότητα των γεγονότων. Ορισμένα είδη γεγονότων ίσως και να μην μπορούν να εξακριβωθούν αν και βρίσκονται στην φύση. Ένα τυπικό παράδειγμα είναι η έρευνα στα παιδιά που έχουν κάποια σπάνια ασθένεια. Αν

μία οικογένεια και με τους δύο γονείς ετεροζυγώτες ως προς το γονίδιο της ασθένειας δεν έχει παιδιά με την ασθένεια αυτή, τότε δεν υπάρχει περίπτωση να εντοπιστούν οι γονείς αυτοί. Η μοναδική περίπτωση να εντοπιστούν είναι αν τουλάχιστον ένα τους παιδί έχει γεννηθεί με την ασθένεια αυτή. Η πραγματική συχνότητα του γεγονότος $A = \{0 \textbf{ ασθενή παιδιά}\}$ είναι μη εξακριβώσιμη. Ως εκ τούτου, η παρατηρούμενη κατανομή είναι η αρχική κατανομή περικεκομμένη στο 1 που είναι μία ειδική περίπτωση σταθμισμένης κατανομής.

Μερική καταστροφή των παρατηρήσεων. Ο Rao παρατήρησε ότι παρατηρήσεις που παράγονται από την φύση (όπως ο αριθμός των αβγών, ο αριθμός των ατυχημάτων κ.τ.λ.) ίσως να καταστραφούν μερικώς ή να εξακριβωθούν μερικώς. Στην περίπτωση αυτή, η παρατηρούμενη κατανομή είναι μία διαστρεβλωμένη εκδοχή της αρχικής κατανομής. Στην περίπτωση που ο μηχανισμός που επιφέρει τη μερική καταστροφή είναι γνωστός, η κατανομή ανάλογα με τις παρατηρούμενες τιμές, είναι μία σταθμισμένη εκδοχή της αρχικής κατανομής.

Δειγματοληψία με άνισες πιθανότητες επιλογής. Σε πολλές πρακτικές καταστάσεις μία απλή τυχαία δειγματοληψία δεν είναι εφικτή. Η δειγματοληψία πραγματοποιείται υπό ορισμένη προσέγγιση. Για παράδειγμα, τα ψάρια συλλέγονται μέσω ενός δικτυού, τα φυτά παρατηρούνται περπατώντας κατά μήκος κάποιας γραμμής ενώ τα πουλιά παρατηρούνται βάσει των ήχων που κάνουν. Η συγκεκριμένη δειγματοληπτική προσέγγιση οδηγεί σε στάθμιση, η οποία μεταβάλλει την αρχική κατανομή.

Κυρίως ο Cox (δες Cox, 1969) ασχολήθηκε με την εκτίμηση του μέσου της αρχικής κατανομής βασιζόμενη σε σταθμισμένες λόγω μεγέθους παρατηρήσεις. Για περισσότερη σαφήνεια, έστω Y τυχαία μεταβλητή με αντίστοιχη συνάρτηση πιθανότητας πυκνότητας $f(y)$ και συνάρτηση κατανομής $F(y)$. Έστω επίσης ότι η αντίστοιχη σταθμισμένη τυχαία μεταβλητή είναι η Y^W με συνάρτηση πιθανότητας πυκνότητας

$f^W(y)$ και συνάρτηση κατανομής $F^W(y)$. Στην περίπτωση που $w(y) = y$ ισχύει ότι

$$\begin{aligned}\mathsf{E}[(Y^W)^r] &= \int_{-\infty}^{\infty} y^r f^W(y)dy \\ &= \int_{-\infty}^{\infty} y^r \frac{yf(y)}{\mu}dy \\ &= \frac{1}{\mu}\int_{-\infty}^{\infty} y^{r+1} f(y)dy \\ &= \frac{\mu_{r+1}}{\mu},\end{aligned}$$

όπου μ_r είναι η r ροπή της Y. Τότε προκύπτει άμεσα ότι

$$\mathsf{E}\Big[\frac{1}{Y^W}\Big] = \frac{1}{\mu}. \tag{1.1}$$

Αντίστοιχα προκύπτει επίσης ότι

$$\mathsf{E}\Big[\big(\frac{1}{Y^W}\big)^2\Big] = \frac{\mu_{-1}}{\mu}. \tag{1.2}$$

Συνδυάζοντας τις (1.1) και (1.2) προκύπτει

$$\begin{aligned}\mathsf{Var}\Big[\frac{1}{Y^W}\Big] &= \mathsf{E}\Big[\Big(\frac{1}{Y^W}\Big)^2\Big] - \Big[\mathsf{E}\Big(\frac{1}{Y^W}\Big)\Big]^2 \\ &= \frac{\mu_{-1}}{\mu} - \frac{1}{\mu^2} \\ &= \frac{1}{\mu^2}(\mu\mu_{-1} - 1).\end{aligned} \tag{1.3}$$

Από την (1.1) καταλαβαίνουμε ότι ένας αμερόληπτος εκτιμητής για το $1/\mu$ είναι ο

$$\frac{1}{n}\sum_i \frac{1}{Y_i^W}.$$

Βάσει αυτού, ο Cox πρότεινε ως εκτιμητή για την μέση τιμή μ της Y τον

$$\widehat{\mu} = \frac{n}{\sum_i 1/Y_i^W}.$$

Από το Κεντρικό Οριακό Θεώρημα, ο εκτιμητής έχει ασυμπτωτική κανονική κατανομή[1] με μέσο μ και διασπορά $\mu\mu_{-1} - 1/n$. Όμως ο εκτιμητής του μ είναι στην

[1] Στην πραγματικότητα όπως θα φανεί σε επόμενα κεφάλαια, η ασυμπτωτική κατανομή προκύπτει για τον εκτιμητή πολλαπλασιασμένο με κάποιες σταθερές

πραγματικότητα μεροληπτικός. Έτσι λοιπόν, ο Sen (δες Sen, 1987) πρότεινε την χρήση jackknife για την μείωση της μεροληψίας του $\widehat{\mu}$.

Ο Cox (1969) ασχολήθηκε επίσης με την εκτίμηση της $F(y)$. Αυτό είναι κάτι το οποίο θα αναλυθεί και στην παρούσα διπλωματική. Επίσης μελετήθηκαν κάποιες ιδιαίτερες ιδιότητες σταθμισμένων κατανομών. Οι Patil and Rao (1978) μελέτησαν τη σχέση μεταξύ των μέσων της αρχικής κατανομής και διαφορετικών σταθμισμένων κατανομών. Οι Bayarri and Degroot (1987, 1989) και οι Patil and Taillie (1989) μελέτησαν την πληροφορία Fisher στις σταθμισμένες κατανομές και τις αντίστοιχες σχέσεις που προκύπτουν με την αρχική κατανομή.

Παρ' όλ' αυτά, στην παρούσα διπλωματική βασικό αντικείμενο μελέτης είναι η μη παραμετρική εκτίμηση της βασικής συνάρτησης κατανομής F, βάσει τυχαίων δειγμάτων από μία ή περισσότερες μεροληπτικές κατανομές. Με το αντικείμενο αυτό ασχολήθηκε εκτενώς ο Vardi, ο οποίος πρότεινε μεθόδους εκτίμησης της F βάσει επαναληπτικών αλγορίθμων και μελέτησε τις θεωρητικές ιδιότητες των εκτιμητών που προκύπτουν.

Αρχικά, θα μελετηθεί η περίπτωση των δύο δειγμάτων, και εν συνεχεία θα παρουσιαστεί ο αλγόριθμος εκτίμησης της F. Τέλος θα μελετηθεί ένα ημιπαραμετρικό μοντέλο υπό το οποίο οι πυκνότητες πιθανότητας f_0 και f_1 της βασικής και της μεροληπτικής κατανομής, αντίστοιχα, έχουν μία σχέση της μορφής $f_1(x) = e^{\alpha+\beta h(x)} f_0(x)$ για κάποια συνάρτηση $h(x)$.

Είναι σημαντικό να αναφερθεί (αν και δεν θα μελετηθεί στην παρούσα διπλωματική) ότι ενδιαφέρον παρουσιάζει και η περίπτωση που έχουμε διδιάστατες κατανομές (δες Patil, Rao and Ratnaparkhic, 2010). Συγκεκριμένα, έστω X, Y ένα ζεύγος μη αρνητικών τυχαίων μεταβλητών με από κοινού συνάρτηση πυκνότητας πιθανότητας $f(x, y)$ και έστω επίσης $w(x, y)$ μία μη αρνητική συνάρτηση. Τότε η αντίστοιχη σταθμισμένη εκδοχή της $f(x, y)$ δίνεται από τον τύπο

$$f^w(x, y) = \frac{w(x, y) f(x, y)}{\mathsf{E}[w(X, Y)]},$$

υπό την προυπόθεση ότι η $\mathsf{E}[w(X,Y)]$ είναι πεπερασμένη.
Από την σχέση αυτή είναι σαφές ότι προκύπτουν και τα ακόλουθα:

$$f^w(x) = \frac{\mathsf{E}[w(x,Y|X=x)]f(x,y)}{\mathsf{E}[w(X,Y)]}$$

και

$$f^w(y|x) = \frac{w(x,y)f(y|X=x)}{\mathsf{E}[w(x,Y|X=x)]}.$$

Προφανώς, οι δύο παραπάνω συναρτήσεις είναι σταθμισμένες εκδοχές της περιθωριακής και της δεσμευμένης κατανομής της Y δοθείσης της X.

Πληροφοριακά, παρουσιάζονται παρακάτω κάποιες από τις βασικές συναρτήσεις βάρους που χρησιμοποιούνται στην διδιάστατη περίπτωση:

- $w(x,y) = x^\alpha y^\beta$, $\alpha \geq 0, \beta \geq 0$
- $w(x,y) = \max(x,y)$
- $w(x,y) = \min(x,y)$
- $w(x,y) = x^\alpha + y^\beta$, $\alpha \geq 0, \beta \geq 0$

1.3 Η εμπειρική συνάρτηση κατανομής ως μη παραμετρικός ΕΜΠ

Ορισμός 1. Έστω $X_1, ..., X_n \in \Re$. Η εμπειρική συνάρτηση κατανομής των $X_1, \ldots, X_n$ δίνεται από τον τύπο

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,x]}(X_i),$$

για $-\infty < x < \infty$.

Η εμπειρική συνάρτηση κατανομής είναι μία διακριτή κατανομή που αντιστοιχίζει πιθανότητα $1/n$ σε κάθε μία από τις n παρατηρηθείσες τιμές. Σε περίπτωση που κάποια παρατήρηση έχει πολλαπλότητα t, η πιθανότητα που αντιστοιχίζεται σε αυτήν είναι t/n.

Ορισμός 2. Έστω $X_1, \ldots, X_n$ τυχαίο δείγμα από την κατανομή με σ.κ. F. Η μη παραμετρική (εμπειρική) συνάρτηση πιθανοφάνειας της F είναι

$$L(F) = \prod_{i=1}^{n} \mathsf{P}(X = x_i) = \prod_{i=1}^{n} \{F(X_i) - F(X_i-)\},$$

όπου $F(x-) = \mathsf{P}[X < x]$.

Μία άμεση συνέπεια του ορισμού είναι ότι $L(F) = 0$ αν η F είναι συνεχής κατανομή. Για να έχουμε θετική μη παραμετρική πιθανοφάνεια θα πρέπει η κατανομή F να αντιστοιχίζει θετική πιθανότητα σε κάθε μία από τις παρατηρηθείσες τιμές. Στο παρακάτω θεώρημα θα δειχθεί ότι η μη παραμετρική πιθανοφάνεια μεγιστοποιείται στην εμπειρική συνάρτηση κατανομής. Αυτό στην ουσία σημαίνει ότι η εμπειρική συνάρτηση κατανομής είναι ο **μη παραμετρικός εκτιμητής μεγίστης πιθανοφάνειας** για την F (ΜΠΕΜΠ).

Θεώρημα 1 (Owen, 2001)**.** Έστω $X_1, \ldots, X_n$ τυχαίο δείγμα από κάποια κατανομή. Έστω επίσης F_n η εμπειρική συνάρτηση κατανομής και F οποιαδήποτε άλλη συνάρτηση κατανομής. Αν $F \neq F_n$, τότε

$$L(F) < L(F_n).$$

Απόδειξη. Έστω $t_1 < \cdots < t_m$ οι διατεταγμένες παρατηρηθείσες τιμές των $X_1, \ldots, X_n$ και $n_j \geq 1$ το πλήθος των X_i που παίρνουν την τιμή t_j. Τότε $n_j = \sum_{i=1}^{n} I_{\{t_j\}}(X_i)$. Έστω επίσης $p_j = F(t_j) - F(t_j-)$ και $\widehat{p}_j = n_j/n$. Επειδή αν $p_j = 0$ για κάποιο $j = 1, 2, \ldots, m$ τότε $L(F) = 0 \leq L(F_n)$, θεωρούμε ότι $p_j > 0$, $\forall j$, και ότι για τουλάχιστον ένα j ισχύει ότι $p_j \neq \widehat{p}_j$. Γνωρίζουμε όμως ότι $\forall x > 0$ ισχύει $\log(x) \leq x - 1$ με την ισότητα να ισχύει αν και μόνον αν $x = 1$. Έτσι λοιπόν

$$\log\left(\frac{L(F)}{L(F_n)}\right) = \sum_{i=1}^{m} n_j \log\left(\frac{p_j}{\widehat{p}_j}\right) = n \sum_{i=1}^{m} \widehat{p}_j \log\left(\frac{p_j}{\widehat{p}_j}\right) < n \sum_{i=1}^{m} \widehat{p}_j \left(\frac{p_j}{\widehat{p}_j} - 1\right)$$

$$= n\Big\{\sum_{i=1}^{m} p_j - \sum_{i=1}^{m} \widehat{p}_j\Big\} \leq 0.$$

Επομένως

$$L(F) < L(F_n).$$

□

Κεφάλαιο 2

Εκτίμηση υπό μεροληψία λόγω μεγέθους

Στο κομμάτι αυτό της ανάλυσής μας θα βρούμε τον μη παραμετρικό εκτιμητή μεγίστης πιθανοφάνειας μίας συνάρτησης κατανομής F όταν έχουμε στην διάθεσή μας δύο δείγματα: Ένα δείγμα μεγέθους m από την κατανομή F και άλλο ένα μεγέθους n από μία σταθμισμένη εκδοχή της. Πιο συγκεκριμένα θα θεωρήσουμε την περίπτωση $G_Y(x) = \int_0^x u\, dF(u)/\mu$ όπου $\mu = \int_0^\infty u\, dF(u)$.

2.1 Εισαγωγικές έννοιες

Ορισμός 3. Έστω X τυχαία μεταβλητή με συνάρτηση κατανομής F. Θεωρούμε την τυχαία μεταβλητή Y με συνάρτηση κατανομής

$$G_Y(x) = \frac{\int_{-\infty}^x w(y)\, dF_X(y)}{\int_{-\infty}^\infty w(y)\, dF_X(y)}, \quad w(y) \geq 0.$$

Τότε η τυχαία μεταβλητή Y ονομάζεται σταθμισμένη τυχαία μεταβλητή με στάθμη (βάρος) $w(x)$. Στην περίπτωση που η συνάρτηση πυκνότητας πιθανότητας υπάρχει, δίνεται από τον τύπο

$$g_Y(x) = \frac{w(x) f_X(x)}{\int_{-\infty}^\infty w(x) f_X(x)}, \quad w(x) \geq 0.$$

Αν τα μήκη κάποιων αντικειμένων κατανέμονται σύμφωνα με την αθροιστική συνάρτηση κατανομής F, και αν επίσης η πιθανότητα να συλλέξουμε κάποιο αντικείμενο είναι ανάλογη του μήκους του, τότε τα μήκη των αντικειμένων αυτών κατανέμονται σύμφωνα με την συνάρτηση κατανομής που δίνεται από τον ακόλουθο τύπο (δες Vardi, 1982):

$$G_Y(y) \equiv G(y) = \frac{1}{\mu}\int_0^y x dF(x), \quad y \geq 0. \tag{2.1}$$

Εδώ $\mu = \int_0^\infty x dF(x)$ και υποθέτουμε πως $\mu < \infty$. Κατανομές όπως η G ονομάζονται *μεροληπτικές λόγω μεγέθους* κατανομές.

Στη συνέχεια θα ασχοληθούμε με την εύρεση του μη παραμετρικού εκτιμητή μεγίστης πιθανοφάνειας (ΜΠΕΜΠ) για την αθροιστική συνάρτηση κατανομής F με χρήση ενός τυχαίου δείγματος που προέρχεται από την F και ενός ανεξάρτητου τυχαίου δείγματος που προέρχεται από τη σταθμισμένη εκδοχή της G.

2.2 Εκτίμηση

Έστω λοιπόν ότι $\mathcal{X} = (X_1, \ldots, X_m)$ ένα τυχαίο δείγμα από την F με $F(0) = 0$ και $\mu = \int_0^\infty x dF(x) < \infty$ ενώ $\mathcal{Y} = Y_1, \ldots, Y_n$ ένα ανεξάρτητο τυχαίο δείγμα από την G με μορφή όπως η (2.1). Παίρνουμε όλες τις παρατηρήσεις που διαθέτουμε, δηλαδή την ένωση[1] $\mathcal{X} \bigcup \mathcal{Y}$ και τις διατάσσουμε από την μικρότερη στην μεγαλύτερη. Έστω $t_1 < \cdots < t_h$, $h \leq n+m$ καθώς υπάρχει το ενδεχόμενο να υπάρχουν παρατηρήσεις με ίδια τιμή. Θεωρούμε επίσης ότι η πολλαπλότητα του t_i στα X και τα Y είναι αντίστοιχα ξ_i και η_i. Είναι σαφές ότι αν η F είναι συνεχής κατανομή τότε με πιθανότητα ένα θα ισχύει ότι $h = n + m$ και επίσης $\xi_i + \eta_i = 1$.

Η πιθανοφάνεια της F δίνεται από την σχέση

$$L(F) = \prod_{i=1}^{h} \{dF(t_i)\}^{\xi_i} \left\{ \frac{t_i dF(t_i)}{\int_0^\infty u dF(u)} \right\}^{\eta_i}. \tag{2.2}$$

[1] Όπως γνωρίζουμε η ένωση δύο συνόλων διαγράφει τα επαναλαμβανόμενα στοιχεία κάτι το οποίο εδώ δεν πρέπει να γίνει. Επομένως εδώ με την ένωση αναφερόμαστε στο σύνολο όλων των δεδομένων.

Στόχος μας είναι σαφώς να βρούμε την συνάρτηση κατανομής η οποία μεγιστοποιεί την (2.2). Όμως για την μεγιστοποίηση αυτή αρκεί να περιοριστούμε στην εύρεση μίας διακριτής κατανομής η οποία θα έχει θετικα άλματα σε κάθε ένα από τα σημεία $t_1, ..., t_h$ και μόνο εκεί. (Η περίπτωση είναι ανάλογη με αυτήν της Ενότητας 1.3.) Συγκεκριμένα, η μεγιστοποίηση της (2.2) ισοδυναμεί με την μεγιστοποίηση της συνάρτησης

$$L(p_1, ..., p_h) = \prod_{i=1}^{h} p_i^{\xi_i} \left(\frac{t_i p_i}{\sum_{j=1}^{h} t_j p_j} \right)^{\eta_i}, \tag{2.3}$$

όπου $\sum_{j=1}^{h} p_j = 1$, $p_j > 0$ για $j = 1, 2, \ldots, h$ και $p_j = dF(t_j)$ είναι η μάζα πιθανότητας του t_j. Είναι σαφές ότι ο παρονομαστής στο γινόμενο της σχέσης (2.3) εμφανίζεται ως άθροισμα αντί για ολοκλήρωμα καθώς μιλάμε πλέον για μία διακριτή κατανομή. Αν συμβολίσουμε με $\widehat{p} = (\widehat{p}_1, ..., \widehat{p}_h)$ την λύση που προκύπτει από την μεγιστοποίηση της (2.3), τότε ο μη παραμετρικός εκτιμητής μεγίστης πιθανοφάνειας για την F θα δίνεται από την σχέση

$$\widehat{F}(x) = \sum_{j=1}^{h} \widehat{p}_j I_{(-\infty, x]}(t_j) \tag{2.4}$$

και τότε ικανοποιείται $L(\widehat{F}) \geq L(F_1)$ για οποιαδήποτε άλλη συνάρτηση κατανομής F_1.

Θεώρημα 2. Η μοναδική λύση που προκύπτει από την μεγιστοποίηση της (2.3) δίνεται από τον τύπο

$$\widehat{p}_k = \frac{(\xi_k + \eta_k)\,\widehat{\mu}}{n t_k + m\widehat{\mu}}, \quad k = 1, \ldots, h, \tag{2.5}$$

όπου $\widehat{\mu}$ είναι η μοναδική λύση της εξίσωσης

$$\sum_{k=1}^{h} \frac{(\xi_k + \eta_k)\, t_k}{n t_k + m a} = 1 \tag{2.6}$$

ως προς a.

Απόδειξη. Η μεγιστοποίηση της (2.3) θα γίνει υπό τον περιορισμό $\sum_{i=1}^{h} p_i = 1$. Ο

λογάριθμος της συνάρτησης πιθανοφάνειας είναι

$$\begin{aligned} l(p_1,...,p_h) &= \log[L(p_1,...,p_h)] \\ &= \log\left[\prod_{i=1}^{h} p_i^{\xi_i}\left(\frac{t_i p_i}{\sum_{j=1}^{h} t_j p_j}\right)^{\eta_i}\right] \\ &= \sum_{i=1}^{h}\xi_i \log(p_i) + \sum_{i=1}^{h}\eta_i \log(t_i p_i) - \sum_{i=1}^{h}\eta_i \log(\sum_{j=1}^{h} t_j p_j). \end{aligned}$$

Εισάγοντας έναν πολλαπλασιαστή Lagrange ορίζουμε την συνάρτηση

$$g(p_1,\ldots,p_h) = \sum_{i=1}^{h}\xi_i \log(p_i) + \sum_{i=1}^{h}\eta_i \log(t_i p_i) - \sum_{i=1}^{h}\eta_i \log(\sum_{j=1}^{h} t_j p_j) - \lambda(\sum_{i=1}^{h} p_i - 1).$$

Παραγωγίζοντας ως προς p_k και εξισώνοντας με το μηδέν προκύπτει

$$\frac{\xi_k + \eta_k}{p_k} - \frac{t_k}{\sum_{j=1}^{h} t_j p_j}\sum_{i=1}^{h}\eta_i - \lambda = 0. \tag{2.7}$$

Πολλαπλασιάζοντας την (2.7) με p_k και αθροίζοντας ως προς k θα προκύψει

$$\sum_{k=1}^{h}(\xi_k + \eta_k) - \sum_{i=1}^{h}\eta_i - \lambda = 0,$$

το οποίο συνεπάγεται ότι $\lambda = \sum_{k=1}^{h}\xi_k$.

Επομένως από την (2.7) θα έχουμε

$$\frac{\xi_k+\eta_k}{p_k}-\frac{t_k}{\sum_{j=1}^{h}t_jp_j}\sum_{i=1}^{h}\eta_i-\sum_{k=1}^{h}\xi_k=0\Rightarrow$$
$$\frac{\xi_k+\eta_k}{p_k}=\sum_{k=1}^{h}\xi_k+\frac{t_k}{\sum_{j=1}^{h}t_jp_j}\sum_{i=1}^{h}\eta_i\Rightarrow$$
$$\frac{\xi_k+\eta_k}{p_k}=\frac{(\sum_{j=1}^{h}t_jp_j)\sum_{k=1}^{h}\xi_k+t_k\sum_{i=1}^{h}\eta_i}{\sum_{j=1}^{h}t_jp_j}\Rightarrow$$
$$\frac{\xi_k+\eta_k}{p_k}=\frac{(\sum_{j=1}^{h}t_jp_j)m+t_kn}{\sum_{j=1}^{h}t_jp_j}\Rightarrow$$
$$p_k=\frac{\sum_{j=1}^{h}t_jp_j(\xi_k+\eta_k)}{m\sum_{j=1}^{h}t_jp_j+nt_k}\Rightarrow$$
$$p_k=\frac{\mu(\xi_k+\eta_k)}{m\mu+nt_k},$$

όπου $\mu=\sum_{j=1}^{h}t_jp_j$. Επομένως για τον εκτιμητή του p_k θα ισχύει

$$\hat{p}_k=\frac{\hat{\mu}(\xi_k+\eta_k)}{m\hat{\mu}+nt_k}, \tag{2.8}$$

όπου $\hat{\mu}=\sum_{j=1}^{h}t_j\hat{p}_j$. Πολλαπλασιάζοντας την (2.8) με t_k και αθροίζοντας ως προς k θα πάρουμε

$$\sum_{k=1}^{h}\hat{p}_kt_k=\sum_{k=1}^{h}\frac{t_k\hat{\mu}(\xi_k+\eta_k)}{m\hat{\mu}+nt_k}=\hat{\mu}\sum_{k=1}^{h}\frac{t_k(\xi_k+\eta_k)}{m\hat{\mu}+nt_k}\Leftrightarrow$$
$$\sum_{k=1}^{h}\frac{(\xi_k+\eta_k)t_k}{nt_k+m\hat{\mu}}=1.$$

□

Ο παραπάνω εκτιμητής έχει νόημα ακόμα και όταν κάποιο από τα δύο δείγματα δεν υπάρχει. Συγκεκριμένα:

- Όταν $n=0$, δηλαδή δεν υπάρχει το δείγμα Y, τότε είναι σαφές ότι ο εκτιμητής $\widehat{\mu}$ είναι ο μέσος του δείγματος X και επομένως η εκτίμηση για την συνάρτηση κατανομής είναι η εμπειρική συνάρτηση κατανομής βασιζόμενη στο δείγμα X.

- Αντίστοιχα αν $m = 0$, δηλαδή δεν υπάρχει το δείγμα X και είναι διαθέσιμο μόνο το σταθμισμένο δείγμα Y, τότε προφανώς $\xi_k = 0$. Στην περίπτωση αυτή η (2.5) θα γίνει

$$\widehat{p}_k = \frac{\eta_k \widehat{\mu}}{n t_k}.$$

Αθροίζοντας λοιπόν ως προς k θα έχουμε:

$$\sum_{k=1}^{h} \widehat{p}_k = \sum_{k=1}^{h} \frac{\eta_k \widehat{\mu}}{n t_k} = \frac{\widehat{\mu}}{n} \sum_{k=1}^{n} \frac{1}{y_k}.$$

Όμως $\sum_{k=1}^{n} \widehat{p}_k = 1$. Επομένως,

$$1 = \frac{\widehat{\mu}}{n} \sum_{k=1}^{n} \frac{1}{y_k} \Rightarrow$$
$$\widehat{\mu} = \frac{n}{\sum_{k=1}^{n} \frac{1}{y_k}}.$$

Με αυτό λοιπόν καταλαβαίνουμε ότι ο εκτιμητής στην περίπτωση αυτή θα είναι ο αρμονικός μέσος του δείγματος Y. Ο εκτιμητής αυτός είναι εκείνος που προτάθηκε από τον Cox (1969).

Κατ' αντιστοιχία με την (2.4) ο ΜΠΕΜΠ της G είναι ο

$$\widehat{G}(x) = \sum_{i=1}^{h} \widehat{g}_i I_{(-\infty, x]}(t_i),$$

όπου

$$\widehat{g}_i = \frac{t_i \widehat{p}_i}{\widehat{\mu}} = \frac{(\xi_i + \eta_i) t_i}{n t_i + m \widehat{\mu}}.$$

2.3 Ασυμπτωτική συμπεριφορά του ΜΠΕΜΠ

2.3.1 Σύγκλιση κατά πιθανότητα

Έχουμε αποδείξει ότι όταν έχουμε στη διάθεσή μας δύο δείγματα, τότε ο ΜΠΕΜΠ $\widehat{\mu}$ του μ ικανοποιεί τη συνθήκη (2.6). Θα αποδείξουμε αρχικά το ακόλουθο:

Θεώρημα 3. (Gill, Vardi and Wellner, 1988)

$$\widehat{\mu} \overset{\sigma.\beta.}{\to} \mu.$$

Απόδειξη. Για κάθε a ισχύει

$$\begin{aligned}
&\sum_{k=1}^{h} \frac{(\xi_k + \eta_k)\, t_k}{nt_k + ma} - 1 \\
&= \sum_{i=1}^{m} \frac{x_i}{nx_i + ma} + \sum_{i=1}^{n} \frac{y_i}{ny_i + ma} - 1 \\
&= \sum_{i=1}^{m} \frac{x_i}{nx_i + ma} + \frac{1}{n} \sum_{i=1}^{n} \frac{ny_i}{ny_i + ma} - \frac{1}{n} \sum_{i=1}^{n} \frac{ny_i + ma}{ny_i + ma} \\
&= \sum_{i=1}^{m} \frac{x_i}{nx_i + ma} - \frac{ma}{n} \sum_{i=1}^{n} \frac{1}{ny_i + ma} \\
&= m\Big[\frac{1}{m} \sum_{i=1}^{m} \frac{x_i}{nx_i + ma} - \frac{a}{n} \sum_{i=1}^{n} \frac{1}{ny_i + ma}\Big] \\
&= \frac{m}{m+n}\Big[\frac{1}{m} \sum_{i=1}^{m} \frac{x_i}{(1-\lambda)\, x_i + \lambda a} - \frac{a}{n} \sum_{i=1}^{n} \frac{1}{(1-\lambda)\, y_i + \lambda a}\Big],
\end{aligned}$$

όπου $\lambda = m/(m+n) = m/N$ και $N = m + n$. Η παραπάνω σχέση προέκυψε πολλαπλασιάζοντας και τα δύο μέλη με την ποσότητα $m + n$.

Έστω τώρα $\widehat{a}_N$ η λύση της εξίσωσης

$$\widehat{h}_N(a) = \frac{1}{m} \sum_{i=1}^{m} \frac{x_i}{(1-\lambda)\, x_i + \lambda a} - \frac{a}{n} \sum_{\iota=1}^{n} \frac{1}{(1-\lambda)\, y_i + \lambda a} = 0 \tag{2.9}$$

ως προς a. Το $\widehat{a}_N$ συμπίπτει με το $\widehat{\mu}$ της (2.6).

Έστω επίσης

$$h(a) = \mathsf{E}\left(\frac{X}{(1-\lambda)\, X + \lambda a}\right) - a\mathsf{E}\left(\frac{1}{(1-\lambda)\, Y + \lambda a}\right). \tag{2.10}$$

Είναι σαφές ότι

$$h(a) = \int \frac{x}{(1-\lambda)\, x + \lambda a} f(x)\, dx - a \int \frac{1}{(1-\lambda)\, x + \lambda a} \left(\frac{x f(x)}{\mu}\right) dx$$

και αυτό είναι ίσο με μηδέν αν και μόνο αν $\alpha = \mu$. Επίσης, σύμφωνα με τον Ισχυρό Νόμο των Μεγάλων Αριθμών, όταν $N \to \infty$ τότε $\widehat{h}_N(a) \overset{\sigma.\beta.}{\to} h(a)$. Όμως παράλληλα ισχύει ότι

$$\frac{d}{da}\widehat{h}_N(a) = -\frac{\lambda}{m}\sum_{i=1}^{m}\left(\frac{x_i}{\left[(1-\lambda)x_i + \lambda a\right]^2}\right) - \frac{1-\lambda}{n}\sum_{i=1}^{n}\left(\frac{y_i}{\left[(1-\lambda)y_i + \lambda a\right]^2}\right) < 0,\ \forall a.$$

Αναπτύσσοντας κατά Taylor ως προς $\widehat{\mu}$ γύρω από το μ θα πάρουμε

$$\widehat{h}_N(\widehat{\mu}) = \widehat{h}_N(\mu) + (\widehat{\mu} - \mu)\frac{d}{da}\widehat{h}_N(a^*) \tag{2.11}$$

για κάποιο a^* ανάμεσα στα $\widehat{\mu}$ και μ. Όμως εξ ορισμού $\widehat{h}_N(\widehat{\mu}) = 0$ και επομένως

$$\widehat{h}_N(\mu) + (\widehat{\mu} - \mu)\frac{d}{da}\widehat{h}_N(a^*) = 0.$$

Παράλληλα ισχύει ότι $\widehat{h}_N(\mu) \overset{\sigma.\beta.}{\to} h(\mu) = 0$ καθώς επίσης ότι η ποσότητα $\frac{d}{da}\widehat{h}_N(a^*)$ δεν μπορεί να γίνει μηδέν. Άρα λοιπόν όταν $N \uparrow \infty$ απαραιτήτως πρέπει να ισχύει ότι $\widehat{\mu} \overset{\sigma.\beta.}{\to} \mu$. □

2.3.2 Ασυμπτωτική κατανομή του $\widehat{\mu}$

Σε αυτήν την ενότητα θα βρούμε την ασυμπτωτική κατανομή του ΜΠΕΜΠ του μ. Έστω

$$K(x) = \int_0^x \frac{y}{\lambda\mu + (1-\lambda)y} f(y)\, dy,$$

ενώ

$$K = \lim_{x\to\infty} K(x) = \int_0^\infty \frac{y}{\lambda\mu + (1-\lambda)y} f(y)\, dy.$$

Θα αποδειχθεί το ακόλουθο θεώρημα:

Θεώρημα 4. (Vardi, 1982) Για τον ΜΠΕΜΠ του μ ισχύει

$$\sqrt{N}\left(\frac{\widehat{\mu}}{\mu}-1\right)\xrightarrow{d}N\left(0,\frac{1-K}{K\lambda\left(1-\lambda\right)}\right)$$

Απόδειξη. Ξεκινώντας από την (2.9), σε συνδυασμό με την (2.10) και παίρνοντας τα αθροίσματα, προκύπτει ότι

$$\sqrt{m}\Big[\frac{1}{m}\sum_{i=1}^{m}\Big(\frac{x_i}{\left(1-\lambda\right)x_i+\lambda a}\Big)-\mathsf{E}\Big(\frac{X}{\left(1-\lambda\right)X+\lambda a}\Big)\Big]\xrightarrow{d}N\left(0,\mathsf{Var}\Big(\frac{X}{\left(1-\lambda\right)X+\lambda a}\Big)\right)$$

$$\sqrt{n}\Big[\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{1}{\left(1-\lambda\right)y_i+\lambda a}\Big)-\mathsf{E}\Big(\frac{1}{\left(1-\lambda\right)Y+\lambda a}\Big)\Big]\xrightarrow{d}N\left(0,\mathsf{Var}\Big(\frac{1}{\left(1-\lambda\right)Y+\lambda a}\Big)\right)$$

Λαμβάνοντας υπ' όψιν ότι $m/N=\lambda$ και $n/N=1-\lambda$, από τις παραπάνω σχέσεις προκύπτει ισοδύναμα ότι

$$\sqrt{N}\left\{\frac{1}{m}\sum_{i=1}^{m}\left(\frac{x_i}{\left(1-\lambda\right)x_i+\lambda a}\right)-\mathsf{E}\left(\frac{X}{\left(1-\lambda\right)X+\lambda a}\right)\right\}\xrightarrow{d}N\left(0,\frac{1}{\lambda}\mathsf{Var}\Big(\frac{X}{\left(1-\lambda\right)X+\lambda a}\Big)\right)$$

$$\sqrt{N}\left\{\frac{1}{n}\sum_{j=1}^{n}\left(\frac{1}{\left(1-\lambda\right)y_i+\lambda a}\right)-\mathsf{E}\left(\frac{1}{\left(1-\lambda\right)Y+\lambda a}\right)\right\}\xrightarrow{d}N\left(0,\frac{1}{1-\lambda}\mathsf{Var}\Big(\frac{1}{\left(1-\lambda\right)Y+\lambda a}\Big)\right)$$

Έστω τώρα

$$W=\frac{1}{m}\sum_{i=1}^{m}\left(\frac{x_i}{\left(1-\lambda\right)x_i+\lambda a}\right)-\frac{\alpha}{n}\sum_{i=1}^{n}\left(\frac{1}{\left(1-\lambda\right)y_i+\lambda a}\right)$$

με

$$\mathsf{E}\left(W\right)=\mathsf{E}\left(\frac{X}{\left(1-\lambda\right)X+\lambda a}\right)-\alpha\mathsf{E}\left(\frac{1}{\left(1-\lambda\right)Y+\lambda a}\right).$$

Εφαρμόζοντας το Κεντρικό Οριακό Θεώρημα σε συνδυασμό με την ανεξαρτησία των X και Y παίρνουμε

$$\sqrt{N}\left\{W-\mathsf{E}\left(W\right)\right\}\xrightarrow{d}N\left(0,\frac{1}{\lambda}\mathsf{Var}\Big(\frac{X}{\left(1-\lambda\right)X+\lambda a}\Big)+\frac{\alpha^2}{1-\lambda}\mathsf{Var}\Big(\frac{1}{\left(1-\lambda\right)Y+\lambda a}\Big)\right).$$

Ειδικότερα, για $\alpha = \mu$,

$$\sqrt{N}\left\{W - \mathsf{E}\left(W\right)\right\} \xrightarrow{d} N\left(0, \frac{1}{\lambda}\mathsf{Var}\Big(\frac{X}{(1-\lambda)X + \lambda\mu}\Big) + \frac{\mu^2}{1-\lambda}\mathsf{Var}\Big(\frac{1}{(1-\lambda)Y + \lambda\mu}\Big)\right).$$

Στη συνέχεια θα υπολογιστεί η διακύμανση

$$V^2 = \underbrace{\frac{1}{\lambda}\mathsf{Var}\Big(\frac{X}{(1-\lambda)X + \lambda a}\Big)}_{V_1^2} + \underbrace{\frac{\mu^2}{1-\lambda}\mathsf{Var}\Big(\frac{1}{(1-\lambda)Y + \lambda a}\Big)}_{V_2^2}.$$

Αυτό θα γίνει υπολογίζοντας κάθε ένα από τα δύο μέρη χωριστά. Έτσι λοιπόν ισχύει

$$\begin{aligned}
V_1^2 &= \frac{1}{\lambda}\mathsf{Var}\left[\frac{X}{(1-\lambda)X + \lambda\mu}\right] \\
&= \frac{1}{\lambda}\left(\mathsf{E}\left[\frac{X^2}{\left[(1-\lambda)X + \lambda\mu\right]^2}\right] - \mathsf{E}^2\left[\frac{X}{(1-\lambda)X + \lambda\mu}\right]\right) \\
&= \frac{1}{\lambda}\left(\int_0^\infty \frac{x^2}{\left[(1-\lambda)x + \lambda\mu\right]^2} f(x)dx - \left[\int_0^\infty \frac{x}{(1-\lambda)x + \lambda\mu} f(x)dx\right]^2\right) \\
&= \frac{1}{\lambda}\left(\int_0^\infty \frac{x^2}{\left[(1-\lambda)x + \lambda\mu\right]^2} f(x)dx - K^2\right), \qquad (2.12)
\end{aligned}$$

ενώ

$$\begin{aligned}
V_2^2 &= \frac{\mu^2}{1-\lambda}\mathsf{Var}\left[\frac{1}{(1-\lambda)Y + \lambda\mu}\right] \\
&= \frac{\mu^2}{1-\lambda}\left(\mathsf{E}\left[\frac{1}{\left[(1-\lambda)Y + \lambda\mu\right]^2}\right] - \mathsf{E}^2\left[\frac{1}{(1-\lambda)Y + \lambda\mu}\right]\right) \\
&= \frac{\mu^2}{1-\lambda}\left(\int_0^\infty \frac{1}{\left[(1-\lambda)y + \lambda\mu\right]^2} g(y)dy - \left[\int_0^\infty \frac{1}{(1-\lambda)y + \lambda\mu} g(y)dy\right]^2\right) \\
&= \frac{\mu^2}{1-\lambda}\left(\int_0^\infty \frac{y}{\left[(1-\lambda)y + \lambda\mu\right]^2}\frac{f(y)}{\mu}dy - \left[\int_0^\infty \frac{y}{(1-\lambda)y + \lambda\mu}\frac{f(y)}{\mu}dy\right]^2\right) \\
&= \frac{\mu^2}{1-\lambda}\left(\frac{1}{\mu}\int_0^\infty \frac{x}{\left[(1-\lambda)x + \lambda\mu\right]^2} f(x)dx - \frac{1}{\mu^2}\left[\int_0^\infty \frac{x}{(1-\lambda)x + \lambda\mu} f(x)dx\right]^2\right) \\
&= \frac{\mu}{1-\lambda}\left(\int_0^\infty \frac{x}{\left[(1-\lambda)x + \lambda\mu\right]^2} f(x)dx\right) - \frac{1}{1-\lambda}K^2. \qquad (2.13)
\end{aligned}$$

Συνδυάζοντας τις (2.12) και (2.13) θα προκύψει

$$\begin{aligned}
V^2 &= \int_0^\infty \frac{1}{\left[(1-\lambda)\,x+\lambda\mu\right]^2}\Big(\frac{x^2}{\lambda}+\frac{x\mu}{1-\lambda}\Big)f(x)dx-\frac{K^2}{\lambda}-\frac{K^2}{1-\lambda}\\
&= \int_0^\infty \frac{1}{\left[(1-\lambda)\,x+\lambda\mu\right]^2}\Big(\frac{(1-\lambda)\,x^2+\lambda\mu x}{\lambda\,(1-\lambda)}\Big)f(x)dx-\frac{K^2}{\lambda\,(1-\lambda)}\\
&= \int_0^\infty \frac{1}{\left[(1-\lambda)\,x+\lambda\mu\right]^2}\Big(\frac{x\left[(1-\lambda)\,x+\lambda\mu\right]}{\lambda\,(1-\lambda)}\Big)f(x)dx-\frac{K^2}{\lambda\,(1-\lambda)}\\
&= \frac{1}{\lambda\,(1-\lambda)}\int_0^\infty \frac{x}{(1-\lambda)\,x+\lambda\mu}f(x)dx-\frac{K^2}{\lambda\,(1-\lambda)}\\
&= \frac{K}{\lambda\,(1-\lambda)}-\frac{K^2}{\lambda}-\frac{K^2}{1-\lambda}\\
&= \frac{K}{\lambda\,(1-\lambda)}-\frac{K^2}{\lambda\,(1-\lambda)}\\
&= \frac{K\,(1-K)}{\lambda\,(1-\lambda)}.
\end{aligned}$$

Όμως χρησιμοποιώντας τον Ασθενή Νόμο των Μεγάλων Αριθμών (ΑΝΜΑ) επειδή $\alpha^* \xrightarrow{p} \mu$, έχουμε:

$$\begin{aligned}
\frac{d}{d\alpha^*}\widehat{h}_N(\alpha^*) &= -\frac{\lambda}{m}\sum_{i=1}^{m}\frac{x_i}{\left[(1-\lambda)\,x_i+\lambda\alpha^*\right]^2}-\frac{1-\lambda}{n}\sum_{i=1}^{n}\frac{y_i}{\left[(1-\lambda)\,y_i+\lambda\alpha^*\right]^2}\\
&\xrightarrow{p} -\lambda\mathsf{E}\left(\frac{X}{\left[(1-\lambda)\,X+\lambda\mu\right]^2}\right)-(1-\lambda)\,\mathsf{E}\left(\frac{Y}{\left[(1-\lambda)\,Y+\lambda\mu\right]^2}\right)\\
&= -\lambda\int_0^\infty \frac{x}{\left[(1-\lambda)\,x+\lambda\mu\right]^2}f(x)dx-(1-\lambda)\int_0^\infty \frac{y}{\left[(1-\lambda)\,y+\lambda\mu\right]^2}g(y)dy\\
&= -\lambda\int_0^\infty \frac{x}{\left[(1-\lambda)\,x+\lambda\mu\right]^2}f(x)dx-(1-\lambda)\int_0^\infty \frac{x^2}{\left[(1-\lambda)\,x+\lambda\mu\right]^2}\frac{f(x)}{\mu}dx\\
&= \int_0^\infty \frac{1}{\left[(1-\lambda)\,x+\lambda\mu\right]^2}\Big(-\lambda x-\frac{(1-\lambda)\,x^2}{\mu}\Big)f(x)dx\\
&= \int_0^\infty \frac{1}{\left[(1-\lambda)\,x+\lambda\mu\right]^2}\Big(-\frac{x\left[\lambda\mu+(1-\lambda)\,x\right]}{\mu}\Big)f(x)dx\\
&= -\frac{1}{\mu}\int_0^\infty \frac{x}{(1-\lambda)\,x+\lambda\alpha}f(x)dx\\
&= -\frac{K}{\mu}.
\end{aligned}$$

Άρα λοιπόν

$$\frac{d}{d\alpha}\widehat{h}_N\left(\alpha^*\right) \xrightarrow{p} -\frac{K}{\mu}.$$

Χρησιμοποιώντας την σχέση (2.11) θα προκύψει ότι

$$0 = \widehat{h}_N\left(\mu\right) + \left(\widehat{a}_N - \mu\right)\frac{d}{da}\widehat{h}_N\left(a^*\right) \Leftrightarrow$$

$$\left(\widehat{\alpha}_N - \mu\right) = \frac{\widehat{h}_N\left(\mu\right)}{-\frac{d}{d\alpha}\widehat{h}_N\left(\alpha^*\right)} \Leftrightarrow$$

$$\sqrt{N}\left(\widehat{\alpha}_N - \mu\right) = \frac{\sqrt{N}\widehat{h}_N\left(\mu\right)}{-\frac{d}{d\alpha}\widehat{h}_N\left(\alpha^*\right)}.$$

Εφαρμόζοντας το θεώρημα Slutsky και χρησιμοποιώντας τις παραπάνω σχέσεις παίρνουμε

$$\sqrt{N}\left(\widehat{\alpha}_N - \mu\right) \xrightarrow{d} N\Big(0, \frac{\left(1-K\right)\mu^2}{K\lambda\left(1-\lambda\right)}\Big).$$

Τέλος, με εφαρμογή της μεθόδου δέλτα προκύπτει ότι

$$\sqrt{N}\Big(\frac{\widehat{\alpha}_N}{\mu} - 1\Big) \xrightarrow{d} N\Big(0, \frac{1-K}{K\lambda\left(1-\lambda\right)}\Big).$$

□

Να παρατηρήσουμε εδώ ότι ισχύει το ακόλουθο. Έστω μ_i η i ροπή της F και ας υποθέσουμε ότι οι ροπές μ_{-1} και μ_2 υπάρχουν και είναι πεπερασμένες. Τότε ισχύει ότι

$$\lim_{\lambda\to\lambda_0}\frac{\mu^2\left(1-K\right)}{K\lambda\left(1-\lambda\right)} = \begin{cases}\mu^2\left(\mu\mu_{-1}-1\right), & \lambda_0 = 0\\ \mu_2 - \mu^2 = \mathsf{Var}\left(X\right), & \lambda_0 = 1\end{cases}$$

Στην πραγματικότητα αυτή η σχέση δείχνει τη διακύμανση της ασυμπτωτικής κατανομής του εκτιμητή στις περιπτώσεις που τα δεδομένα αποτελούνται από ένα μόνο δείγμα, είτε αυτό είναι από την σταθμισμένη κατανομή είτε από την βασική κατανομή. Βλέπουμε ότι και στην περίπτωση αυτή η ασυμπτωτική κατανομή είναι η κανονική κατανομή.

2.4 Η περίπτωση περισσότερων από δύο δειγμάτων

Έστω τώρα $\mathcal{Y}_i \equiv (Y_{i1}, \ldots, Y_{in_i})$, $i = 1, \ldots, s$, ανεξάρτητα τυχαία δείγματα τα οποία προέρχονται από τις συναρτήσεις κατανομής

$$F_i(t) = W_i(F)^{-1} \int_{-\infty}^{t} w_i(u) dF(u), \quad i = 1, 2, \ldots, s,$$

όπου F είναι μία άγνωστη συνάρτηση κατανομής και

$$W_i(F) = \int_{-\infty}^{\infty} w_i(u) dF(u), \quad i = 1, 2, \ldots, s,$$

ενώ $s \geq 2$. Υποθέτουμε επίσης ότι οι συναρτήσεις w_i είναι γνωστές και μη αρνητικές ενώ ταυτόχρονα ικανοποιείται η συνθήκη $0 < W_i(F) < \infty \ \forall i = 1, 2, \ldots, s$. Το πρόβλημά μας είναι η εύρεση του ΜΠΕΜΠ της F (δες Vardi, 1985) χρησιμοποιώντας τα δείγματα $\mathcal{Y}_1, \ldots, \mathcal{Y}_s$.

2.4.1 Η συνάρτηση πιθανοφάνειας και τα δεδομένα

Έστω $t_1 < t_2 < \cdots < t_h$ όλα τα δεδομένα μας, δηλαδή από την ένωση $\mathcal{Y}_1 \cup \ldots \cup \mathcal{Y}_s$ τοποθετημένα σε αύξουσα σειρά. Εδώ είναι $h \leq \sum_{i=1}^{s} n_i$ λόγω πιθανών ίδιων τιμών (ties). Έστω επίσης η_{ij} η πολλαπλότητα του t_j στο $\mathcal{Y}_i$, όπου $j = 1, 2, \ldots, h$ και $i = 1, \ldots, s$. Τέλος η συνολική πολλαπλότητα του t_j και το μέγεθος του δείγματος $\mathcal{Y}_i$ δίνεται από τους τύπους

$$r_j = \sum_{i=1}^{s} \eta_{ij} \quad \text{και} \quad n_i = \sum_{j=1}^{h} \eta_{ij}.$$

Με βάση τα παραπάνω, η εμπειρική πιθανοφάνεια δίνεται από την σχέση

$$L(F) = \prod_{j=1}^{h} \prod_{i=1}^{s} \left(\frac{w_i(t_j) dF(t_j)}{W_i(F)} \right)^{\eta_{ij}}.$$

Όπως και στην περίπτωση των δύο δειγμάτων, η μεγιστοποίηση της L θα γίνει ψάχνοντας την κλάση των διακριτών κατανομών οι οποίες έχουν θετικά άλματα σε κάθε

ένα από τα σημεία $t_1, \ldots, t_h$. Θεωρούμε δηλαδή ότι $p_j = dF(t_j)\ \forall j$ και επομένως θα πρέπει να μεγιστοποιήσουμε την

$$L(p) = \prod_{j=1}^{h} \prod_{i=1}^{s} \left(\frac{w_{ij} p_j}{W_i(\mathbf{p})} \right)^{\eta_{ij}}, \tag{2.14}$$

υπό τη συνθήκη

$$\sum_j p_j = 1, \quad p_j > 0.$$

Στη (2.14) έχουμε θέσει για απλότητα $w_{ij} \equiv w_i(t_j)$ ενώ $\mathbf{p} \equiv (p_1, \ldots, p_h)$ έτσι ώστε να ισχύει

$$W_i(\mathbf{p}) = \sum_j w_{ij} p_j, \quad i = 1, \ldots, s. \tag{2.15}$$

Αν η μεγιστοποίηση της (2.14) επιτυγχάνεται στο $\widehat{p}$, τότε ο ΜΠΕΜΠ της F είναι

$$\widehat{F}(x) = \sum_{j=1}^{h} \widehat{p}_j I_{(-\infty, x]}(t_j).$$

2.5 Μία εναλλακτική προσέγγιση

Από την Θεωρία μέτρου γνωρίζουμε τα ακόλουθα (δες Shao, 2003):

Έστω $(\Omega, \mathcal{F}, \nu)$ ένας χώρος με μέτρο και f μία μη αρνητική συνάρτηση Borel. Τότε το

$$\lambda(A) = \int_A f d\nu, \quad A \in \mathcal{F} \tag{2.16}$$

είναι ένα μέτρο στο $(\Omega, \mathcal{F})$. Προφανώς, ισχύει

$$\nu(A) = 0 \Rightarrow \lambda(A) = 0. \tag{2.17}$$

Γενικά όταν η (2.17) ισχύει για δύο μέτρα λ και ν, τότε λέμε ότι το λ είναι *απολύτως συνεχές* ως προς το ν και γράφουμε $\lambda \ll \nu$.

Θεώρημα 5. (Radon-Nikodym) Έστω ν και λ δύο μέτρα στον $(\Omega, \mathcal{F})$ όπου ν είναι ένα σ-πεπερασμένο μέτρο. Αν $\lambda \ll \nu$, τότε υπάρχει μία μη αρνητική συνάρτηση Borel f στον Ω τέτοια ώστε να ισχύει η (2.16). Η συνάρτηση f ονομάζεται παράγωγος Radon-Nikodym του λ ως προς το ν και συμβολίζεται με $d\lambda/d\nu$.

2.5.1 Εκτίμηση της συνάρτησης κατανομής

Λήμμα 1. (Davidov and Iliopoulos, 2008) Ας υποθέσουμε ότι έχουμε στην διάθεσή μας s δείγματα, με το i δείγμα να προέρχεται από την κατανομή με συνάρτηση κατανομής

$$F_i(x) = \int_{-\infty}^{x} \frac{w_i(u)}{W_i} dF(u), \tag{2.18}$$

όπου

$$W_i = \int_{-\infty}^{\infty} w_i(u) dF(u).$$

Τότε:

1. Για $s = 1$ και w_1 θετική συνάρτηση σε όλο το στήριγμα της F, έχουμε

$$F(x) = \frac{\int_{-\infty}^{x} (w_1(u))^{-1} dF_1(u)}{\int_{-\infty}^{\infty} (w_1(u))^{-1} dF_1(u)}. \tag{2.19}$$

2. Έστω $s > 1$ και $F^*(x) = \sum_{i=1}^{s} \lambda_i F_i(x)$ ένας οποιοσδήποτε κυρτός συνδυασμός των F_i με $\lambda_i > 0$. Αν το σύνολο $\cup_{i=1}^{s} \{w_i(x) > 0\}$ συμπίπτει με το στήριγμα της F, τότε

$$F(x) = \frac{\int_{-\infty}^{x} \left(\sum_{i=1}^{s} \lambda_i w_i(u)/W_i\right)^{-1} dF^*(u)}{\int_{-\infty}^{\infty} \left(\sum_{i=1}^{s} \lambda_i w_i(u)/W_i\right)^{-1} dF^*(u)}. \tag{2.20}$$

Απόδειξη. 1. Όταν $s = 1$, η (2.18) θα πάρει την μορφή

$$F_1(x) = \int_{-\infty}^{x} \frac{w_1(u)}{W_1} dF(u).$$

Η παράγωγος Radon-Nicodym της F_1 ως προς την F είναι $w_1(x)/W_1$. Επομένως, όταν $w_1(x) > 0$, η παράγωγος Radon-Nicodym της F ως προς την F_1 είναι η $W_1/w_1(x)$. Άρα, με βάση αυτό θα πάρουμε ότι

$$F(x) = \frac{\int_{-\infty}^{x} \frac{1}{w_1(u)} dF_1(u)}{\frac{1}{W_1}}.$$

Όμως έχουμε ότι

$$\int_{-\infty}^{\infty} \frac{1}{w_1(x)} dF_1(x) = \int_{-\infty}^{\infty} \frac{1}{w_1(x)} \frac{w_1(x)}{W_1} dF(x) = \frac{1}{W_1}.$$

Επομένως με αντικατάσταση προκύπτει η σχέση (2.19).

2. Στην περίπτωση που $s > 1$ θα έχουμε

$$F^*(x) = \sum_{i=1}^{s} \lambda_i F_i(x) = \int_{-\infty}^{x} \left(\sum_{i=1}^{s} \lambda_i \frac{w_i(u)}{W_i} \right) dF(u).$$

Ορίζοντας

$$w(x) = \sum_{i=1}^{s} \lambda_i \frac{w_i(x)}{W_i}$$

και θεωρώντας ότι το σύνολο $\cup_{i=1}^{s} \{w_i(x) > 0\}$ συμπίπτει με το στήριγμα της F, εφαρμόζουμε την (2.19) για την περίπτωση όπου $s = 1$ και έτσι θα πάρουμε την (2.20).

□

Βασιζόμενοι στην (2.19) μπορούμε να βρούμε μία εύκολη διαδικασία εκτίμησης στην περίπτωση που $s = 1$. Συγκεκριμένα αν $Y_1, \ldots, Y_n$ είναι ένα τυχαίο δείγμα από την F_i, τότε ο εκτιμητής

$$\frac{1}{n} \sum_{i=1}^{n} I_{(-\infty, x]}(Y_i)$$

εκτιμά την συνάρτηση κατανομής F_i. Τότε λοιπόν ένας άμεσος εκτιμητής για την F είναι ο

$$\widehat{F}(x) = \frac{\sum_{i=1}^{n} (w(y_i))^{-1} I_{(-\infty, x]}(y_i)}{\sum_{i=1}^{n} (w(y_i))^{-1}}.$$

Ο εκτιμητής αυτός είναι ο εκτιμητής του Cox. Φυσικά στην περίπτωση που έχουμε αρκετά δείγματα τα πράγματα είναι πιο περίπλοκα. Μπορούμε να εκτιμήσουμε την F^* συνδυάζοντας τις αντίστοιχες εμπειρικές συναρτήσεις κατανομής και θέτοντας $\lambda_i = n_i/n$ όπου $n = \sum_{i=1}^{s} n_i$. Παρ' όλ' αυτά, σε αντίθεση με την περίπτωση που $s = 1$, η F μπορεί να εκτιμηθεί αν και μόνον αν το γράφημά της είναι συνδεδεμένο. Στην ουσία αυτό σημαίνει ότι για κάθε $1 \leq i \leq s$ υπάρχει ένα $1 \leq j \leq s$ τέτοιο ώστε να ισχύει $P_F[w_i(x) > 0, w_j(x) > 0] > 0$. (Δες και Ενότητα 2.6.)

2.5.2 Ένας μετασχηματισμός της πιθανοφάνειας

Έστω $\mathcal{Y}_1, \ldots, \mathcal{Y}_s$ ανεξάρτητα τυχαία δείγματα όπως στην προηγούμενη ενότητα. Συμβολίζουμε με $z_j,\ j = 1, 2, \ldots, n$, όπου $n = \sum_{j=1}^{s} n_j$ το σύνολο όλων των παρατηρήσεων. Τότε η εμπειρική πιθανοφάνεια δίνεται από την σχέση

$$\prod_{i=1}^{s} \prod_{z_j \in \mathcal{Y}_i} dF_i(z_j) = \prod_{i=1}^{s} \prod_{z_j \in \mathcal{Y}_i} \frac{dF(z_j) w_i(z_j)}{\int_{-\infty}^{\infty} w_i(u) dF(u)}.$$

Όμως ο ΜΠΕΜΠ της F είναι μία διακριτή συνάρτηση κατανομής με άλματα στις παρατηρηθείσες τιμές. Αυτό σημαίνει ότι μπορούμε να αντικαταστήσουμε το $\int_{-\infty}^{\infty} w_i(u) dF(u)$ με το $\sum_{j=1}^{n} p_j w_{ij}$ όπου $p_j = dF(z_j)$ και $w_{ij} = w_i(z_j)$. Άμεσα προκύπτει ότι η εμπειρική πιθανοφάνεια είναι ανάλογη του

$$\prod_{j=1}^{n} p_j \prod_{i=1}^{s} \frac{1}{\left(\sum_{j=1}^{n} p_j w_{ij}\right)^{n_i}}. \tag{2.21}$$

Στόχος μας είναι να μεγιστοποιήσουμε την (2.21) ως προς $\mathbf{p} \in \mathcal{P}$ όπου

$$\mathcal{P} = \left\{ \mathbf{p} : \mathbf{p}^T \mathbf{1}_n = 1,\ \mathbf{p} > 0 \right\}.$$

Με $\mathbf{1}_n \in \Re^n$ συμβολίζεται το διάνυσμα με όλες τις συντεταγμένες ίσες με την μονάδα ενώ $\mathbf{p} > 0$ σημαίνει ότι όλα τα στοιχεία του διανύσματος $\mathbf{p}$ είναι θετικά.
Η (2.21) μπορεί να γραφτεί

$$\begin{aligned}
\prod_{j=1}^{n} p_j \prod_{i=1}^{s} \frac{1}{\left(\sum_{j=1}^{n} p_j w_{ij}\right)^{n_i}} &= \exp\left[\log\left\{\prod_{j=1}^{n} p_j \prod_{i=1}^{s} \frac{1}{\left(\sum_{j=1}^{n} p_j w_{ij}\right)^{n_i}}\right\}\right] \\
&= \exp\left\{\sum_{j=1}^{n} \log(p_j) - \sum_{i=1}^{s} n_i \log\left[\sum_{j=1}^{n} p_j w_{ij}\right]\right\} \\
&= \exp\left\{\sum_{j=1}^{n} \theta_j - \sum_{i=1}^{s} n_i \Omega_i(\boldsymbol{\theta})\right\}, \qquad (2.22)
\end{aligned}$$

όπου $\theta_j = \log(p_j)$ και $\Omega_i(\boldsymbol{\theta}) = \log\left[\sum_{j=1}^{n} w_{ij} \exp(\theta_j)\right]$.
Με βάση την (2.22) μπορούμε να πούμε ότι έχουμε μία εκθετική οικογένεια κατανομών με κανονική παράμετρο την $\boldsymbol{\theta}$ κάτι το οποίο σημαίνει ότι μεγιστοποίηση της (2.21) ισοδυναμεί με μεγιστοποίηση της (2.22).

Λήμμα 2. Ο λογάριθμος της (2.22) είναι μία κοίλη συνάρτηση αλλά όχι αυστηρά κοίλη.

Απόδειξη. Ο λογάριθμος της πιθανοφάνειας δίνεται από την σχέση

$$l(\boldsymbol{\theta}) = \sum_{j=1}^{n} \theta_j - \sum_{i=1}^{s} n_i \Omega_i(\boldsymbol{\theta}). \tag{2.23}$$

Έστω

$$\mathbf{H}_i = \nabla^2_{\boldsymbol{\theta}} \Omega_i(\boldsymbol{\theta}) \quad \text{και} \quad \gamma_i^T = (w_{i1} \exp(\theta_1), \ldots, w_{in} \exp(\theta_n)).$$

Το σύμβολο ∇_θ είναι ο διανυσματικός διαφορικός τελεστής μιας συνάρτησης. Τότε έχουμε τα ακόλουθα:

$$\frac{\partial \Omega_i(\boldsymbol{\theta})}{\partial \theta_\mu} = \frac{w_{i\mu} \exp(\theta_\mu)}{\sum_{j=1}^{n} w_{ij} \exp(\theta_j)},$$

$$\begin{aligned} \frac{\partial^2 \Omega_i(\boldsymbol{\theta})}{\partial \theta_\mu^2} &= \frac{w_{i\mu} \exp(\theta_\mu) \sum_{j=1}^{n} w_{ij} \exp(\theta_j) - \left(w_{i\mu} \exp(\theta_\mu)\right)^2}{\left(\sum_{j=1}^{n} w_{ij} \exp(\theta_j)\right)^2} \\ &= \frac{\gamma_{i\mu}}{W_i} - \frac{\gamma_{i\mu}^2}{W_i^2} \end{aligned} \tag{2.24}$$

και

$$\frac{\partial^2 \Omega_i(\boldsymbol{\theta})}{\partial \theta_\mu \partial \theta_\nu} = -\frac{w_{i\mu} \exp(\theta_\mu) w_{i\nu} \exp(\theta_\nu)}{\left(\sum_{j=1}^{n} w_{ij} \exp(\theta_j)\right)^2} = -\frac{\gamma_{i\mu} \gamma_{i\nu}}{W_i^2}. \tag{2.25}$$

Συνδυάζοντας τις (2.24) και (2.25) θα πάρουμε

$$\nabla^2_{\boldsymbol{\theta}} \Omega_i(\boldsymbol{\theta}) = \begin{cases} \dfrac{\gamma_{i\mu}}{W_i} - \dfrac{\gamma_{i\mu}^2}{W_i^2}, & \mu = \nu \\ \\ -\dfrac{\gamma_{i\mu} \gamma_{i\nu}}{W_i^2}, & \mu \neq \nu. \end{cases}$$

Από αυτό μπορούμε να συμπεράνουμε ότι

$$\nabla^2_{\boldsymbol{\theta}} \Omega_i(\boldsymbol{\theta}) = \frac{1}{W_i} \mathrm{diag}\{\gamma_i\} - \frac{1}{W_i^2} \gamma_i \gamma_i^T,$$

όπου $\mathrm{diag}\{\gamma_i\}$ είναι ο διαγώνιος πίνακας με κύρια διαγώνιο τα στοιχεία του γ_i.

Για κάθε $x \in \Re^n$ έχουμε λοιπόν ότι

$$\begin{aligned}
\mathbf{x}^T\mathbf{H}_i\mathbf{x} &= \mathbf{x}^T\left(\frac{1}{W_i}\mathrm{diag}\left\{\gamma_i\right\} - \frac{1}{W_i^2}\gamma_i\gamma_i^T\right)\mathbf{x} \\
&= \frac{1}{W_i}\mathbf{x}^T\mathrm{diag}\left\{\gamma_i\right\}\mathbf{x} - \frac{1}{W_i^2}\mathbf{x}^T\gamma_i\gamma_i^T\mathbf{x} \\
&= \frac{1}{W_i}\sum_{j=1}^{n}\gamma_{ij}x_j^2 - \frac{1}{W_i^2}\Big(\sum_{j=1}^{n}\gamma_{ij}x_j\Big)^2 \\
&= \frac{1}{W_i^2}\left(W_i\sum_{j=1}^{n}\gamma_{ij}x_j^2 - \Big[\sum_{j=1}^{n}\gamma_{ij}x_j\Big]^2\right) \\
&= \frac{1}{W_i^2}\left(\sum_{j=1}^{n}\gamma_{ij}\sum_{j=1}^{n}\gamma_{ij}x_j^2 - \Big[\sum_{j=1}^{n}\gamma_{ij}x_j\Big]^2\right). \qquad (2.26)
\end{aligned}$$

Από την (2.26) χρησιμοποιώντας την ανισότητα Cauchy-Schwarz βλέπουμε ότι ο $\mathbf{H}_i$ είναι ένας θετικά ημιορισμένος πίνακας. Συνεπώς οι $\Omega_i(\boldsymbol{\theta})$ είναι κυρτές για κάθε $1 \leq i \leq s$ που σημαίνει ότι ο λογάριθμος της πιθανοφάνειας είναι κοίλη συνάρτηση ως προς $\boldsymbol{\theta}$. Παρ' όλ' αυτά αν $\mathbf{x} = c\mathbf{1}$ τότε η (2.26) ισούται με μηδέν, κάτι το οποίο σημαίνει ότι ο πίνακας $\mathbf{H}_i$ είναι μεν θετικά ημιορισμένος αλλά όχι θετικά ορισμένος. Αυτό σημαίνει πως ο λογάριθμος της πιθανοφάνειας δεν είναι αυστηρά κοίλη συνάρτηση. □

2.6 Ύπαρξη και μοναδικότητα του ΜΠΕΜΠ

Για καθαρά λόγους ευκολίας θεωρούμε αρχικά την περίπτωση που έχουμε στην διάθεσή μας $s = 2$ δείγματα, $\mathcal{Y}_1$ και $\mathcal{Y}_2$. Τότε ο λογάριθμος της πιθανοφάνειας στη (2.23) γράφεται ως εξής:

$$\begin{aligned}
l(\mathbf{p}) &= \sum_{j=1}^{n}\log(p_j) - \sum_{i=1}^{2}n_i\log\left(\sum_{j=1}^{n}p_jw_{ij}\right) \\
&= \sum_{j=1}^{n}\log(p_j) - n_1\log\left(\sum_{j=1}^{n}p_jw_{1j}\right) - n_2\log\left(\sum_{j=1}^{n}p_jw_{2j}\right). \qquad (2.27)
\end{aligned}$$

Έστω τώρα $\mathbf{p} = (\mathbf{p}_1, \mathbf{p}_2)$ με $\mathbf{p}_1, \mathbf{p}_2$ τα υποδιανύσματα των πιθανοτήτων τα οποία σχετίζονται με το $\mathcal{Y}_1$ και $\mathcal{Y}_2$ αντίστοιχα. Έστω επίσης $\mathbf{w}_1 = (\mathbf{w}_{11}, \mathbf{w}_{12})$ όπου $\mathbf{w}_{11}$

είναι τα βάρη για το δείγμα $\mathcal{Y}_1$ και $\mathbf{w}_{12}$ είναι τα βάρη για το δείγμα $\mathcal{Y}_2$ υπολογισμένα όμως με την συνάρτηση $\mathbf{w}_1(x)$. Με παρόμοιο τρόπο ορίζουμε το $\mathbf{w}_2 = (\mathbf{w}_{21}, \mathbf{w}_{22})$. Τότε ο λογάριθμος της συνάρτησης πιθανοφάνειας ισούται με το άθροισμα $l_1(\mathbf{p})+l_2(\mathbf{p})$ όπου

$$l_1(\mathbf{p}) = \sum_{j=1}^{n_1} \log(p_{1j}) - n_1 \log\left(\sum_{j=1}^{n_1} p_{1j} w_{11j} + \sum_{j=1}^{n_2} p_{2j} w_{12j}\right) \tag{2.28}$$

και

$$l_2(\mathbf{p}) = \sum_{j=1}^{n_2} \log(p_{2j}) - n_2 \log\left(\sum_{j=1}^{n_1} p_{1j} w_{21j} + \sum_{j=1}^{n_2} p_{2j} w_{22j}\right). \tag{2.29}$$

Αν τώρα $\mathbf{w}_{12} = 0$, τότε $l_1(\mathbf{p}) = l_1(\mathbf{p}_1)$, ενώ η συνάρτηση πιθανοφάνειας ισούται με

$$\sum_{j=1}^{n_1} \log(p_{1j}) - n_1 \log\left(\sum_{j=1}^{n_1} p_{1j} w_{11j}\right). \tag{2.30}$$

Θεώρημα 6. Θεωρούμε τις ακόλουθες ισότητες:

$$l_1(\mathbf{p}) = l_1(\mathbf{p}_1) \quad \text{και} \quad l_2(\mathbf{p}) = l_2(\mathbf{p}_2)$$

1. Αν και οι δύο παραπάνω ισότητες ικανοποιούνται τότε ο ΜΠΕΜΠ της F υπάρχει αλλά δεν είναι μοναδικός.
2. Αν μόνο μία εκ των δύο ικανοποιείται τότε ο ΜΠΕΜΠ της F δεν υπάρχει.
3. Αν καμμία δεν ικανοποιείται τότε ο ΜΠΕΜΠ της F υπάρχει και είναι μοναδικός.

Απόδειξη. Στην πρώτη περίπτωση που και οι δύο παραπάνω ισότητες ικανοποιούνται τότε $l(\mathbf{p}) = l_1(\mathbf{p}_1) + l_2(\mathbf{p}_2)$. Στην περίπτωση αυτή λοιπόν θα έχουμε ότι το $\max\{l(\mathbf{p}) : \mathbf{p} \in \mathcal{P}\}$ θα ισούται με

$$\max_{\alpha \in (0,1)} \left\{\max\{l_1(\mathbf{p}_1) : \mathbf{p}_1 \in \mathcal{P}_{1,\alpha}\} + \max\{l_2(\mathbf{p}_2) : \mathbf{p}_2 \in \mathcal{P}_{2,1-\alpha}\}\right\}, \tag{2.31}$$

όπου $\mathcal{P}_{1,\alpha} = \left\{\mathbf{p}_1 : \mathbf{p}_1^T \mathbf{1}_{n_1} = \alpha,\ \mathbf{p}_1 > 0\right\}$ ενώ $\mathcal{P}_{2,1-\alpha} = \left\{\mathbf{p}_2 : \mathbf{p}_2^T \mathbf{1}_{n_2} = 1-\alpha,\ \mathbf{p}_2 > 0\right\}$. Για κάθε $\alpha \in (0,1)$ και με χρήση των πολλαπλασιαστών Lagrange θα πάρουμε τα ακόλουθα:

$$\widehat{p}_{1j} = \frac{\alpha}{w_{11j}} \left(\sum_{j=1}^{n_1} \frac{1}{w_{11j}} \right)^{-1}, \quad j = 1, \ldots, n_1 \tag{2.32}$$

$$\widehat{p}_{2j} = \frac{1-\alpha}{w_{22j}} \left(\sum_{j=1}^{n_1} \frac{1}{w_{22j}} \right)^{-1}, \quad j = 1, \ldots, n_2. \tag{2.33}$$

(Για την απόδειξη αυτών δες στην Ενότητα 2.6.1.)
Από τις (2.32) και (2.33) παίρνουμε ότι

$$\max \{l_1(\mathbf{p}_1) : \mathbf{p}_1 \in \mathcal{P}_{1,\alpha}\} = -\sum_{j=1}^{n_1} \log(w_{11j}) - n_1 \log(n_1)$$

$$\max \{l_2(\mathbf{p}_2) : \mathbf{p}_2 \in \mathcal{P}_{1,\alpha}\} = -\sum_{j=1}^{n_2} \log(w_{22j}) - n_2 \log(n_2).$$

Άρα το $\max \{l_1(\mathbf{p}_1) + l_2(\mathbf{p}_2) : \mathbf{p}_1 \in \mathcal{P}_{1,\alpha} \textit{ και } \mathbf{p}_2 \in \mathcal{P}_{2,1-\alpha}\}$ είναι ανεξάρτητο του α, κάτι το οποίο σημαίνει ότι ο εκτιμητής δεν είναι μοναδικός.

Έστω τώρα $l_1(\mathbf{p}) \neq l_1(\mathbf{p}_1)$ αλλά $l_2(\mathbf{p}) = l_2(\mathbf{p}_2)$. Τότε το $\max \{l(\mathbf{p}) : \mathbf{p} \in \mathcal{P}\}$ θα ισούται με

$$\max \{l_1(\mathbf{p}) + l_2(\mathbf{p}_2) : \mathbf{p}_1 \in \mathcal{P}_{1,\alpha} \textit{ και } \mathbf{p}_2 \in \mathcal{P}_{2,1-\alpha}\}.$$

Όμως όπως είδαμε το $\max \{l_2(\mathbf{p}_2) : \mathbf{p}_2 \in \mathcal{P}_{2,1-\alpha}\}$ είναι ανεξάρτητο από το α. Επίσης το $l_1(\mathbf{p})$ δίνεται από την σχέση (2.28), η οποία είναι μία φθίνουσα συνάρτηση ως προς οποιοδήποτε διάνυσμα $\mathbf{p}_2$. Έτσι η (2.28) αυξάνει καθώς το $\mathbf{p}_2 \to 0$ ή αντίστοιχα όταν το $\alpha \to 1$. Αυτό λοιπόν σημαίνει ότι η μεγιστοποίηση θα γίνει σε άκρο του συνόλου $\mathcal{P}$, δηλαδή εκτός παραμετρικού χώρου, κάτι το οποίο σημαίνει ότι ο ΜΠΕΜΠ δεν υπάρχει.

Τέλος έστω ότι καμμία συνθήκη δεν ικανοποιείται, δηλαδή ισχύει $l_1(\mathbf{p}) \neq l_1(\mathbf{p}_1)$ και $l_2(\mathbf{p}) \neq l_2(\mathbf{p}_2)$. Η πιθανοφάνεια που δίνεται από την (2.21), έστω $L(\mathbf{p})$, τείνει στο μηδέν αν κάποιο από τα $\mathbf{p}_j \to 0$. Επομένως $L(\mathbf{p}) = 0 \ \forall \mathbf{p} \in \overline{\mathcal{P}} \backslash \mathcal{P}$ όπου $\overline{\mathcal{P}}$ είναι το περίβλημα του $\mathcal{P}$. Όμως η $L(\mathbf{p})$ είναι συνεχής στο $\overline{\mathcal{P}}$ και άρα θα έχει ένα μέγιστο

στο σύνολο αυτό. Όμως αφού $L(\mathbf{p}) > 0\ \forall \mathbf{p} \in \mathcal{P}$, το μέγιστο θα είναι στο $\mathcal{P}$ και όχι στο $\overline{\mathcal{P}} \backslash \mathcal{P}$ και αυτό σημαίνει ότι ο ΜΠΕΜΠ υπάρχει. Θα αποδείξουμε τώρα ότι το μέγιστο αυτό είναι και μοναδικό.

Έστω $\mathbf{p}$ *και* $\mathbf{p}'$ δύο μέγιστα και επίσης θ *και* θ' οι λογάριθμοι αυτών. Τότε τα θ *και* θ' μεγιστοποιούν την (2.23). Σύμφωνα με το Λήμμα 2 όμως, η (2.23) είναι κοίλη, κάτι το οποίο σημαίνει ότι για κάθε $0 \leq \lambda \leq 1$ η ποσότητα $\lambda\theta + (1-\lambda)\theta'$ είναι επίσης ένα μέγιστο για την (2.23) αλλά και για την (2.27). Άρα λοιπόν προκύπτει ότι η $l\left(\lambda\theta + (1-\lambda)\,\theta'\right)$ είναι σταθερά ως συνάρτηση του $\lambda \in [0,1]$ και επομένως:

$$\frac{d^k}{d\lambda^k}\left(l\left(\lambda\theta + (1-\lambda)\,\theta'\right)\right) = 0, \quad \forall k \in \mathcal{N} \text{ και } \lambda \in [0,1]$$

Συγκεκριμένα για $k = 2$ και $\lambda = 1$ θα πάρουμε

$$n_1\Delta_1 + n_2\Delta_2 = 0, \tag{2.34}$$

όπου

$$\Delta_i = \frac{\left\{\sum_{j=1}^n w_{ij}(\theta_j - \theta'_j)^2 e^{\theta_j}\right\}\sum_{j=1}^n w_{ij}e^{\theta_j} - \left\{\sum_{j=1}^n w_{ij}(\theta_j - \theta'_j)e^{\theta_j}\right\}^2}{\left(\sum_{j=1}^n w_{ij}e^{\theta_j}\right)^2}. \tag{2.35}$$

Εφαρμόζοντας την ανισότητα Cauchy-Schwarz στην (2.35) θα πάρουμε ότι $\Delta_1, \Delta_2 \geq 0$ το οποίο ταυτόχρονα με την (2.34) θα μας δώσει ότι $\Delta_1 = \Delta_2 = 0$. Η ισότητα στην ανισότητα Cauchy-Schwarz όμως ισχύει μόνο όταν $\theta_j - \theta'_j = c$ ή ισοδύναμα όταν $p_j = \exp(c)p'_j$ για κάποια σταθερά $c \in \Re$. Αθροίζοντας ως προς j έχουμε

$$\begin{aligned}
&p_j = \exp(c)p'_j \Leftrightarrow \\
&\sum_j p_j = \sum_j \exp(c)p'_j \Leftrightarrow \\
&1 = \exp(c) \Leftrightarrow \\
&c = 0,
\end{aligned}$$

κάτι το οποίο σημαίνει ότι $p_j = p'_j$, $\forall j$. Άρα ο ΜΠΕΜΠ είναι μοναδικός. □

Παρατήρηση: Από το Θεώρημα 6 μπορούμε να δούμε ότι:

1. Αν $\mathbf{w}_1^T \mathbf{w}_2 = 0$, τότε ο ΜΠΕΜΠ δεν είναι μοναδικός. Στην περίπτωση αυτή φυσικά ισχύει ότι $\mathbf{w}_{12} = 0$ και $\mathbf{w}_{21} = 0$.

2. Αν ακριβώς ένα εκ των $\mathbf{w}_{12}$, $\mathbf{w}_{21}$ είναι μηδενικά τότε ο ΜΠΕΜΠ δεν υπάρχει.

3. Αν και το $\mathbf{w}_{12}$ αλλά και το $\mathbf{w}_{21}$ είναι μη μηδενικά τότε ο ΜΠΕΜΠ είναι μοναδικός.

2.6.1 Απόδειξη των σχέσεων (2.32, 2.33)

Το αποτέλεσμα της σχέσης (2.32) βασίζεται στην μεγιστοποίηση της πιθανοφάνειας (2.30) υπό τη συνθήκη $\sum_{j=1}^{n_1} p_{1j} = \alpha$. Χρησιμοποιώντας πολλαπλασιαστές Lagrange ορίζουμε την συνάρτηση

$$\begin{aligned} g\left(p_{1j}\right) &= l_1\left(\mathbf{p}_1\right) - \lambda\left(\sum_{j=1}^{n_1} p_{1j} - \alpha\right) \\ &= \sum_{j=1}^{n_1} \log(p_{1j}) - n_1 \log\left\{\sum_{j=1}^{n_1} p_{1j} w_{11j}\right\} - \lambda\left(\sum_{j=1}^{n_1} p_{1j} - \alpha\right). \end{aligned}$$

Παραγωγίζουμε ως προς p_{1j} και παίρνουμε

$$\begin{aligned} \frac{\partial g\left(p_{1j}\right)}{\partial p_{1j}} &= \frac{1}{p_{1j}} - n_1 \frac{w_{11j}}{\sum_{j=1}^{n_1} p_{1j} w_{11j}} - \lambda = 0 \\ &\Rightarrow 1 - n_1 \frac{w_{11j} p_{1j}}{\sum_{j=1}^{n_1} p_{1j} w_{11j}} - \lambda p_{1j} = 0 \\ &\Rightarrow \sum_{j=1}^{n_1} 1 - \sum_{j=1}^{n_1} n_1 \frac{w_{11j} p_{1j}}{\sum_{j=1}^{n_1} p_{1j} w_{11j}} - \sum_{j=1}^{n_1} \lambda p_{1j} = 0 \\ &\Rightarrow n_1 - \frac{n_1}{\sum_{j=1}^{n_1} p_{1j} w_{11j}} \sum_{j=1}^{n_1} p_{1j} w_{11j} - \lambda = 0 \\ &\Rightarrow n_1 - n_1 - \lambda = 0 \\ &\Rightarrow \lambda = 0. \end{aligned}$$

Άρα λοιπόν

$$\frac{1}{p_{1j}} - n_1 \frac{w_{11j}}{\sum_{j=1}^{n_1} p_{1j} w_{11j}} = 0 \Rightarrow \sum_{j=1}^{n_1} p_{1j} w_{11j} = n_1 \widehat{p}_{1j} w_{11j}, \tag{2.36}$$

όπου από αυτήν προκύπτει ότι

$$\widehat{p}_{1j} = \frac{\sum_{j=1}^{n_1} p_{1j} w_{11j}}{n_1 w_{11j}}. \tag{2.37}$$

Αθροίζοντας ως προς j την τελευταία σχέση θα πάρουμε

$$\alpha = \sum_{j=1}^{n} \widehat{p}_{1j} = \sum_{j=1}^{n} \frac{\sum_{j=1}^{n_1} p_{1j} w_{11j}}{n_1 w_{11j}} \Rightarrow \alpha = \frac{\sum_{j=1}^{n_1} p_{1j} w_{11j}}{n_1} \sum_{j=1}^{n_1} \frac{1}{w_{11j}}.$$

Χρησιμοποιώντας όμως την (2.36) παίρνουμε

$$\alpha = \frac{n_1 \widehat{p}_{1j} w_{11j}}{n_1} \sum_{j=1}^{n_1} \frac{1}{w_{11j}} \Rightarrow \alpha = \widehat{p}_{1j} w_{11j} \sum_{j=1}^{n_1} \frac{1}{w_{11j}},$$

από όπου προκύπτει ότι

$$\widehat{p}_{1j} = \frac{\alpha}{w_{11j}} \left(\sum_{j=1}^{n_1} \frac{1}{w_{11j}} \right)^{-1}.$$

Παρόμοια αποδεικνύεται και η (2.33).

2.6.2 Παραδείγματα

Παράδειγμα 1. (Μη ύπαρξη ΜΠΕΜΠ) Έστω ότι έχουμε στην διάθεσή μας $s = 2$ δείγματα, $y_1 = \{13, 15\}$ και $y_2 = \{5, 8\}$ για τα οποία ξέρουμε ότι οι συναρτήσεις βάρη δίνονται από τις σχέσεις

$$w_1(z) = I_{[10,20]}(z) \quad \text{και} \quad w_2(z) = 1.$$

Στην ουσία αυτό σημαίνει ότι εμείς έχουμε ένα δείγμα μεγέθους $n_1 = 2$ από την περικεκομμένη κατανομή F στο διάστημα $[10, 20]$ και οι παρατηρηθείσες τιμές είναι οι $13, 15$ και επίσης έχουμε άλλο ένα δείγμα πάλι μεγέθους $n_2 = 2$ αλλά προέρχεται

από την κατανομή F και έχει τις τιμές $5, 8$. Με βάση αυτές τις τιμές μπορούμε να δούμε εύκολα ότι η πιθανοφάνεια θα δίνεται από τον τύπο:

$$L(\mathbf{p}) = p_1 p_2 \frac{p_3 p_4}{(p_3 + p_4)^2},$$

όπου $\mathbf{p} = (p_1, p_2, p_3, p_4)$. Να παρατηρήσουμε απλά ότι η σχέση αυτή προέκυψε σύμφωνα με την (2.15) όπου $W_1(\mathbf{p}) = W_2(\mathbf{p}) = 1$ και $W_3(\mathbf{p}) = W_4(\mathbf{p}) = p_3 + p_4$. Η παραπάνω όμως συνάρτηση πιθανοφάνειας δεν μεγιστοποιείται, κάτι το οποίο σημαίνει ότι ο ΜΠΕΜΠ δεν υπάρχει. Για να το δούμε αυτό παίρνουμε αρχικά τον λογάριθμο της συνάρτησης πιθανοφάνειας, θέτοντας $p_4 = 1 - p_1 - p_2 - p_3$.

$$L(\mathbf{p}) = p_1 p_2 \frac{p_3(1 - p_1 - p_2 - p_3)}{(p_3 + [1 - p_1 - p_2 - p_3)]^2} = p_1 p_2 \frac{p_3(1 - p_1 - p_2 - p_3)}{(1 - p_1 - p_2)^2} \Leftrightarrow$$
$$l(\mathbf{p}) = \log(p_1) + \log(p_2) + \log(p_3) + \log(1 - p_1 - p_2 - p_3) - 2\log(1 - p_1 - p_2).$$

Παραγωγίζοντας ως προς p_3 θα πάρουμε

$$\frac{\partial l(\mathbf{p})}{\partial p_3} = \frac{1}{p_3} - \frac{1}{1 - p_1 - p_2 - p_3} = 0 \Leftrightarrow$$
$$1 - p_1 - p_2 - p_3 - p_3 = 0 \Leftrightarrow$$
$$p_3 = \frac{1 - p_1 - p_2}{2}.$$

Παραγωγίζοντας όμως ως προς p_1 και αντικαθιστώντας το p_3 προκύπτει

$$\begin{aligned}
\frac{\partial l(\mathbf{p})}{\partial p_1} &= \frac{1}{p_1} - \frac{1}{1 - p_1 - p_2 - p_3} + \frac{2}{1 - p_1 - p_2} \\
&= \frac{1}{p_1} - \frac{1}{1 - p_1 - p_2 - \frac{1 - p_1 - p_2}{2}} + \frac{2}{1 - p_1 - p_2} \\
&= \frac{1}{p_1} - \frac{2}{1 - p_1 - p_2} + \frac{2}{1 - p_1 - p_2} \\
&= \frac{1}{p_1} \neq 0.
\end{aligned}$$

Άρα ο ΜΠΕΜΠ δεν υπάρχει.

Παράδειγμα 2. (Μη μοναδικότητα ΜΠΕΜΠ) Θεωρούμε τα δεδομένα του Παραδείγματος 1 με μόνη διαφορά το βάρος που έχει το δεύτερο δείγμα. Έστω ότι

$$w_2(z) = I_{[3,9]}(z).$$

Στην περίπτωση αυτή καθένα από τα $W_i(\mathbf{p})$ εξαρτάται μόνο από το δείγμα i. Συγκεκριμένα εδώ θα ισχύει ότι $W_1(\mathbf{p}) = W_2(\mathbf{p}) = p_1 + p_2$ και $W_3(\mathbf{p}) = W_4(\mathbf{p}) = p_3 + p_4$. Επομένως η συνάρτηση πιθανοφάνειας είναι η

$$L(\mathbf{p}) = \frac{p_1 p_2}{(p_1+p_2)^2} \frac{p_3 p_4}{(p_3+p_4)^2}.$$

Όμως αν θέσουμε $p_1 + p_2 = \alpha$ και $p_3 + p_4 = 1 - \alpha$ τότε

$$L(\mathbf{p}) = \frac{p_1(\alpha - p_1)}{\alpha^2} \frac{p_3(1-\alpha-p_3)}{(1-\alpha)^2}.$$

Η συνάρτηση αυτή όμως μεγιστοποιείται για οποιοδήποτε $\widehat{\mathbf{p}}$ της μορφής

$$\widehat{\mathbf{p}} = \left(\frac{\alpha}{2}, \frac{\alpha}{2}, \frac{1-\alpha}{2}, \frac{1-\alpha}{2}\right),$$

όπου $\alpha \in (0,1)$. Συγκεκριμένα με αντικατάσταση προκύπτει ότι

$$L(\mathbf{p}) = \frac{\frac{\alpha}{2}\frac{\alpha}{2}}{\alpha^2} \frac{\frac{1-\alpha}{2}\frac{1-\alpha}{2}}{(1-\alpha)^2} = \frac{1}{16},$$

που σημαίνει ότι το αποτέλεσμα είναι ανεξάρτητο του α. Άρα λοιπόν ο ΜΠΕΜΠ δεν είναι μοναδικός.

Παράδειγμα 3 (Μοναδικότητα ΜΠΕΜΠ)**.** Αν στο Παράδειγμα 1 αλλάξουμε την πρώτη τιμή του δεύτερου δείγματος από 5 σε 12, τότε η πιθανοφάνεια δεν θα έχει τον ίδιο τύπο όπως προηγουμένως αλλά θα δίνεται από την σχέση:

$$L(\mathbf{p}) = p_1 p_2 \frac{p_3 p_4}{(p_2+p_3+p_4)^2}.$$

Τώρα έχουμε $W_1(\mathbf{p}) = W_2(\mathbf{p}) = p_2 + p_3 + p_4$ και $W_3(\mathbf{p}) = W_4(\mathbf{p}) = 1$. Θέτοντας $p_1 = 1 - p_2 - p_3 - p_4$ προκύπτει

$$L(\mathbf{p}) = p_1 p_2 \frac{p_3 p_4}{(p_2+p_3+p_4)^2} = (1-p_2-p_3-p_4) p_2 \frac{p_3 p_4}{(p_2+p_3+p_4)^2} \Leftrightarrow$$
$$l(\mathbf{p}) = \log(1-p_2-p_3-p_4) + \log(p_2) + \log(p_3) + \log(p_4) - 2\log(p_2+p_3+p_4).$$

Παραγωγίζοντας ως προς p_2 θα πάρουμε

$$\begin{aligned}
&\frac{\partial l(\mathbf{p})}{\partial p_2} = -\frac{1}{1-p_2-p_3-p_4} + \frac{1}{p_2} - \frac{2}{p_2+p_3+p_4} = 0 \Leftrightarrow \\
&\frac{1}{p_2} = \frac{2}{p_2+p_3+p_4} + \frac{1}{1-p_2-p_3-p_4} \Leftrightarrow \\
&\frac{1}{p_2} = \frac{2(1-p_2-p_3-p_4)+p_2+p_3+p_4}{(p_2+p_3+p_4)(1-p_2-p_3-p_4)} \Leftrightarrow \\
&\frac{1}{p_2} = \frac{2-p_2-p_3-p_4}{(p_2+p_3+p_4)(1-p_2-p_3-p_4)} \Leftrightarrow \\
&p_2 = \frac{(p_2+p_3+p_4)(1-p_2-p_3-p_4)}{2-p_2-p_3-p_4}.
\end{aligned}$$

Παραγωγίζοντας ως προς p_3 και p_4 αντίστοιχα βλέπουμε ότι $p_2 = p_3 = p_4$. Άρα με αντικατάσταση στην τελευταία σχέση προκύπτει

$$\begin{aligned}
p_2 &= \frac{(p_2+p_3+p_4)(1-p_2-p_3-p_4)}{2-p_2-p_3-p_4} \Leftrightarrow \\
p_2 &= \frac{(p_2+p_2+p_2)(1-p_2-p_2-p_2)}{2-p_2-p_2-p_2} \Leftrightarrow \\
p_2 &= \frac{(3p_2)(1-3p_2)}{2-3p_2} \Leftrightarrow \\
2-3p_2 &= 3(1-3p_2) \Leftrightarrow \\
p_2 &= \frac{1}{6}.
\end{aligned}$$

Έτσι λοιπόν η λύση είναι η $(\widehat{p}_1, \widehat{p}_2, \widehat{p}_3, \widehat{p}_4) = (1/2, 1/6, 1/6, 1/6)$. Στην πραγματικότητα αυτό που συνέβη εδώ και πήραμε τον μοναδικό ΜΠΕΜΠ είναι ότι συνδυάσαμε τα δύο δείγματα μια και αυτά επικαλύπτονται.

2.6.3 Η περίπτωση πολλών δειγμάτων

Όλες οι συνθήκες οι οποίες περιγράφηκαν στο Θεώρημα 6 μπορούν να γενικευτούν και στην περίπτωση που έχουμε στη διάθεσή μας περισσότερα από δύο δείγματα. Για τον λόγο αυτό θεωρούμε κατάλληλο υποσύνολο J από τα $\{1, 2, \ldots, s\}$. Τότε ο λογάριθμος της πιθανοφάνειας θα πάρει την μορφή $l(\mathbf{p}) = l_J(\mathbf{p}) + l_{J^C}(\mathbf{p})$ όπου

$$l_J(\mathbf{p}) = \log\left\{\prod_{i\in J}\prod_{z_j\in\mathcal{Y}_i} dF_i(z_j)\right\}$$

είναι το κομμάτι του λογαρίθμου της πιθανοφάνειας που σχετίζεται με το σύνολο $\cup_{i\in J}\mathcal{Y}_i$. Αντίστοιχα ορίζεται και ο λογάριθμος της πιθανοφάνειας $l_{J^C}(\mathbf{p})$. Ορίζουμε τώρα το διάνυσμα $\mathbf{p} = \{\mathbf{p}_J, \mathbf{p}_{J^C}\}$ το οποίο είναι μία διαίρεση του διανύσματος $\mathbf{p}$ όπου το $\mathbf{p}_J$ είναι το υποδιάνυσμα που σχετίζεται με την ένωση $\cup_{i\in J}\mathcal{Y}_i$. Με τον ίδιο τρόπο ορίζεται το υποδιάνυσμα $\mathbf{p}_{J^C}$. Άρα στην ουσία έχουμε ανάλογα πράγματα όπως και στην περίπτωση που το $s = 2$. Συγκεκριμένα:

1. Αν για κάποιο J έχουμε ότι $l_J(\mathbf{p}) = l_J(\mathbf{p}_J)$ και $l_{J^C}(\mathbf{p}) = l_{J^C}(\mathbf{p}_{J^C})$ τότε ο ΜΠΕΜΠ υπάρχει αλλά δεν είναι μοναδικός.

2. Αν μόνο μία εκ των δύο παραπάνω ισοτήτων ισχύει τότε ο ΜΠΕΜΠ δεν υπάρχει.

3. Τέλος αν για κάθε J έχουμε ότι $l_J(\mathbf{p}) \neq l_J(\mathbf{p}_J)$ τότε ο ΜΠΕΜΠ υπάρχει και είναι μοναδικός.

Κεφάλαιο 3

Παρουσίαση αλγορίθμου και παραδείγματα

Στο κεφάλαιο αυτό θα παρουσιαστούν αλγόριθμοι για τον υπολογισμό του ΜΠΕΜΠ με βάση $s \geq 2$ ανεξάρτητα σταθμισμένα τυχαία δείγματα. Θα γίνουν συγκρίσεις χρησιμοποιώντας προσομοιωμένα δείγματα και κατά δεύτερον θα εκτιμηθούν οι συναρτήσεις κατανομής από τις οποίες προέρχονται αυτά.

3.1 Ο επαναληπτικός αλγόριθμος του Vardi

Συνοψίζοντας τα αποτελέσματα του προηγούμενου κεφαλαίου, είναι φανερό ότι όταν διαθέτουμε s το πλήθος σταθμισμένα δείγματα μεγέθους n_i από κατανομή F_i, τότε για την πιθανοφάνεια $L(F)$ ισχύει

$$L(F) \propto \prod_{j=1}^{n} p_j \prod_{i=1}^{s} \frac{1}{\left(\sum_{j=1}^{n} p_j w_{ij}\right)^{n_i}},$$

όπου $n = \sum_{i=1}^{s} n_i$ και $\mathbf{p} = dF(t_i) \in \mathcal{P} = \{\mathbf{p} : \mathbf{p}^T \mathbf{1}_n = 1,\ \mathbf{p} > 0\}$. Στην περίπτωση που ο ΜΠΕΜΠ υπάρχει και είναι μοναδικός, η μεγιστοποίηση της πιθανοφάνειας θα γίνει ως προς $\mathbf{p}$ για

$$\widehat{p}_j = \left(\sum_{i=1}^{s} n_i \frac{w_{ij}}{\widehat{W}_i}\right)^{-1}, \quad j = 1, 2, \ldots, n, \tag{3.1}$$

όπου $\widehat{W}_i = \sum_{j=1}^{n} \widehat{p}_j w_{ij}$ για $i = 1, 2, \ldots, s$. Φυσικά όμως οι τιμές του διανύσματος $\mathbf{W} = (W_1, \ldots, W_s)$ δεν είναι γνωστές με αποτέλεσμα μην είναι εφικτή η εκτίμηση των πιθανοτήτων από την (3.1). Συνδυάζοντας όμως τη σχέση αυτή, με την προϋπόθεση $\widehat{W}_i = \sum_{j=1}^{n} \widehat{p}_j w_{ij}$, θα πάρουμε ένα σύστημα από s μη γραμμικές εξισώσεις με s αγνώστους

$$\sum_{j=1}^{n} \sum_{i=1}^{s} n_i \frac{w_{ij}}{\widehat{W}_i} \begin{pmatrix} n_1 w_{1j} \\ \vdots \\ n_K w_{sj} \end{pmatrix} = \begin{pmatrix} \widehat{W}_1 \\ \vdots \\ \widehat{W}_s \end{pmatrix} \tag{3.2}$$

Λύνοντας τις εξίσωσεις (3.2) και αντικαθιστώντας τις τιμές του $\widehat{\mathbf{W}}$ στην (3.1), φτάνουμε στην εκτίμηση του ΜΠΕΜΠ, όπως αρχικά προτάθηκε από τον Vardi (1982, 1985). Ο Vardi χρησιμοποιεί μία ελαφρώς διαφορετική παραμετροποίηση. Αρχικά θέτει $q_j = p_j w_{sj}$ και μεγιστοποιεί την πιθανοφάνεια όταν $W_s = \sum_{j=1}^{n} p_j w_{sj} = 1$. Επισης ορίζοντας $V_i = W_i / W_s$ για $i = 1, 2, \ldots, s-1$, λύνεται το σύστημα. Φυσικά η λύση της (3.2) δεν είναι άμεση καθώς έχουμε σύστημα μη γραμμικών εξισώσεων. Για το λόγο αυτό, ο Vardi πρότεινε την επίλυση των $s-1$ εξισώσεων με αριθμητικές μεθόδους. Συγκεκριμένα, δίνονται κάποιες αρχικές τιμές στο διάνυσμα $\mathbf{V}$, έστω $\left(V_1^{(0)}, \ldots, V_{s-1}^{(0)}\right)$ αυτές και έτσι λύνεται η πρώτη εξίσωση ως προς V_1 κρατώντας όλες τις άλλες τιμές σταθερές. Η λύση $V_1^{(1)}$ αντικαθιστά την τιμή $V_1^{(0)}$ και λύνεται η δεύτερη εξίσωση ως προς V_2. Αντίστοιχα η λύση $V_2^{(1)}$ αντικαθιστά την τιμή $V_2^{(0)}$ κρατώντας όλα τα άλλα σταθερά και αυτή η διαδικασία συνεχίζεται για $s-1$ βήματα όπου και το τελικό διανυσμα θα έχει γίνει $\left(V_1^{(1)}, \ldots, V_{s-1}^{(1)}\right)$ και ένας κύκλος θα έχει ολοκληρωθεί. Η διαδικασία αυτή επαναλαμβάνεται μέχρι το διάνυσμα $\mathbf{V}$ να ικανοποιεί ένα προκαθορισμένο κριτήριο σύγκλισης. Ο Vardi πρότεινε την χρήση της μεθόδου της διχοτόμησης αλλά και άλλες αριθμητικές μέθοδοι μπορούν να δώσουν λύση.

3.2 Ο επαναληπτικός αλγόριθμος των Davidov and Iliopoulos

Θεωρούμε την ακόλουθη επαναληπτική διαδικασία:
Δίνουμε κάποιες αρχικές τιμές στο διάνυσμα $\mathbf{W}$ έστω $W^{(0)} = \left(W_1^{(0)}, \ldots, \mathbf{W}_s^{(0)}\right)$.
Τότε υπολογίζονται οι πιθανότητες σύμφωνα με την (3.1) ως εξής:

$$\widehat{pr}_j^{(m)} = \left(\sum_{i=1}^{s} n_i \frac{w_{ij}}{\widehat{W}_i}\right)^{-1}, \; j = 1, 2, \ldots, n,$$

ενώ οι εκτιμηθείσες πιθανότητες θα δίνονται όπως φαίνονται παρακάτω:

$$\widehat{p}_j^{(m)} = \frac{\widehat{pr}_j^{(m)}}{\sum_{j=1}^{n} \widehat{pr}_j^{(m)}}, \quad j = 1, 2, \ldots, n.$$

Στο επόμενο βήμα εκτιμώνται εκ νέου οι ποσότητες

$$\widehat{W}_i^{(m+1)} = \sum_{j=1}^{n} \widehat{p}_j^{(m)} w_{ij}, \quad i = 1, 2, \ldots, s.$$

Ο αλγόριθμος σταματάει όταν οι πιθανότητες συγκλίνουν στο σημείο που μεγιστοποιεί την πιθανοφάνεια και αυτό θα θεωρηθεί ότι έχει γίνει όταν δεν θα υπάρχει ουσιαστική μεταβολή από το ένα βήμα στο επόμενο. Συγκεκριμένα η εκτίμηση για τις πιθανότητες θα είναι η $\widehat{p} = \widehat{p}^{(m)}$ όταν

$$\left\|\widehat{p}^{(m)} - \widehat{p}^{(m-1)}\right\| \leq \varepsilon \quad \text{ή} \quad \left\|\widehat{W}^{(m)} - \widehat{W}^{(m-1)}\right\| \leq \varepsilon,$$

όπου το ε είναι προκαθορισμένο και αρκετά μικρό. Φυσικά $\varepsilon > 0$ ενώ με $\|\cdot\|$ συμβολίζεται κάποια προκαθορισμένη νόρμα στον $\Re^n$. Ο συγκεκριμένος αλγόριθμος δεν χρειάζεται την επίλυση κάποιας εξίσωσης ώστε να δώσει αποτέλεσμα, κάτι το οποίο τον κάνει αρκετά πιο γρήγορο σε σχέση με τον αλγόριθμο που πρότεινε ο Vardi. Συγκεκριμένα οι Davidov and Iliopoulos (2010) μέσω προσομοιώσεων έδειξαν ότι ο απαιτούμενος χρόνος μέχρι την σύγκλιση είναι έως και 95% μικρότερος από αυτόν του αλγορίθμου του Vardi.

Θεώρημα 7. Αν ο ΜΠΕΜΠ υπάρχει και είναι μοναδικός τότε για οποιαδήποτε αρχική τιμή του διανύσματος $W^{(0)} > 0$, ο παραπάνω αλγόριθμος θα συγκλίνει στον ΜΠΕΜΠ.

Απόδειξη. Στην (2.26) αποδείχθηκε ότι ο λογάριθμος της πιθανοφάνειας που δίνεται από την σχέση

$$l(\theta) = \sum_{j=1}^{n} \theta_j - \sum_{i=1}^{s} n_i \Omega_i(\theta)$$

είναι μία κοίλη συνάρτηση. Θεωρούμε τώρα ότι θέλουμε να βρούμε το

$$\max \left\{ l(\theta) : \sum_{j=1}^{n} \exp(\theta_j) = 1 \, , \Omega_1 = \widehat{\Omega}_1^{(m)}, \ldots, \Omega_s = \widehat{\Omega}_s^{(m)} \right\},$$

όπου $\widehat{\Omega}_i^{(m)}$ είναι η τιμή του Ω_i στην m-οστή επανάληψη. Αυτό είναι ένα *κυρτό πρόβλημα βελτιστοποίησης* το οποίο λύνεται μοναδικά στα σημεία $\widehat{\theta}_j^{(m)} = \log\left(\widehat{p}_j^{(m)}\right)$. Αν λοιπόν αρχικά οριστεί κάποιο $W^{(0)} > 0$ και θεωρήσουμε $\widehat{\theta}^{(1)}, \widehat{\theta}^{(2)}, \ldots$ την ακολουθία η οποία δημιουργείται από τον παραπάνω αλγόριθμο, τότε θα ισχύει

$$l(\widehat{\theta}^{(m)}) \leq l(\widehat{\theta}^{(m+1)}) \leq l(\widehat{\theta})$$

για όλα τα $m \in \mathbb{N}$, όπου $\widehat{\theta}$ είναι η μοναδική τιμή που μεγιστοποιεί την πιθανοφάνεια. Ως εκ τούτου η ακολουθία $l(\widehat{\theta}^{(m)})$ είναι μονότονη αλλά και φραγμένη και συγκλίνει σε κάποια ποσότητα $D \leq l(\widehat{\theta})$. Λόγω όμως της κοιλότητας της l προκύπτει ότι $D = l(\widehat{\theta})$. Τελικά λόγω της συνέχειας της l και της μοναδικότητας του μεγίστου συνεπάγεται ότι $\lim_{m\uparrow\infty}(\widehat{\theta}^{(m)}) = \widehat{\theta}$. □

3.3 Παραδείγματα

Στην ενότητα αυτή θα αξιολογηθεί η αποδοτικότητα του ΜΠΕΜΠ σε σχέση με τον ΕΜΠ. Με τον όρο αποδοτικότητα αναφερόμαστε στον λόγο δύο Μέσων Τετραγωνικών Σφαλμάτων (ΜΤΣ). Αρχικά θα προσομοιωθούν τυχαία δείγματα από συγκεκριμένα παραμετρικά μοντέλα και θα εκτιμηθεί η πρώτη και η δεύτερη ροπή. Θα γίνουν

εκτιμήσεις Monte-Carlo βάσει του ΜΠΕΜΠ και του ΕΜΠ και θα υπολογιστεί η αποδοτικότητα του ενός σε σχέση με τον άλλο. Σε κάθε περίπτωση θα παρουσιαστεί ένας πίνακας με τις πραγματικές τιμές των ροπών, τις εκτιμηθείσες ροπές καθώς και τις τιμές της αποδοτικότητας. Να παρατηρηθεί ότι η εκτίμηση των ροπών στην περίπτωση του ΜΠΕΜΠ θα γίνει σύμφωνα με τους ακόλουθους τύπους

1. $\widehat{\mathsf{E}[X]} = \sum t_i \widehat{p}_i$
2. $\widehat{\mathsf{E}[X^2]} = \sum t_i^2 \widehat{p}_i$

όπου t_i είναι οι παρατηρηθείσες τιμές και $\widehat{p}_i$ οι αντίστοιχες πιθανότητες όπως προκύπτουν από τον αλγόριθμο.
Πριν από αυτά όμως θα παρουσιαστεί ένα παράδειγμα το οποίο υπάρχει στο άρθρο του Vardi (1985) και μελετώνται τέσσερα ανεξάρτητα δείγματα, κάθε ένα με ξεχωριστή συνάρτηση βάρους.

Παράδειγμα 1. Έστω ότι μας ενδιαφέρει η μελέτη ενός μη ελεγχόμενου φυσικού φαινομένου. Για τον λόγο αυτό τέσσερεις ερευνητές ανεξάρτητα ο ένας από τον άλλο καταγράφουν μετρήσεις ώστε να εκτιμηθεί η συνάρτηση κατανομής F του φαινομένου αυτού. Συγκεκριμένα για κάθε ερευνητή ισχύουν τα ακόλουθα:

- Ο πρώτος ερευνητής, εξ' αιτίας περιορισμένων πειραματικών συνθηκών, μπορεί να παρατηρήσει το φαινόμενο μόνο στο εύρος 10 έως 20. Έξω από τα όρια αυτά, ακόμα και να συμβεί το φαινόμενο, θα περάσει απαρατήρητο.
- Ο δεύτερος ερευνητής έχει ελαφρώς καλύτερο εξοπλισμό από τον πρώτο. Για τον λόγο αυτόν στο εύρος 10 έως 20 παρατηρεί πλήρως το φαινόμενο ενώ έξω από τα όρια αυτά έχει πιθανότητα 50% να το παρατηρήσει.
- Ο τρίτος ερευνητής έχει στην διάθεσή του τον πιο σύγχρονο εξοπλισμό με αποτέλεσμα να μπορεί να παρατηρήσει το φαινόμενο σε όλο του το εύρος.

- Τέλος ο τέταρτος ερευνητής μπορεί να παρατηρήσει το φαινόμενο ανάλογα με την τιμή του, δηλαδή ανάλογα με το πόσο μεγάλη ή μικρή τιμή παίρνει, αλλά σε όλο του το εύρος.

Τα δεδομένα φαίνονται στον ακόλουθο πίνακα:

Ερευνητής	**Παρατηρήσεις**
1	13, 15, 16, 18
2	9, 11, 17, 18
3	8, 11, 13, 16, 16, 17, 22
4	15, 19, 22, 22, 25

Το ερώτημα που προκύπτει είναι πώς μπορούν να συνδυαστούν οι τιμές των δειγμάτων αυτών ώστε να προκύψει η εκτίμηση της συνάρτησης κατανομής. Είναι προφανές από την ανάλυση που έγινε παραπάνω ότι κάθε ένας από τους ερευνητές παρατηρηρεί το φαινόμενο αλλά οι παρατηρήσεις που προκύπτουν δεν είναι από την βασική κατανομή. Κάθε μία από τις κατανομές είναι μία μεροληπτική εκδοχή της η οποία έχει διαφορετική συνάρτηση βάρους. Συγκεκριμένα, σύμφωνα με τον τρόπο που κάθε επιστήμονας παρατηρεί το φαινόμενο οι συναρτήσεις βάρους είναι οι ακόλουθες:

$$w_1(x) = I_{[10,20]}(x) \qquad w_2(x) = \frac{1 + I_{[10,20]}(x)}{2}$$
$$w_3(x) = 1 \qquad w_4(x) = x$$

όπου

$$I_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A. \end{cases}$$

Με χρήση των παραπάνω συναρτήσεων και των δειγμάτων που παρατηρήθηκαν προκύπτει ο Πίνακας 3.1. Να παρατηρήσουμε απλά ότι για να εκτιμηθούν οι παραπάνω ποσότητες χρειάστηκαν 9 επαναλήψεις του αλγορίθμου ενώ η εκτίμηση του διανύσματος W είναι

$$\widehat{W} = (0.5951107, 0.7975553, 1, 15.95222).$$

Πίνακας 3.1: ΜΠΕΜΠ

Τιμές	Πιθανότητες
8	0.0832
9	0.0811
11	0.0902
13	0.0877
15	0.0853
16	0.1263
17	0.0831
18	0.082
19	0.0405
22	0.1829
25	0.0577

Παράδειγμα 2. Έστω η τυχαία μεταβλητή X η οποία ακολουθεί την εκθετική κατανομή με παράμετρο θ και συνάρτηση πυκνότητας που δίνεται από τον τύπο:

$$f(x|\theta) = \frac{1}{\theta} e^{-x/\theta}, \quad x > 0.$$

Είναι εύκολο να δειχθεί ότι η σταθμισμένη εκδοχή της X με βάρος $w(x) = x^k$ ακολουθεί την κατανομή γάμμα με παραμέτρους $(k+1, \theta)$ $\forall k \in \mathbb{N}$.

Ας υποθέσουμε ότι έχουμε τρία ανεξάρτητα τυχαία δείγματα μεγέθους n_1, n_2 και n_3 τα οποία προέρχονται αντίστοιχα από την εκθετική κατανομή με παράμετρο θ, την κατανομή γάμμα με παραμέτρους $(2, \theta)$ και την κατανομή γάμμα με παραμέτρους $(3, \theta)$. Προφανώς οι δύο τελευταίες κατανομές αποτελούν σταθμισμένες εκδοχές της πρώτης με βάρη $w_1(x) = x$ και $w_2(x) = x^2$ αντίστοιχα. Έστω ότι τα τ.δ. είναι τα $X_1, X_2, ..., X_{n_1}$, $Y_1, Y_2, ..., Y_{n_2}$, $Z_1, Z_2, ..., Z_{n_3}$. Χρησιμοποιώντας την από κοινού κατανομή των τριών δειγμάτων θα υπολογιστεί αρχικά ο Εκτιμητής Μεγίστης Πιθανοφάνειας για την άγνωστη παράμετρο θ.

Ισχύει λοιπόν ότι:

$$\begin{aligned} f(\mathbf{x},\mathbf{y},\mathbf{z}|\theta) &= \prod_{i=1}^{n_1} f_X(x_i) \prod_{j=1}^{n_2} f_Y(y_j) \prod_{s=1}^{n_3} f_Z(z_s) \\ &= \prod_{i=1}^{n_1} \frac{1}{\theta} e^{x_i/\theta} \prod_{j=1}^{n_2} \frac{y_j e^{-y_j/\theta}}{\theta^2 \Gamma(2)} \prod_{s=1}^{n_3} \frac{z_s^2 e^{-z_s/\theta}}{\theta^3 \Gamma(3)} \\ &\propto \theta^{-n_1} e^{-\frac{1}{\theta}\sum_{i=1}^{n_1} x_i} \theta^{-2n_2} e^{-\frac{1}{\theta}\sum_{j=1}^{n_2} y_j} \theta^{-3n_3} e^{-\frac{1}{\theta}\sum_{s=1}^{n_3} z_s}. \end{aligned}$$

Για παρατηρηθέντα δείγματα X, Y, Z, η συνάρτηση πιθανοφάνειας είναι

$$L(\theta) \propto \theta^{-(n_1+2n_2+3n_3)} e^{-\frac{1}{\theta}(\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j + \sum_{s=1}^{n_3} z_s)},$$

ενώ για τον λογάριθμο ισχύει ότι

$$l(\theta) \propto -n_1 \log(\theta) - \frac{1}{\theta}\sum_{i=1}^{n_1}(x_i) - 2n_2 \log(\theta) - \frac{1}{\theta}\sum_{j=1}^{n_2}(y_j) - 3n_3 \log(\theta) - \frac{1}{\theta}\sum_{s=1}^{n_3}(z_s).$$

Παραγωγίζοντας ως προς θ ισχύει

$$\frac{dl(\theta)}{d\theta} = -\frac{n_1}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n_1}(x_i) - \frac{2n_2}{\theta} + \frac{1}{\theta^2}\sum_{j=1}^{n_2}(y_j) - \frac{3n_3}{\theta} + \frac{1}{\theta^2}\sum_{s=1}^{n_3}(z_s).$$

Εξισώνοντας με το μηδέν και λύνοντας ως προς θ προκύπτει

$$-n_1 + \frac{1}{\theta}\sum_{i=1}^{n_1} x_i - 2n_2 + \frac{1}{\theta}\sum_{j=1}^{n_2} y_j - 3n_3 + \frac{1}{\theta}\sum_{s=1}^{n_3} z_s = 0 \Rightarrow$$

$$-(n_1+2n_2+3n_3) + \frac{1}{\theta}\left[\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j + \sum_{s=1}^{n_3} z_s\right] = 0 \Rightarrow$$

$$\frac{1}{\theta}\left[\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j + \sum_{s=1}^{n_3} z_s\right] = (n_1+2n_2+3n_3) \Rightarrow$$

$$\widehat{\theta} = \frac{\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j + \sum_{s=1}^{n_3} z_s}{n_1+2n_2+3n_3}.$$

Προσομοιώνουμε τρία ανεξάρτητα τυχαία δείγματα από τις κατανομές γάμμα $G(i, \theta)$, $i = 1, 2, 3$, με χρήση του R. Για $b = 500$ επαναλήψεις εκτιμώνται οι δύο πρώτες ροπές χρησιμοποιώντας τους δύο διαφορετικούς εκτιμητές, με χρήση της μεθόδου Monte Carlo. Επίσης υπολογίζεται η αποδοτικότητα του ενός εκτιμητή σε σχέση με τον άλλο. Τα αποτελέσματα της προσομοίωσης φαίνονται στους πίνακες που ακολουθούν.

Πίνακας 3.2: $n_1 = n_2 = n_3 = 10$ με $\theta = 3$

	Πραγματικές τιμές	**ΕΜΠ**	**ΜΠΕΜΠ**	**αποδοτικότητα**
$\mathsf{E}[X]$	3	2.98	3.0640	0.38
$\mathsf{E}[X^2]$	18	17.91	18.48	0.65

Καλά αποτελέσματα όμως προκύπτουν και στην περίπτωση που έχουμε δεδομένα μόνο από τις σταθμισμένες κατανομές και όχι από την βασική κατανομή. Συγκεκριμένα από τις προσομοιώσεις προέκυψε ο Πίνακας 3.3

Πίνακας 3.3: $n_1 = 0, n_2 = n_3 = 30$ με $\theta = 3$

	Πραγματικές τιμές	**ΕΜΠ**	**ΜΠΕΜΠ**	**αποδοτικότητα**
$\mathsf{E}[X]$	3	2.95	3.24	0.73
$\mathsf{E}[X^2]$	18	18.1	18.42	0.87

Παράδειγμα 3. Στη συνέχεια θα παρουσιαστεί ένα παράδειγμα όπου η βασική κατανομή έχει συνάρτηση πιθανότητας που δίνεται από τον τύπο

$$f(x|\theta) = -\frac{\theta^x}{x \log(1-\theta)}$$

με $x = 1, 2, 3, \ldots$ και $\theta \in (0, 1)$. Πρόκειται για την λογαριθμική κατανομή με μέση τιμή

$$\mathsf{E}[X] = \sum_{x=1}^{\infty} x f(x|\theta) = \sum_{x=1}^{\infty} x \left(-\frac{\theta^x}{x \log(1-\theta)} \right) = -\frac{1}{\log(1-\theta)} \sum_{x=1}^{\infty} \theta^x = \frac{\alpha\theta}{1-\theta},$$

όπου $\alpha = -1/\log(1-\theta)$. Η δεύτερη ροπή της είναι

$$\mathsf{E}\left[X^2\right] = \sum_{x=1}^{\infty} x^2 f\left(x|\theta\right) = \sum_{x=1}^{\infty} x^2 \left(-\frac{\theta^x}{x\log\left(1-\theta\right)}\right) = \alpha \sum_{x=1}^{\infty} x\theta^x = \frac{\alpha\theta}{\left(1-\theta\right)^2}.$$

Στην περίπτωση αυτή, η αντίστοιχη μεροληπτική λόγω μεγέθους κατανομή έχει συνάρτηση πιθανότητας

$$g\left(x|\theta\right) = \left(1-\theta\right)\theta^{x-1}, \quad x = 1, 2, 3\ldots$$

Αυτή είναι η συνάρτηση πιθανότητας μίας γεωμετρικής κατανομής με πιθανότητα επιτυχίας $1-\theta$.

Έστω λοιπόν ότι έχουμε δύο δείγματα, ένα από την βασική κατανομή (λογαριθμική) και ένα από την μεροληπτική λόγω μεγέθους εκδοχή της (γεωμετρική). Τότε είναι εφικτό να εκτιμηθεί η άγνωστη ποσότητα θ μεγιστοποιώντας την συνάρτηση πιθανοφάνειας ή αντίστοιχα τον λογάριθμό της. Συγκεκριμένα αν θεωρήσουμε ότι το αρχικό δείγμα X έχει μέγεθος n_1 και το σταθμισμένο δείγμα Y έχει μέγεθος n_2, τότε η συνάρτηση πιθανοφάνειας είναι η

$$L(\theta) = \prod_{i=1}^{n_1} \frac{\alpha\theta^{x_i}}{x_i} \prod_{j=1}^{n_2} (1-\theta)\theta^{y_j-1}, \quad \theta \in \Theta = (0,1).$$

Από την σχέση αυτή προκύπτει εύκολα ότι

$$l(\theta) = c + n_1 \log\alpha + \big(\sum_{i=1}^{n_1} x_i + \sum_{j=1}^{n_2} y_j - n_2\big)\log\theta + n_2\log(1-\theta), \qquad (3.3)$$

όπου c μία ποσότητα που δεν περιέχει την άγνωστη παράμετρο θ.

Προσομοιώνοντας δύο ανεξάρτητα τυχαία δείγματα από λογαριθμική και γεωμετρική κατανομή, εκτιμώνται οι δύο πρώτες ροπές με χρήση της μεθόδου Monte Carlo για $b = 500$ επαναλήψεις. Ο ΕΜΠ υπολογίζεται με την μέθοδο Newton-Raphson μεγιστοποιώντας την (3.3) ενώ ο ΜΠΕΜΠ υπολογίζεται βάσει του επαναληπτικού αλγορίθμου των Davidov and Iliopoulos. Επίσης υπολογίζεται και η αποδοτικότητα του ενός εκτιμητή σε σχέση με τον άλλο. Τα αποτελέσματα των προσομοιώσεων για διάφορα μεγέθη δειγμάτων παρουσιάζονται στους πίνακες που ακολουθούν.

Πίνακας 3.4: $n_1 = n_2 = 20$ με $\theta = 0.4$

	Πραγματικές τιμές	**ΕΜΠ**	**ΜΠΕΜΠ**	**αποδοτικότητα**
$\mathsf{E}[X]$	1.31	1.30	1.32	0.88
$\mathsf{E}[X^2]$	2.18	2.17	2.19	0.91

Πίνακας 3.5: $n_1 = n_2 = 10$ με $\theta = 0.4$

	Πραγματικές τιμές	**ΕΜΠ**	**ΜΠΕΜΠ**	**αποδοτικότητα**
$\mathsf{E}[X]$	1.31	1.30	1.32	0.86
$\mathsf{E}[X^2]$	2.18	2.21	2.22	0.91

Πίνακας 3.6: $n_1 = 0, n_2 = 30$ με $\theta = 0.4$

	Πραγματικές τιμές	**ΕΜΠ**	**ΜΠΕΜΠ**	**αποδοτικότητα**
$\mathsf{E}[X]$	1.31	1.30	1.31	0.74
$\mathsf{E}[X^2]$	2.18	2.18	2.19	0.93

Παράδειγμα 4. Τέλος θα παρουσιαστεί ένα παράδειγμα όπου η βασική κατανομή ακολουθεί την κατανομή βήτα με παραμετρους $\alpha > 0$, $\beta > 0$. Τότε η μεροληπτική κατανομή με βάρος $w(x) = x^k$ ακολουθεί την κατανομή βήτα με παραμέτρους $(k + \alpha, \beta)$ $\forall k \in \mathbb{N}$. Και εδώ θα εκτιμηθούν οι δύο πρώτες ροπές. Έστω δύο δείγματα όπου το ένα προέρχεται από την κατανομή βήτα$(2, 3)$ και το δεύτερο από την αντίστοιχη μεροληπτική λόγω μεγέθους εκδοχή της βήτα$(2 + 1, 3)$. Να σημειωθεί ότι αν θεωρήσουμε ότι το πρώτο δείγμα έχει μέγεθος n_1 ενώ το δεύτερο έχει μέγεθος n_2 και συμβολίσουμε τις άγνωστες παραμέτρους με α και β, τότε η συνάρτηση πιθανοφάνειας είναι

$$L(\alpha, \beta) = \prod_{i=1}^{n_1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x_i^{\alpha-1} (1 - x_i)^{\beta-1} \prod_{j=1}^{n_2} \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} y_j^{\alpha} (1 - y_j)^{\beta-1}, \qquad (3.4)$$

ενώ ο λογάριθμός της

$$l(\alpha,\beta) = c + n_1 \log\Big(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\Big) + (\alpha-1)\sum_{i=1}^{n_1}\log(x_i) + n_2\log\Big(\frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)}\Big)$$

$$+\alpha\sum_{j=1}^{n_2}\log(y_i) + (\beta-1)\sum_{k=1}^{n_1+n_2}\log(1-z_k),$$

όπου z_k συμβολίζουμε το από κοινού δείγμα ενώ c είναι μία ποσότητα που δεν εξαρτάται από τα α, β.

Προσομοιώνοντας δύο ανεξάρτητα τυχαία δείγματα από κατανομές βήτα(2,3) και βήτα(3,3) εκτιμώνται οι δύο πρώτες ροπές με χρήση της μεθόδου Monte Carlo για $b = 500$ επαναλήψεις. Όπως και στο προηγούμενο παράδειγμα, ο ΕΜΠ προκύπτει μεγιστοποιώντας την συνάρτηση πιθανοφάνειας (3.4) με την μέθοδο Newton-Raphson ενώ ο ΜΠΕΜΠ υπολογίζεται χρησιμοποιώντας τον επαναληπτικό αλγόριθμο των Davidov and Iliopoulos. Τέλος εκτιμάται η αποδοτικότητα του ενός εκτιμητή σε σχέση με τον άλλο. Τα αποτελέσματα των προσομοιώσεων παρουσιάζονται στους πίνακες που ακολουθούν.

Πίνακας 3.7: $n_1 = 20, n_2 = 20$ με $\alpha = 2, \beta = 3$

	Πραγματικές τιμές	**ΕΜΠ**	**ΜΠΕΜΠ**	**αποδοτικότητα**
$\mathsf{E}[X]$	0.4	0.39	0.39	0.94
$\mathsf{E}[X^2]$	0.2	0.198	0.199	0.95

Πίνακας 3.8: $n_1 = 10, n_2 = 10$ με $\alpha = 2, \beta = 3$

	Πραγματικές τιμές	**ΕΜΠ**	**ΜΠΕΜΠ**	**αποδοτικότητα**
$\mathsf{E}[X]$	0.4	0.41	0.40	0.97
$\mathsf{E}[X^2]$	0.2	0.21	0.21	0.98

Τέλος παρουσιάζονται τα αποτελέσματα των προσομοιώσεων στην περίπτωση που τα μοναδικά δεδομένα που διαθέτουμε προέρχονται από την μεροληπτική κατανομή.

Πίνακας 3.9: $n_1 = 0, n_2 = 20$ με $\alpha = 2, \beta = 3$

	Πραγματικές τιμές	**ΕΜΠ**	**ΜΠΕΜΠ**	**αποδοτικότητα**
$\mathsf{E}[X]$	0.4	0.41	0.42	0.95
$\mathsf{E}[X^2]$	0.2	0.21	0.21	0.98

Βάσει των παραπάνω παραδειγμάτων διαπιστώνουμε ότι ο ΜΠΕΜΠ μπορεί να δώσει αρκετά καλές εκτιμήσεις σε σύγκριση με τον ΕΜΠ. Σαφώς στην περίπτωση που γνωρίζουμε τις κατανομές από τις οποίες προέρχονται τα δείγματα, ο ΕΜΠ είναι καλύτερος. Επειδή συχνά όμως δεν ξέρουμε ποιές είναι οι κατανομές αυτές, μπορούμε να χρησιμοποιούμε τον ΜΠΕΜΠ γνωρίζοντας ότι η απώλεια αποδοτικότητας δεν είναι γενικά πολύ μεγάλη.

Κεφάλαιο 4

Εισαγωγή παραμέτρων στο μοντέλο

4.1 Εισαγωγή

Στην ανάλυση που έγινε μέχρι τώρα υποθέταμε ότι η συνάρτηση βάρους w είναι γνωστή. Γενικά όμως κάτι τέτοιο είναι πολύ σπάνιο. Ως εκ τούτου, στα πλαίσια της μοντελοποίησης, μπορούμε να υποθέσουμε ότι είναι γνωστή η συναρτησιακή μορφή της w και να εισαγάγουμε σε αυτήν άγνωστες παραμέτρους (δες Davidov, Fokianos and Iliopoulos, 2009). Στην περίπτωση αυτή, η ποσότητα $w_i(z_j)$ θα αντικατασταθεί με $w_i(z_j, \varphi_i)$ όπου $\varphi_i \in \Re^{d_i}$.

Έστω λοιπόν $s+1$ το πλήθος ανεξάρτητα τυχαία δείγματα που προέρχονται από κάποιες άγνωστες κατανομές με πυκνότητες $f_0, f_1, \ldots, f_s$, δηλαδή

$$Y_{01}, \ldots, Y_{0n_0} \sim f_0(y)$$

$$\vdots$$

$$Y_{s1}, \ldots, Y_{sn_s} \sim f_s(y).$$

Ο βασικός έλεγχος που μας ενδιαφέρει είναι για την υπόθεση

$$f_0 = \cdots = f_s.$$

Στην περίπτωση όπου $f_j(y)$ είναι η συνάρτηση πυκνότητας της κανονικής κατανομής με μέσο μ_j και διακύμανση σ^2, $N(\mu_j, \sigma^2)$, ο έλεγχος της ισότητας των κατανομών

ανάγεται στον έλεγχο της ισότητας των μέσων τιμών όπως γίνεται στην ανάλυση διακύμανσης κατά έναν παράγοντα. Θεωρώντας το πρώτο επίπεδο ως "επίπεδο αναφοράς" θα έχουμε

$$\begin{aligned}\frac{f_j(y)}{f_0(y)} &= \frac{\frac{1}{\sigma\sqrt{2\pi}}\exp\big(-\frac{(y-\mu_j)^2}{2\sigma^2}\big)}{\frac{1}{\sigma\sqrt{2\pi}}\exp\big(-\frac{(y-\mu_0)^2}{2\sigma^2}\big)} \\ &= \exp\Big(-\frac{y^2}{2\sigma^2}+\frac{y\mu_j}{\sigma^2}-\frac{\mu_j^2}{2\sigma^2}+\frac{y^2}{2\sigma^2}-\frac{y\mu_0}{\sigma^2}+\frac{\mu_0^2}{2\sigma^2}\Big) \\ &= \exp\Big[\frac{\mu_0^2-\mu_j^2}{2\sigma^2}+\frac{\mu_j-\mu_0}{\sigma^2}y\Big]\ , \quad j=1,2\ldots s.\end{aligned}$$

Θέτοντας στην συνέχεια

$$\alpha_j = \frac{\mu_0^2-\mu_j^2}{2\sigma^2}\ ,\ \beta_j = \frac{\mu_j-\mu_0}{\sigma^2} \tag{4.1}$$

βλέπουμε ότι ο παραπάνω λόγος έχει την μορφή

$$\frac{f_j(y)}{f_0(y)} = \exp\left(\alpha_j+\beta_j y\right). \tag{4.2}$$

Επομένως ο έλεγχος της υπόθεσης

$$H_0 : \mu_0 = \mu_1 = \ldots = \mu_s$$

έναντι της εναλλακτικής $\mu_i \neq \mu_j$, για κάποια $i, j \in \{0, 1, \ldots, s\}$ είναι ισοδύναμος με τον έλεγχο

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_s = 0,$$

κατά $\beta_i \neq 0$ για κάποιο $i \in \{1, \ldots, s\}$. Είναι σαφές ότι στην περίπτωση που ισχύει ότι $\beta_j = 0$, τότε από την (4.1) συνεπάγεται ότι και $\alpha_j = 0$, $j = 1, 2 \ldots, s$.

Φυσικά όμως στη γενική περίπτωση τα τυχαία δείγματα δεν ακολουθούν κανονική κατανομή. Στην περίπτωση αυτή, αγνοώντας την υπόθεση της κανονικότητας των δεδομένων, θα μπορούσαμε να ισχυριστούμε ότι ισχύει μία παρόμοια σχέση αντίστοιχη της (4.2), η οποία έχει την μορφή

$$\frac{f_j(y)}{f_0(y)} = \exp\left\{\alpha_j+\beta_j h(y)\right\}, \quad j=1,2\ldots s, \tag{4.3}$$

όπου h μία αυθαίρετη αλλά γνωστή συνάρτηση.

Η σχέση (4.3) συνδέεται με την πολλαπλή λογιστική παλινδρόμηση. Έστω κατηγορική τυχαία μεταβλητή J τέτοια ώστε $\mathsf{P}(J=j)=\pi_j$ ενώ ισχύει $\sum_{j=0}^{s}\pi_j=1$. Έστω επίσης ότι $\mathsf{P}(Y=y|J=j)=f_j(y)$ όπου $j=0,1,\ldots,s$. Στην περίπτωση αυτή χρησιμοποιώντας το θεώρημα Bayes παρατηρούμε ότι

$$\mathsf{P}(J=j|Y=y)=\frac{\exp\left(\alpha_j^*+\beta_j h\left(y\right)\right)}{1+\sum_{k=0}^{s-1}\exp\left(\alpha_k^*+\beta_k h\left(y\right)\right)}.$$

Τότε,

$$\begin{aligned}
\frac{f_j(y)}{f_0(y)} &= \frac{\mathsf{P}(Y=y|J=j)}{\mathsf{P}(Y=y|J=0)}\\
&= \frac{\mathsf{P}\left[J=j|Y=y\right]\mathsf{P}\left[Y=y\right]/\mathsf{P}\left[J=j\right]}{\mathsf{P}(J=0|Y=y)\mathsf{P}(Y=y)/\mathsf{P}(J=0)}\\
&= \frac{\pi_0}{\pi_j}\Big(\frac{\exp\left(\alpha_j^*+\beta_j h\left(y\right)\right)}{1+\sum_{k=0}^{s-1}\exp\left(\alpha_k^*+\beta_k h\left(s\right)\right)}\Big)\Big/\Big(\frac{\exp\left(\alpha_0^*+\beta_0 h\left(y\right)\right)}{1+\sum_{k=0}^{s-1}\exp\left(\alpha_k^*+\beta_k h\left(y\right)\right)}\Big)\\
&= \frac{\pi_0}{\pi_j}\frac{\exp\left(\alpha_j^*+\beta_j h\left(y\right)\right)}{\exp\left(\alpha_0^*+\beta_0 h\left(y\right)\right)}\\
&= \exp\Big(\underbrace{\alpha_j^*-\alpha_0^*+\log\left(\frac{\pi_0}{\pi_j}\right)}_{\alpha_j}+\left(\beta_j-\beta_0\right)h\left(y\right)\Big).
\end{aligned}$$

Επειδή όμως ισχύει $\alpha_0^*=\beta_0=0$, θα έχουμε

$$\frac{f_j\left(y\right)}{f_0\left(y\right)}=\exp\left(\alpha_j^*+\log\left(\frac{\pi_0}{\pi_j}\right)+\beta_j h\left(y\right)\right),\quad j=1,\ldots,s.$$

Να παρατηρηθεί ότι η παραπάνω σχέση μπορεί να χρησιμοποιηθεί ώστε να γίνουν οι εκτιμήσεις για τις τιμές των αγνώστων παραμέτρων α_j και β_j με χρήση του αλγορίθμου που χρησιμοποιείται και στην λογιστική παλινδρόμηση. Όπως φαίνεται και από τον τύπο αυτόν δεν υπάρχει διαφορά στις τιμές των παραμέτρων β_j σε σχέση με τη λογιστική παλινδρόμηση ενώ για τις παραμέτρους α_j προστίθεται ο όρος $\log(\pi_0/\pi_j)$, όπου π_0 η κατηγορία αναφοράς. Στην περίπτωση όπου $h\left(y\right)\neq y$, τότε αρκεί να μετασχηματιστούν όλα τα δεδομένα, να αποκτήσουν την μορφή της συνάρτησης $h\left(y\right)$ και μέσω λογιστικής παλινδρόμησης να εκτιμηθούν οι άγνωστες παράμετροι.

Είναι εύκολο να βρεθεί η σχέση που συνδέει μεταξύ τους τις άγνωστες παραμέτρους α_j και β_j. Συγκεκριμένα, σύμφωνα με την (4.3),

$$f_j(y) = \exp\left(\alpha_j + \beta_j h(y)\right) f_0(y).$$

Ολοκληρώνοντας σε όλο το πεδίο που ορίζονται οι συναρτήσεις έστω $\mathcal{A}$ (ή αντίστοιχα αθροίζοντας) προκύπτουν τα ακόλουθα

$$\begin{aligned} 1 &= \int_{\mathcal{A}} \exp\left(\alpha_j + \beta_j h(y)\right) f_0(y)\, dy \\ &= e^{\alpha_j} \int_{\mathcal{A}} \exp\left(\beta_j h(y)\right) f_0(y)\, dy \\ &= e^{\alpha_j} M_h(\beta_j), \end{aligned}$$

όπου M_h είναι η ροπογεννήτρια της $h(Y)$ όταν $Y \sim f_0$. Από την σχέση αυτή προκύπτει τελικά

$$\alpha_j = -\log\left[M_h(\beta_j)\right],$$

για β_j τέτοια ώστε η M_h να είναι πεπερασμένη.

Συνδυάζοντας όλα τα δεδομένα από τα $s+1$ τυχαία δείγματα, ενδιαφέρον έχει η μελέτη των παρακάτω προβλημάτων:

1. Μη παραμετρική εκτίμηση της $F_0(y)$, η οποία είναι η συνάρτηση κατανομής που συνδέεται με την $f_0(x)$.

2. Έλεγχοι υποθέσεων για τα β, π.χ. $H_0 : \beta_1 = \cdots = \beta_s = 0$.

3. Εκτίμηση των παραμέτρων $\alpha = (\alpha_1, \ldots, \alpha_s)'$ και $\beta = (\beta_1, \ldots, \beta_s)'$ και ασυμπτωτικές ιδιότητες των αντίστοιχων εκτιμητών.

4.2 Εκτίμηση παραμέτρων

Η εκτίμηση της συνάρτησης κατανομής όπως επίσης και των αγνώστων παραμέτρων θα γίνει όπως και στην περίπτωση που η συνάρτηση βάρους ήταν γνωστή. Σε ό,τι

ακολουθεί θα συμβολίζεται με y_{ij} η j παρατήρηση από το i δείγμα για $i = 0, 1, \ldots, s$ και $j = 1, \ldots, n_i$. Επίσης, με $y_1, \ldots, y_n$, $n = \sum_{i=0}^{s} n_i$, θα συμβολίζονται τα δεδομένα από την ένωση των $s+1$ δειγμάτων. Έστω $p_i = dF_0(y_i)$, $i = 1, \ldots, n$, οι πιθανότητες που αντιστοιχούν στα παρατηρηθέντα σημεία. Τότε η συνάρτηση πιθανοφάνειας είναι

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{p}) &= \prod_{i=1}^{n} p_i \prod_{j=1}^{n_1} \exp\left(\alpha_1 + \beta_1 h(y_{1j})\right) \ldots \prod_{j=1}^{n_s} \exp\left(\alpha_s + \beta_s h(y_{sj})\right) \\ &= \prod_{i=1}^{n} p_i \prod_{k=1}^{s} \prod_{j=1}^{n_k} \exp\left(\alpha_k + \beta_k h(y_{kj})\right). \end{aligned} \tag{4.4}$$

Η μεγιστοποίηση της παραπάνω πιθανοφάνειας ως προς $\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{p}$ θα γίνει υπό τους περιορισμούς

$$\sum_{i=1}^{n} p_i = 1, \tag{4.5}$$

$$\sum_{i=1}^{n} p_i \left(w_1(y_i) - 1\right) = 0, \ldots, \sum_{i=1}^{n} p_i \left(w_s(y_i) - 1\right) = 0, \tag{4.6}$$

όπου $w_j(y) = \exp\left(\alpha_j + \beta_j h(y)\right)$, $j = 1, 2, \ldots, s$.

Έστω $l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{p}) \equiv \log L(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{p})$. Χρησιμοποιώντας πολλαπλασιαστές Lagrange ορίζουμε τη συνάρτηση:

$$g(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{p}) = l(\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{p}) - \lambda_0 \left(\sum_{i=1}^{n} p_i - 1 \right) - \sum_{k=1}^{s} \lambda_k \sum_{i=1}^{n} p_i \left(w_k(y_i) - 1\right).$$

Παραγωγίζοντας ως προς p_j προκύπτει ότι

$$\frac{dg(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\lambda}, \mathbf{p})}{dp_j} = \frac{1}{p_j} - \lambda_0 - \sum_{k=1}^{s} \lambda_k \left(w_k(y_j) - 1\right) = 0 \tag{4.7}$$

ή, ισοδύναμα,

$$1 - p_j \lambda_0 - p_j \sum_{k=1}^{s} \lambda_k \left(w_k(y_j) - 1\right) = 0.$$

Αθροίζοντας για όλα τα j έχουμε ότι

$$\sum_{j=1}^{n} 1 - \sum_{j=1}^{n} p_j \lambda_0 - \sum_{j=1}^{n} p_j \sum_{k=1}^{s} \lambda_k \left(w_k(y_j) - 1\right) = 0 \Rightarrow$$

$$n - \lambda_0 - \sum_{k=1}^{s} \lambda_k \sum_{j=1}^{n} p_j \left(w_k \left(y_j \right) - 1 \right) = 0.$$

Από την τελευταία εξίσωση λαμβάνοντας υπ' όψιν την (4.6), προκύπτει ότι

$$\lambda_0 = n.$$

Εν συνεχεία θεωρούμε ότι όλα τα λ_j είναι συναρτήσεις του n και συγκεκριμένα ότι έχουν την μορφή $\lambda_j = \nu_j n$, $j = 1, 2, \ldots, s$. Στην περίπτωση αυτή από την σχέση (4.7) προκύπτει εύκολα ότι

$$\begin{aligned} p_j = \frac{1}{\lambda_0 + \sum_{k=1}^{s} \lambda_k \left(w_k \left(y_j \right) - 1 \right)} &= \frac{1}{n + \sum_{k=1}^{s} n \nu_k \left(w_k \left(y_j \right) - 1 \right)} \\ &= \frac{1}{n} \frac{1}{1 + \sum_{k=1}^{s} \nu_k \left(w_k \left(y_j \right) - 1 \right)}. \end{aligned} \tag{4.8}$$

Με αντικατάσταση της (4.8) στην (4.6) προκύπτει:

$$\frac{1}{n} \sum_{i=1}^{n} \frac{w_j \left(y_i \right) - 1}{1 + \sum_{k=1}^{s} \nu_k \left(w_k \left(y_i \right) - 1 \right)} = 0, \quad j = 1, 2, \ldots, s. \tag{4.9}$$

Επόμενο βήμα είναι να αντικατασταθεί στην συνάρτηση πιθανοφάνειας η ποσότητα p_j. Έτσι η συνάρτηση πιθανοφάνειας θα γίνει μία συνάρτηση των $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$. Επομένως θα μεγιστοποιηθεί και θα βρεθούν οι εκτιμητές για τις άγνωστες αυτές ποσότητες. Με την αντικατάσταση αυτή προκύπτει ότι ο λογάριθμος της πιθανοφάνειας θα πάρει τη μορφή

$$l(\boldsymbol{\alpha}, \boldsymbol{\beta}) \equiv \log L \left(\boldsymbol{\alpha}, \boldsymbol{\beta} \right) = c - \sum_{i=1}^{n} \log \left(1 + \sum_{k=1}^{s} \nu_k \left(w_k \left(y_i \right) - 1 \right) \right) + \sum_{k=1}^{s} \sum_{j=1}^{n_k} \left(\alpha_k + \beta_k h \left(y_{kj} \right) \right),$$

όπου c είναι μία ποσότητα ανεξάρτητη των $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$.

Βρίσκοντας τώρα τις μερικές παραγώγους ως προς α_j και χρησιμοποιώντας την σχέση (4.9) προκύπτει ότι

$$\nu_j = \frac{n_j}{n}, \quad j = 1, 2, \ldots, s.$$

Με αντικατάσταση τελικά στην (4.8) θα προκύψει ότι:

$$\begin{aligned} p_j &= \frac{1}{n}\frac{1}{1+\sum_{k=1}^{s}\nu_k\left(w_k\left(y_j\right)-1\right)} \\ &= \frac{1}{n}\frac{n}{n+\sum_{k=1}^{s}n_k\left(w_k\left(y_j\right)-1\right)} \\ &= \frac{1}{n+\sum_{k=1}^{s}\Big(n_k w_k\left(y_j\right)-n_k\Big)} \\ &= \frac{1}{\underbrace{n-\sum_{k=1}^{s}n_k}_{n_0}+\sum_{k=1}^{s}n_k\left(w_k\left(y_j\right)\right)} \\ &= \frac{1}{n_0+\sum_{k=1}^{s}n_k w_k\left(y_j\right)}, \end{aligned}$$

από όπου παίρνουμε τελικά ότι:

$$p_j = \frac{1}{n_0}\frac{1}{1+\sum_{k=1}^{s}r_k w_k\left(y_j\right)}, \quad j=1,2,\ldots,n,$$

όπου $r_k = n_k/n_0$, $k=1,2,\ldots,s$.

Αντικαθιστώντας τα ν_j στον λογάριθμο της συνάρτησης πιθανοφάνειας παίρνουμε

$$l(\boldsymbol{\alpha},\boldsymbol{\beta}) = c - \sum_{i=1}^{n}\log\Big(1+\sum_{k=1}^{s}r_k w_k(y_i)\Big) + \sum_{k=1}^{s}\sum_{j=1}^{n_k}\Big(\alpha_k+\beta_k h\left(y_{kj}\right)\Big).$$

Από την παραπάνω σχέση παραγωγίζοντας ως προς α_j και β_j αντίστοιχα, καταλήγουμε στις εξισώσεις-σκόρ απο τις οποίες θα βρεθούν οι εκτιμήσεις μεγίστης πιθανοφάνειας για τις παραμέτρους $\boldsymbol{\alpha}$ και $\boldsymbol{\beta}$. Οι εξισώσεις αυτές δίνονται παρακάτω για $j=1,2,\ldots,s$:

$$\frac{\partial l(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \alpha_j} = -\sum_{i=1}^{n}\frac{r_j w_j\left(y_i\right)}{1+\sum_{k=1}^{s}r_k w_k\left(y_i\right)} + n_j = 0$$

$$\frac{\partial l(\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial \beta_j} = -\sum_{i=1}^{n}\frac{r_j h\left(y_i\right) w_j\left(y_i\right)}{1+\sum_{k=1}^{s}r_k w_k\left(y_i\right)} + \sum_{i=1}^{n_j}h\left(y_{ji}\right) = 0$$

Επομένως θα ισχύει

$$\begin{aligned}\widehat{p}_j &= \frac{1}{n_0}\frac{1}{1+\sum_{k=1}^{s} r_k \exp\left[\widehat{\alpha}_k + \widehat{\beta}_k h\left(y_j\right)\right]} \\ &= \left\{\sum_{k=0}^{s} n_k \exp[\widehat{\alpha}_k + \widehat{\beta}_k h(y_j)]\right\}^{-1}, \quad j = 1, 2 \ldots, n,\end{aligned} \tag{4.10}$$

ενώ ο ημιπαραμετρικός εκτιμητής μεγίστης πιθανοφάνειας για την συνάρτηση κατανομής F_0 δίνεται από την σχέση:

$$\widehat{F_0}(x) = \frac{1}{n_0}\sum_{i=1}^{n}\frac{I_{(-\infty,x]}(y_i)}{1+\sum_{k=1}^{s} r_k \exp\left(\widehat{\alpha}_k + \widehat{\beta}_k h(y_i)\right)}.$$

Να παρατηρήσουμε στο σημείο αυτό ότι οι εκτιμητές για τις υπόλοιπες συναρτήσεις κατανομής δίνονται από την σχέση

$$\widehat{F}_j(x) = \frac{1}{n_0}\sum_{i=1}^{n}\frac{I_{(-\infty,x]}(y_i)\exp\left(\widehat{\alpha}_j + \widehat{\beta}_j h(y_i)\right)}{1+\sum_{k=1}^{s} r_k \exp\left(\widehat{\alpha}_k + \widehat{\beta}_k h(y_i)\right)}, \quad j = 1, \ldots, s. \tag{4.11}$$

Παρατήρηση

Θέτοντας $z_j = h(y_j)$, να παρατηρηθεί ότι ο εκτιμητής υπάρχει και είναι μοναδικός αν και μόνον αν ο πίνακας

$$\begin{pmatrix} 1 & 1 & \ldots & 1 \\ z_1 & z_2 & \ldots & z_n \end{pmatrix}$$

είναι πλήρους τάξης, ενώ ταυτόχρονα

$$\left(\min_{j\in\mathcal{Y}_i} z_j, \max_{j\in\mathcal{Y}_i} z_j\right) \bigcap \left(\min_{j\in\bigcup\limits_{k\neq i}\mathcal{Y}_k} z_j, \max_{j\in\bigcup\limits_{k\neq i}\mathcal{Y}_k} z_j\right) \neq \emptyset.$$

Στην πραγματικότητα η τελευταία σχέση απλά σημαίνει ότι δεν επιτρέπεται κάποιο δείγμα να ξεχωρίζει από τα υπόλοιπα. Τα δείγματα μεταξύ τους δηλαδή επικαλύπτονται. Είναι παρόμοιο με αυτό που γινόταν και στην περίπτωση που η συνάρτηση βάρους ήταν γνωστή. Να υπενθυμίσουμε ότι οι εκτιμητές των παραμέτρων $\boldsymbol{\alpha}, \boldsymbol{\beta}$ εκτιμώνται με τους ίδιους αλγορίθμους που προκύπτουν και οι εκτιμητές όταν έχουμε πολλαπλή λογιστική παλινδρόμηση, καθώς οι δύο εξισώσεις πιθανοφάνειας ουσιαστικά ταυτίζονται.

4.3 Ασυμπτωτικά αποτελέσματα

4.3.1 Η περίπτωση των δύο δειγμάτων

Στην ενότητα αυτή θα μελετηθεί η περίπτωση που έχουμε δύο δείγματα. Το πρώτο που προέρχεται από την βασική κατανομή και έχει μέγεθος n_0 ενώ το δεύτερο προέρχεται από την σταθμισμένη κατανομή και έχει μέγεθος n_1. Για λόγους απλότητας στους συμβολισμούς θα γράφουμε α, β αντί για α_1, β_1. Σύμφωνα με τις (4.4) και (4.10), ο λογάριθμος της πιθανοφάνειας είναι

$$l(\alpha,\beta) = n_1\alpha + n_1\overline{z}_1\beta - \sum_{k=1}^{n}\log(n_0 + n_1 e^{\alpha+\beta z_k}),$$

όπου $\overline{z}_1 = \sum_{j=1}^{n_1} h(y_{1j})/n_1$. Επομένως οι μερικές παράγωγοι ως προς α και β είναι

$$\frac{\partial l(\alpha,\beta)}{\partial\alpha} = n_1\Big(1 - \sum_{k=1}^{n}\frac{e^{\alpha+\beta z_k}}{n_0 + n_1 e^{\alpha+\beta z_k}}\Big)$$

$$\frac{\partial l(\alpha,\beta)}{\partial\beta} = n_1\Big(\overline{z}_1 - \sum_{k=1}^{n}\frac{z_k e^{\alpha+\beta z_k}}{n_0 + n_1 e^{\alpha+\beta z_k}}\Big).$$

Έστω $r = n_1/n_0$. Τότε ισχύει

$$\frac{n_0}{n} = \frac{1}{1+r} \quad \text{και} \quad \frac{n_1}{n} = \frac{r}{1+r}.$$

Θα θεωρήσουμε για ευκολία ότι καθώς $n_1, n_0 \to \infty$ ο λόγος τους r παραμένει σταθερός. Να σημειωθεί όμως ότι τα ακόλουθα αποτελέσματα ισχύουν ακόμη και αν $n_1/n_0 \to r \in (0,1)$.

Θα μελετηθεί αρχικά η πρώτη παράγωγος. Αν συμβολίσουμε με z_{0i} τις παρατηρήσεις του πρώτου δείγματος και με z_{1i} τις παρατηρήσεις του δεύτερου δείγματος θα

πάρουμε τα ακόλουθα:

$$\begin{aligned}\frac{1}{n}\frac{\partial l(\alpha,\beta)}{\partial \alpha} &= \frac{n_1}{n} - \frac{n_1}{n}\sum_{i=1}^{n_0}\frac{e^{\alpha+\beta z_{0i}}}{n_0+n_1e^{\alpha+\beta z_{0i}}} - \frac{n_1}{n}\sum_{i=1}^{n_1}\frac{e^{\alpha+\beta z_{1i}}}{n_0+n_1e^{\alpha+\beta z_{1i}}}\\ &= \frac{n_1}{n} - \frac{n_1}{n}\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{e^{\alpha+\beta z_{0i}}}{1+\frac{n_1}{n_0}e^{\alpha+\beta z_{0i}}} - \frac{n_1}{n}\frac{1}{n_0}\sum_{i=1}^{n_1}\frac{e^{\alpha+\beta z_{1i}}}{1+\frac{n_1}{n_0}e^{\alpha+\beta z_{1i}}}\\ &= \frac{r}{1+r} - \frac{r}{1+r}\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{e^{\alpha+\beta z_{0i}}}{1+re^{\alpha+\beta z_{0i}}} - \frac{r}{1+r}r\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{e^{\alpha+\beta z_{1i}}}{1+re^{\alpha+\beta z_{1i}}}\\ &= \frac{r}{1+r} - \frac{r}{1+r}\Bigg(\underbrace{\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{e^{\alpha+\beta z_{0i}}}{1+re^{\alpha+\beta z_{0i}}}}_{\overline{U}_0}\Bigg) - \frac{r^2}{1+r}\Bigg(\underbrace{\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{e^{\alpha+\beta z_{1i}}}{1+re^{\alpha+\beta z_{1i}}}}_{\overline{U}_1}\Bigg).\end{aligned}$$

Η σχέση αυτή μας δίνει τελικά:

$$\frac{1}{n}\frac{\partial l(\alpha,\beta)}{\partial \alpha} = \frac{r}{1+r}\Big(1-\overline{U}_0 - r\overline{U}_1\Big)$$

Αντίστοιχα, για την παράγωγο ως προς β θα προκύψουν τα εξής:

$$\begin{aligned}\frac{1}{n}\frac{\partial l(\alpha,\beta)}{\partial \beta} &= \frac{n_1}{n}\overline{z}_1 - \frac{n_1}{n}\sum_{i=1}^{n_0}\frac{z_{0i}e^{\alpha+\beta z_{0i}}}{n_0+n_1e^{\alpha+\beta z_{0i}}} - \frac{n_1}{n}\sum_{i=1}^{n_1}\frac{z_{1i}e^{\alpha+\beta z_{1i}}}{n_0+n_1e^{\alpha+\beta z_{1i}}}\\ &= \frac{r}{1+r}\overline{z}_1 - \frac{r}{1+r}\Bigg(\underbrace{\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{z_{0i}e^{\alpha+\beta z_{0i}}}{1+re^{\alpha+\beta z_{0i}}}}_{\overline{V}_0}\Bigg) - \frac{r^2}{1+r}\Bigg(\underbrace{\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{z_{1i}e^{\alpha+\beta z_{1i}}}{1+re^{\alpha+\beta z_{1i}}}}_{\overline{V}_1}\Bigg).\end{aligned}$$

Η σχέση αυτή μας δίνει τελικά:

$$\frac{1}{n}\frac{\partial l(\alpha,\beta)}{\partial \beta} = \frac{r}{1+r}\Big(\overline{Z}_1 - \overline{V}_0 - r\overline{V}_1\Big).$$

Σύμφωνα όμως με το Κεντρικό Οριακό Θεώρημα ισχύει

$$\sqrt{n}\left\{\frac{1}{n}\begin{pmatrix}\frac{\partial l(\alpha,\beta)}{\partial \alpha}\\ \frac{\partial l(\alpha,\beta)}{\partial \beta}\end{pmatrix} - \boldsymbol{\mu}\right\} \xrightarrow{d} N\Big(\mathbf{0},\mathbf{V}\Big),$$

όπου $\boldsymbol{\mu}$ κατάλληλο διάνυσμα μέσων τιμών και $\mathbf{V}$ κατάλληλος πίνακας διασπορών-συνδιασπορών.

Έστω $S_n(\alpha,\beta) = \frac{1}{n}\nabla l(\alpha,\beta)$. Αναπτύσσοντας κατά Taylor ως προς $(\widehat{\alpha},\widehat{\beta})$ γύρω από το (α,β) θα προκύψει ότι

$$S_n(\widehat{\alpha},\widehat{\beta}) = S_n(\alpha,\beta) + \mathbf{H}_n(\alpha^*,\beta^*)\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix} \tag{4.12}$$

για κάποιο (α^*,β^*) ανάμεσα στα $(\widehat{\alpha},\widehat{\beta})$ και (α,β) ενώ ο $\mathbf{H}_n$ είναι ο πίνακας των δεύτερων παραγώγων (Εσσιανός). Λαμβάνοντας υπ' όψιν ότι $S_n(\widehat{\alpha},\widehat{\beta}) = 0$, η (4.12) θα γίνει

$$\sqrt{n}\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix} = \mathbf{H}_n^{-1}(\alpha^*,\beta^*)\sqrt{n}S_n(\alpha,\beta) \Leftrightarrow$$

$$\sqrt{n}\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix} = \mathbf{H}_n^{-1}(\alpha^*,\beta^*)\,\mathbf{H}(\alpha,\beta)\,\mathbf{H}^{-1}(\alpha,\beta)\sqrt{n}S_n(\alpha,\beta).$$

Όμως

$$\mathbf{H}_n^{-1}(\alpha^*,\beta^*)\,\mathbf{H}_n(\alpha,\beta) \xrightarrow{p} \mathbf{I}_2,$$

όπου $\mathbf{I}_2$ είναι ο μοναδιαίος πίνακας τάξης 2, αφού

$$\mathbf{H}_n(\alpha^*,\beta^*) = \frac{1}{n}\nabla^2 l(\alpha^*,\beta^*) \xrightarrow{p} \mathbf{H}(\alpha,\beta).$$

Επίσης ισχύει ότι

$$\mathbf{H}^{-1}(\alpha,\beta)\sqrt{n}S_n(\alpha,\beta) \xrightarrow{d} N\Big(\mathbf{0},\mathbf{H}^{-1}(\alpha,\beta)\,\mathbf{V}(\alpha,\beta)\mathbf{H}^{-1}(\alpha,\beta)\Big).$$

Τελικά προκύπτει

$$\sqrt{n}\begin{pmatrix}\widehat{\alpha}-\alpha\\ \widehat{\beta}-\beta\end{pmatrix} \xrightarrow{d} N(\mathbf{0},\mathbf{H}^{-1}(\alpha,\beta)\,\mathbf{V}(\alpha,\beta)\mathbf{H}^{-1}(\alpha,\beta)). \tag{4.13}$$

4.3.2 Υπολογισμός του πίνακα H

Στην ενότητα αυτή θα υπολογιστούν τα στοιχεία του πίνακα $\mathbf{H}$. Συγκεκριμένα έχουμε

$$\begin{aligned}
\frac{1}{n}\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha^2} &= -\frac{n_1}{n}\sum_{k=1}^{n}\frac{e^{\alpha+\beta z_k}\left(n_0+n_1e^{\alpha+\beta z_k}\right)-n_1e^{\alpha+\beta z_k}e^{\alpha+\beta z_k}}{\left(n_0+n_1e^{\alpha+\beta z_k}\right)^2}\\
&= -\frac{n_1n_0}{n}\sum_{k=1}^{n}\frac{e^{\alpha+\beta z_k}}{\left(n_0+n_1e^{\alpha+\beta z_k}\right)^2}\\
&= -\frac{n_1}{nn_0}\sum_{k=1}^{n}\frac{e^{\alpha+\beta z_k}}{\left(1+re^{\alpha+\beta z_k}\right)^2}\\
&= -\frac{r}{n}\sum_{k=1}^{n}\frac{e^{\alpha+\beta z_k}}{\left(1+re^{\alpha+\beta z_k}\right)^2}\\
&= -r\left(\frac{1}{1+r}\frac{1}{n_0}\sum_{i=1}^{n_0}\frac{e^{\alpha+\beta z_{0i}}}{\left(1+re^{\alpha+\beta z_{0i}}\right)^2}+\frac{r}{1+r}\frac{1}{n_1}\sum_{i=1}^{n_1}\frac{e^{\alpha+\beta z_{1i}}}{\left(1+re^{\alpha+\beta z_{1i}}\right)^2}\right)\\
&\xrightarrow{p} -r\left(\frac{1}{1+r}\mathsf{E}\left[\frac{e^{\alpha+\beta z_0}}{\left(1+re^{\alpha+\beta z_0}\right)^2}\right]+\frac{r}{1+r}\mathsf{E}\left[\frac{e^{\alpha+\beta z_1}}{\left(1+re^{\alpha+\beta z_1}\right)^2}\right]\right)\\
&= -\frac{r}{1+r}\left(\int_0^\infty\frac{e^{\alpha+\beta h(x)}}{\left(1+re^{\alpha+\beta h(x)}\right)^2}f_0(x)dx+r\int_0^\infty\frac{e^{\alpha+\beta h(x)}}{\left(1+re^{\alpha+\beta h(x)}\right)^2}e^{\alpha+\beta h(x)}f_0(x)dx\right)\\
&= -\frac{r}{1+r}\int_0^\infty\frac{e^{\alpha+\beta h(x)}+re^{\alpha+\beta h(x)}e^{\alpha+\beta h(x)}}{\left(1+re^{\alpha+\beta h(x)}\right)^2}f_0(x)dx\\
&= -\frac{r}{1+r}\int_0^\infty\frac{e^{\alpha+\beta h(x)}}{1+re^{\alpha+\beta h(x)}}f_0(x)dx.
\end{aligned}$$

Συνοψίζοντας,

$$\frac{1}{n}\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha^2}\xrightarrow{p}-\frac{r}{1+r}\int_0^\infty\frac{e^{\alpha+\beta h(x)}}{1+re^{\alpha+\beta h(x)}}f_0(x)dx.$$

Με αντίστοιχο τρόπο αποδεικνύεται επίσης ότι

$$\frac{1}{n}\frac{\partial^2 l(\alpha,\beta)}{\partial\beta^2}\xrightarrow{p}-\frac{r}{1+r}\int_0^\infty\frac{h^2(x)e^{\alpha+\beta h(x)}}{1+re^{\alpha+\beta h(x)}}f_0(x)dx$$

και

$$\frac{1}{n}\frac{\partial^2 l(\alpha,\beta)}{\partial\alpha\partial\beta}\xrightarrow{p}-\frac{r}{1+r}\int_0^\infty\frac{h(x)e^{\alpha+\beta h(x)}}{1+re^{\alpha+\beta h(x)}}f_0(x)dx.$$

4.3.3 Η περίπτωση των πολλών δειγμάτων

Κατ' αντιστοιχία με τα δύο δείγματα, ισχύουν ανάλογα συμπεράσματα και στην περίπτωση που το πλήθος των δειγμάτων είναι $s > 2$. Αν $w_j(x) = \exp(\alpha_j + \beta_j h(x))$, τότε καθώς το $n \to \infty$ ισχύει:

$$-\frac{1}{n}\nabla^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta}) \to \mathbf{S}$$

όπου $\mathbf{S}$ είναι ένας $2s \times 2s$ πίνακας με εισόδους στις θέσεις j, j' που ικανοποιούν τις ακόλουθες σχέσεις:

$$\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \alpha_j^2} \xrightarrow{p} -\frac{r_j}{1 + \sum_{k=1}^{s} r_k} \int_0^\infty \frac{\left[1 + \sum_{k \neq j}^{s} r_k w_j(x)\right] w_j(x)}{1 + \sum_{k=1}^{s} r_k w_j(x)} f_0(x) dx$$

$$\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_j^2} \xrightarrow{p} -\frac{r_j}{1 + \sum_{k=1}^{s} r_k} \int_0^\infty \frac{\left[1 + \sum_{k \neq j}^{s} r_k w_j(x)\right] h^2(x) w_j(x)}{1 + \sum_{k=1}^{s} r_k w_j(x)} f_0(x) dx$$

$$\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \alpha_j \partial \beta_j} \xrightarrow{p} -\frac{r_j}{1 + \sum_{k=1}^{s} r_k} \int_0^\infty \frac{\left[1 + \sum_{k \neq j}^{s} r_k w_j(x)\right] h(x) w_j(x)}{1 + \sum_{k=1}^{s} r_k w_j(x)} f_0(x) dx$$

$$\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \alpha_j \partial \alpha_j'} \xrightarrow{p} \frac{r_j r_j'}{1 + \sum_{k=1}^{s} r_k} \int_0^\infty \frac{w_j(x) w_{j'}(x)}{1 + \sum_{k=1}^{s} r_k w_j(x)} f_0(x) dx$$

$$\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \beta_j \partial \beta_j'} \xrightarrow{p} \frac{r_j r_j'}{1 + \sum_{k=1}^{s} r_k} \int_0^\infty \frac{h^2(x) w_j(x) w_{j'}(x)}{1 + \sum_{k=1}^{s} r_k w_j(x)} f_0(x) dx$$

$$\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial \alpha_j \partial \beta_j'} \xrightarrow{p} \frac{r_j r_j'}{1 + \sum_{k=1}^{s} r_k} \int_0^\infty \frac{h(x) w_j(x) w_{j'}(x)}{1 + \sum_{k=1}^{s} r_k w_j(x)} f_0(x) dx$$

Όπως και στην Ενότητα 4.3.1, τελικά αποδεικνύεται ότι

$$\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha} \\ \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \end{pmatrix} \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

όπου $\boldsymbol{\Sigma} = \mathbf{S}^{-1}\mathbf{V}\mathbf{S}^{-1}$ με $\mathbf{V} = \mathsf{Cov}\Big[\frac{1}{\sqrt{n}}\nabla l(\boldsymbol{\alpha}, \boldsymbol{\beta})\Big]$.

Στην περίπτωση που έχουμε δύο δείγματα, οι Qin and Zhang (1997) έδωσαν μία απλούστερη μορφή για τον πίνακα $\boldsymbol{\Sigma}$. Θεώρησαν αρχικά

$$A_k = \int \frac{h^k(x)\exp(\alpha + \beta h(x))}{1 + r\exp(\alpha + \beta h(x))} f_0(x)dx, \quad \text{για } k = 0, 1, 2,$$

όπου $r \equiv r_1$. Επίσης θεώρησαν τον πίνακα

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_0 & \mathbf{A}_1 \\ \mathbf{A}_1 & \mathbf{A}_2 \end{pmatrix}.$$

Έπειτα από πράξεις προκύπτει ότι

$$\boldsymbol{\Sigma} = \mathbf{S}^{-1}\mathbf{V}\mathbf{S}^{-1} = \frac{1+r}{r}\left[\mathbf{A}^{-1} - \begin{pmatrix} 1+r & 0 \\ 0 & 0 \end{pmatrix}\right].$$

4.4 Έλεγχος για την ισότητα των κατανομών

Ένα από τα κυριότερα προβλήματα στην περίπτωση που παρατηρούμε δείγματα από δύο ή περισσότερες κατανομές είναι ο έλεγχος της υπόθεσης ισότητας των κατανομών αυτών. Όπως αναφέρθηκε στην Ενότητα 4.1, στην περίπτωσή μας ο έλεγχος μεταφράζεται σε έλεγχο της υπόθεσης

$$H_0 : \beta_1 = \cdots = \beta_s = 0.$$

Ο έλεγχος μπορεί να βασιστεί στους εκτιμητές $\widehat{\beta}_1, \ldots, \widehat{\beta}_s$ των $\beta_1, \ldots, \beta_s$. Από την (4.13) προκύπτει ότι υπό την H_0

$$\sqrt{n}\,\widehat{\boldsymbol{\beta}} \to N_s(\mathbf{0}, \boldsymbol{\Sigma}_0).$$

Έπειτα από αρκετές πράξεις μπορεί να δειχθεί ότι

$$\boldsymbol{\Sigma}_0 = \frac{1}{\mathsf{Var}_{f_0}[h(Y)]}\mathbf{M}^{-1},$$

όπου $\mathbf{M}$ είναι ο πίνακας με ij στοιχείο

$$\mathbf{M}_{ij} = \frac{r_i\Big\{\delta_{ij}\Big(1+\sum_{k=1}^{s} r_k\Big) - r_j\Big\}}{\Big(1+\sum_{k=1}^{s} r_k\Big)^2},$$

όπου το δ_{ij} είναι το δέλτα του Kronecker. Η H_0 απορρίπτεται για μεγάλες τιμές της στατιστικής συνάρτησης

$$T = n\,\widehat{\beta}^T\,\widehat{\mathbf{\Sigma}}_0^{-1}\,\widehat{\beta},$$

όπου $\widehat{\mathbf{\Sigma}}_0$ είναι ένας συνεπής εκτιμητής του $\mathbf{\Sigma}_0$. Στην πραγματικότητα, το μόνο που χρειάζεται να εκτιμηθεί είναι η $\mathsf{Var}_{f_0}[h(Y)]$ μια που ο πίνακας $\mathbf{M}$ δεν εξαρτάται από τις παραμέτρους. Η διασπορά αυτή μπορεί να εκτιμηθεί είτε από όλα τα δεδομένα (μια που υπό την H_0 προέρχονται όλα από την F_0) είτε μόνο από το δείγμα που γνωρίζουμε ότι προέρχεται από τη βασική κατανομή.

Υπό την H_0 η T έχει ασυμπτωτική κατανομή χ^2_s. Επομένως η H_0 απορρίπτεται σε επίπεδο σημαντικότητας α αν $T > \chi^2_{s,\alpha}$ όπου $\chi^2_{s,\alpha}$ είναι το α-άνω ποσοστιαίο σημείο της κατανομής χ^2_s.

4.5 Παραδείγματα

Στην ενότητα αυτή θα παρουσιαστούν κάποια αποτελέσματα όπως προκύπτουν βάσει προσομοιωμένων δειγμάτων. Θεωρούμε λοιπόν ότι η κατανομή αναφοράς είναι η f_0 ενώ f_i είναι όλες οι υπόλοιπες κατανομές.

4.5.1 Ομοιόμορφη κατανομή

Στην πρώτη περίπτωση θα χρησιμοποιηθούν τρία δείγματα τα οποία προέρχονται από την ομοιόμορφη κατανομή στο $(0,1)$. Επομένως $f_0(x) = f_1(x) = f_2(x) = 1$ για $x \in (0,1)$. Χρησιμοποιώντας το μοντέλο (4.3) (με $h(y) = y$) θα εκτιμηθούν οι άγνωστες παράμετροι. Βάσει αυτού του μοντέλου ισχύει $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = 0$. Για να

εκτιμηθούν οι παράμετροι αυτές θα χρησιμοποιηθούν προσομοιωμένα δείγματα. Συγκεκριμένα, προσομοιώνουμε τρία ανεξάρτητα τυχαία δείγματα από την ομοιόμορφη κατανομή στο $(0,1)$ και χρησιμοποιώντας την μέθοδο Monte Carlo για 500 επαναλήψεις, εκτιμώνται οι άγνωστες παράμετροι. Επίσης εκτιμώνται και τα αντίστοιχα τυπικά σφάλματα. Αυτό που έχει σημασία κυρίως είναι οι τιμές των παραμέτρων β, καθώς $\beta_1 = \beta_2 = 0$ σημαίνει ότι οι τρεις κατανομές δεν διαφέρουν. Στον Πίνακα 4.1 παρουσιάζονται τα αποτελέσματα της προσομοίωσης για διαφορετικά μεγέθη δειγμάτων.

Πίνακας 4.1: Τρία δείγματα από την ομοιόμορφη κατανομή στο $(0,1)$ $(h(y) = y)$

Μεγέθη δειγμάτων			**Παράμετροι**				**Εκτιμήσεις**			
n_0	n_1	n_2	α_1	α_2	β_1	β_2	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
200	200	200	0	0	0	0	0.003	-0.004	-0.006	0.009
							(0.181)	(0.183)	(0.363)	(0.366)
100	100	100	0	0	0	0	-0.019	0.039	-0.003	0.005
							(0.244)	(0.486)	(0.24)	(0.495)

Από τον παραπάνω πίνακα είναι σαφές ότι οι τιμές των $\widehat{\beta}$ είναι στατιστικά μη σημαντικές, κάτι το οποίο μας οδηγεί στο συμπέρασμα ότι οι κατανομές είναι ίδιες. Αυτό είναι κάτι αναμενόμενο.

4.5.2 Κανονική κατανομή

Θεωρούμε κατ' αρχάς ότι η βασική κατανομή είναι κανονική κατανομή με μέση τιμή 0 και τυπική απόκλιση 1 ενώ οι δύο άλλες κατανομές έχουν την ίδια διακύμανση αλλά διαφορετικές μέσες τιμές (2 και 3 αντίστοιχα). Όπως και στο προηγούμενο παράδειγμα, η εκτίμηση των παραμέτρων του μοντέλου (4.3) (με $h(y) = y$) θα γίνει με την μέθοδο Monte Carlo για 500 επαναλήψεις. Συγκεκριμένα προσομοιώνονται τρία ανεξάρτητα τυχαία δείγματα από τις κατανομές $N(0,1)$, $N(2,1)$ και $N(3,1)$ και εκτιμώνται οι παράμετροι α και β καθώς και τα τυπικά τους σφάλματα. Τα αποτελέσματα

της προσομοίωσης σε αυτή την περίπτωση παρουσιάζονται στον Πίνακα 4.2. Να σημειωθεί ότι οι εκτιμήσεις αυτές είναι πολύ κοντά με τις πραγματικές, ενώ σύμφωνα με τις τιμές των τυπικών αποκλίσεων απορρίπτεται η υπόθεση $\beta_i = 0$, δηλαδή ότι οι πληθυσμοί είναι ίδιοι. Να παρατηρηθεί στο σημείο αυτό ότι η αλλαγή των μεγεθών των δειγμάτων δεν επηρεάζει πολύ τις εκτιμήσεις των παραμέτρων.

Πίνακας 4.2: $f_0(y) \sim N(0,1)$, $f_1(y) \sim N(2,1)$ και $f_2(y) \sim N(3,1)$ $(h(y) = y)$

Μεγέθη δειγμάτων			**Παράμετροι**				**Εκτιμήσεις**			
n_0	n_1	n_2	α_1	α_2	β_1	β_2	$\widehat{\alpha}_1$	$\widehat{\alpha}_2$	$\widehat{\beta}_1$	$\widehat{\beta}_2$
200	200	200	-4.5	-2	3	2	-4.561	-2.052	3.052	2.048
							(0.387)	(0.23)	(0.236)	(0.197)
200	300	100	-4.5	-2	3	2	-4.548	-2.014	3.021	2.01
							(0.424)	(0.193)	(0.229)	(0.176)

4.5.3 Εκθετική κατανομή

Στο παράδειγμα αυτό υποθέτουμε ότι έχουμε στην διάθεσή μας δύο δείγματα τα οποία προέρχονται από εκθετικές κατανομές με διαφορετική παράμετρο. Έστω ότι το πρώτο δείγμα προέρχεται από εκθετική κατανομή με παράμετρο 1 ενώ το δεύτερο δείγμα από εκθετική με παράμετρο 2. Τότε είναι εύκολο να διαπιστωθεί ότι οι πραγματικές τιμές των παραμέτρων κατά την εφαρμογή του μοντέλου (4.3) (με $h(y) = y$) είναι $\alpha_1 = \log(2) = 0.693$ και $\beta_1 = -1$. Για την εκτίμηση των παραμέτρων αυτών προσομοιώνουμε δύο ανεξάρτητα τυχαία δείγματα από εκθετική κατανομή με παραμέτρους 1 και 2 και εφαρμόζουμε την μέθοδο Monte Carlo για 500 επαναλήψεις. Τα αποτελέσματα της προσομοίωσης παρουσιάζονται στον Πίνακα 4.3. Σύμφωνα με τις εκτιμήσεις των παραμέτρων και των τυπικών σφαλμάτων απορρίπτεται η υπόθεση $\beta_1 = 0$, κάτι το οποίο σημαίνει ότι οι δύο πληθυσμοί δεν είναι ίδιοι.

Πίνακας 4.3: $f_0(y) \sim E(1)$ και $f_1(y) \sim E(2)$ $(h(y) = y)$

Μεγέθη δειγμάτων		**Παράμετροι**		**Εκτιμήσεις**	
n_0	n_1	α_1	β_1	$\widehat{\alpha}_1$	$\widehat{\beta}_1$
200	200	0.693	-1	0.697	-1.014
				(0.119)	(0.185)
300	200	0.693	-1	0.692	-1.003
				(0.101)	(0.164)

4.5.4 Παράδειγμα με πραγματικά δεδομένα

Στο αυτό το παράδειγμα, θα παρουσιαστεί η χρησιμότητα του μεροληπτικού μοντέλου

$$f_j(y) = \exp\left(\alpha_j + \beta_j h(y)\right) f_0(y), \quad j = 1, 2, 3, 4 \tag{4.14}$$

χρησιμοποιώντας ένα γνωστό σύνολο δεδομένων, όταν $h(y) = \log(y)$. Πρόκειται για μετρήσεις σε κρανία από Αιγυπτίους άνδρες σε πέντε διαφορετικές χρονικές περιόδους που εκτείνονται από το 4000 π.Χ. έως το 150 μ.Χ. Σε κάθε μία χρονική περίοδο μελετήθηκαν οι ακόλουθες 4 μεταβλητές

MB μέγιστο πλάτος κρανίου (Maximal Breadth of Skull)

BH Basibregmatic Height of Skull

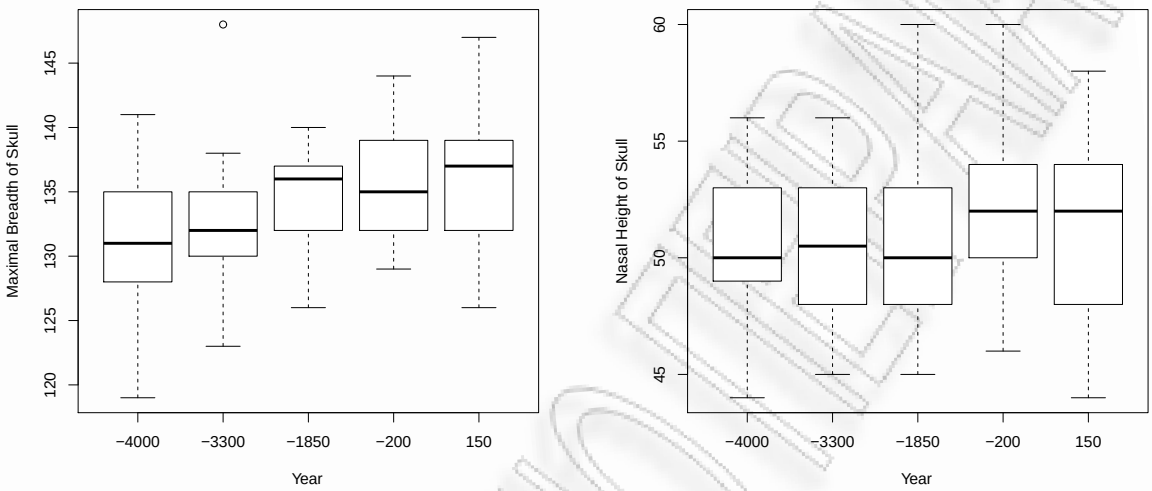
BL μήκος του κρανίου (Basialveolar Length of Skull)

NH ρινικό ύψος κρανίου (Nasal Height of Skull)

Σκοπός είναι να διαπιστωθεί αν υπάρχει κάποια αλλαγή στο μέγεθος του κρανίου σε διαφορετικές χρονικές περιόδους. Αν και η ανάλυση αυτών των δεδομένων θεωρείται ένα πρόβλημα πολυμεταβλητής ανάλυσης, εμείς θα επικεντρωθούμε σε δύο μόνο μεταβλητές, δηλαδή το μέγιστο πλάτος και το ρινικό ύψος του κρανίου. Στο σχήμα που ακολουθεί παρουσιάζονται τα θηκογράμματα αυτών των δύο μεταβλητών.

Από αυτά φαίνεται να υπάρχει κάποια αύξηση και των δύο μεγεθών καθώς περνάει ο χρόνος. Ωστόσο, για την δεύτερη μεταβλητή το μέγεθος της αύξησης φαίνεται λιγότερο σημαντικό. Αυτά τα γεγονότα πρέπει να επιβεβαιωθούν και με την χρήση

Πίνακας 4.4: Θηκογράμματα των μεταβλητών MB και NH



του μεροληπτικού μοντέλου. Για την εφαρμογή του μοντέλου (4.14) θεωρούμε ότι η $f_0(\cdot)$ είναι η πυκνότητα πιθανότητας της κατανομής για το έτος 4000 π.Χ. Αντίστοιχα συμβολίζονται με $f_j(\cdot),\ j = 1, 2, 3, 4$ οι πυκνότητες πιθανότητας για τα έτη 3300 π.Χ., 1850 π.Χ., 200 π.Χ. και 150 μ.Χ.. Οι παράμετροι που έχουν το πιο σημαντικό ρόλο είναι οι $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T$.

Συγκεκριμένα, για την μεταβλητή MB, οι εκτιμηθείσες τιμές των παραμέτρων $\boldsymbol{\beta}$ είναι $\widehat{\boldsymbol{\beta}} = (6.3702, 20.4540, 27.3862, 31.5796)^T$ κάτι το οποίο δείχνει την αύξηση στο μέγιστο πλάτος κρανίου με το πέρασμα των χρόνων. Αντίστοιχα για την μεταβλητή NH, έχουμε $\widehat{\boldsymbol{\beta}} = (-1.6021, -0.0589, 7.4395, 3.9992)^T$. Από τις τιμές αυτές φαίνεται να υπάρχει μία αύξηση στις τιμές των παραμέτρων αλλά η τελευταία παράμετρος δεν αυξάνεται αλλά μειώνεται.

Αυτό που έχει σημασία όμως είναι η διαφορά των κατανομών. Για το λόγο αυτό θα γίνει ο έλεγχος της υπόθεσης

$$H_0 :\ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0.$$

Σύμφωνα με την Ενότητα 4.4 προκύπτει ότι οι τιμές της στατιστικής συνάρτησης T είναι 33.9 και 4.97 για τις μεταβλητές MB και NH αντίστοιχα ενώ οι τιμές p είναι 10^{-6} και 0.29. Βάσει των τιμών p οδηγούμαστε στο συμπέρασμα ότι η υπόθεση $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ απορρίπτεται σε επίπεδο σημαντικότητας 5% για την μεταβλητή MB, κάτι το οποίο σημαίνει ότι οι κατανομές διαφέρουν. Αντίθετα, στην περίπτωση της μεταβλητής NH, δεν μπορούμε να απορρίψουμε την υπόθεση $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ σε επίπεδο σημαντικότητας 5%.

4.6 Η επιλογή της συνάρτησης h(y)

Σε όλα τα παραδείγματα που μελετήθηκαν στην Ενότητα 4.5, δεν αναφερθήκαμε καθόλου στην επιλογή της συνάρτησης $h(y)$. Πρέπει να τονιστεί ότι, γενικά, η συνάρτηση αυτή είναι άγνωστη και επιλέγεται αυθαίρετα από εμάς. Στην πραγματικότητα παίζει πολύ σημαντικό ρόλο στην ανάλυση και κακή επιλογή της μπορεί να οδηγήσει σε τελείως λανθασμένα συμπεράσματα. Ας δούμε το παρακάτω παράδειγμα.

Παράδειγμα 1. Ας υποθέσουμε ότι έχουμε στην διάθεσή μας δύο δείγματα τα οποία προέρχονται από κανονικές κατανομές. Συγκεκριμένα έστω $f_0(y) \sim N(0, \sigma_0^2)$, ενώ $f_1(y) \sim N(0, \sigma_1^2)$. Τότε

$$\begin{aligned}
\frac{f_1(y)}{f_0(y)} &= \frac{\frac{1}{\sigma_1\sqrt{2\pi}}\exp\{-y^2/2\sigma_1^2\}}{\frac{1}{\sigma_0\sqrt{2\pi}}\exp\{-y^2/2\sigma_0^2\}} \\
&= \frac{\sigma_0}{\sigma_1}\exp\Big\{-\frac{y^2}{2\sigma_1^2}+\frac{y^2}{2\sigma_0^2}\Big\} \\
&= \exp\Big\{\log\Big(\frac{\sigma_0}{\sigma_1}\Big)+\frac{1}{2}\Big(\frac{1}{\sigma_0^2}-\frac{1}{\sigma_1^2}\Big)y^2\Big\} \\
&= \exp\{\alpha_1+\beta_1 y^2\},
\end{aligned}$$

όπου $\alpha_1 = \log(\sigma_0/\sigma_1)$ και $\beta_1 = (1/\sigma_0^2 - 1/\sigma_1^2)/2$. Αυτό σημαίνει ότι $h(y) = y^2$.

Για $\sigma_0 = 4$ και $\sigma_1 = 2$, θα προσομοιώσουμε δύο δείγματα, ένα από την βασική κατανομή και ένα από την μεροληπτική εκδοχή της και με την μέθοδο Monte Carlo

για 500 επαναλήψεις θα εκτιμήσουμε τις παραμέτρους α_1 και β_1. Τα αποτελέσματα της προσομοίωσης παρουσιάζονται στον ακόλουθο πίνακα:

Πίνακας 4.5: $f_0(y) \sim N(0,4)$ και $f_1(y) \sim N(0,2)$, $(h(y) = y^2)$

Μεγέθη δειγμάτων		**Παράμετροι**		**Εκτιμήσεις**	
n_0	n_1	α_1	β_1	$\widehat{\alpha}_1$	$\widehat{\beta}_1$
300	200	0.693	-0.094	0.694	-0.095
				(0.089)	(0.013)

Ένα 95% διάστημα εμπιστοσύνης για την παράμετρο β_1 είναι το $(-0.122, -0.069)$ το οποίο δεν συμπεριλαμβάνει την τιμή 0 και αυτό μας οδηγεί στο συμπέρασμα ότι οι δύο κατανομές διαφέρουν.

Ας υποθέσουμε τώρα ότι χρησιμοποιείται λανθασμένη συνάρτηση h, π.χ. η $h(y) = y$. Με αυτήν την h, εφαρμόζοντας την ίδια μέθοδο, οι εκτιμήσεις των παραμέτρων φαίνονται στον Πίνακα 4.6.

Πίνακας 4.6: $f_0(y) \sim N(0,4)$ και $f_1(y) \sim N(0,2)$, $(h(y) = y)$

Μεγέθη δειγμάτων		**Παράμετροι**		**Εκτιμήσεις**	
n_0	n_1	α_1	β_1	$\widehat{\alpha}_1$	$\widehat{\beta}_1$
300	200	0.693	-0.094	0.0037	-0.0002
				(0.007)	(0.036)

Με βάση τον πίνακα, ένα 95% διάστημα εμπιστοσύνης για την παράμετρο β_1 είναι το $(-0.071, 0.069)$ το οποίο συμπεριλαμβάνει την τιμή 0, και αυτό μας οδηγεί στο συμπέρασμα ότι οι δύο πληθυσμοί δεν διαφέρουν σε επίπεδο σημαντικότητας 5%. Άρα η λανθασμένη επιλογή της συνάρτησης h επηρέασε την απόφαση μας για το αν διαφέρουν ή όχι οι δύο κατανομές.

Αυτό επομένως που είναι απαραίτητο να γίνει, είναι ένας έλεγχος καλής προσαρμογής ώστε να αποφασιστεί κατά πόσο η επιλογή της συνάρτησης h είναι καλή. Ένας τέτοιος έλεγχος προτάθηκε από τον Zhang (2002) και βασίζεται στη μέθοδο bootstrap.

Έστω

$$F_{in_i}(x) = \frac{1}{n_i}\sum_{j=1}^{n_i} I_{(-\infty,x]}(y_{ij}), \quad i = 0, 1, \ldots, s, \tag{4.15}$$

οι εμπειρικές συναρτήσεις κατανομής που βασίζονται στα $s+1$ δείγματα. Έστω επίσης

$$\Delta_{in}(x) = n^{1/2}\{\widehat{F}_i(x) - F_{in_i}(x)\}, \quad \Delta_{in} = \sup_x |\Delta_{in}(x)|, \tag{4.16}$$

όπου $\widehat{F}_i$ είναι η ημιπαραμετρική εκτίμηση της F_i βάσει του μοντέλου (4.3). Ορίζουμε την στατιστική συνάρτηση

$$\Delta_n = \frac{1}{s+1}\sum_{i=0}^{s} r_i \Delta_{in}, \tag{4.17}$$

όπου $r_i = n_i/n_0,\ i = 0, \ldots, s$. Στη συνέχεια εφαρμόζουμε τη μέθοδο bootstrap στα δεδομένα μας. Για την προσομοίωση ενός δείγματος bootstrap από την F_i χρησιμοποιείται η εκτίμησή της, $\widehat{F}_i$, στην (4.11). Η διαδικασία επαναλαμβάνεται B φορές.

Ως bootstrap τιμή p ορίζεται η ποσότητα

$$\frac{1}{B}\sum_{b=1}^{B} I(\Delta_n > \Delta_n^{(b)}),$$

όπου $\Delta_n^{(1)}, \ldots, \Delta_n^{(B)}$ είναι οι τιμές του Δ_n από τα δείγματα bootstrap. Με την μεθοδολογία αυτή θα αποφασιστεί αν το μοντέλο προσαρμόζεται στα δεδομένα για την συγκεκριμένη συνάρτηση h.

Παράδειγμα 1. (συνέχεια)

Στο παράδειγμα 1 φάνηκε πόσο επηρεάζει η συνάρτηση h την απόφασή μας σχετικά με το αν οι κατανομές διαφέρουν μεταξύ τους. Για το λόγο αυτό θα χρησιμοποιηθεί ο έλεγχος καλής προσαρμογής του Zhang ώστε να αποφασιστεί κατά πόσο η επιλογή της συνάρτησης αυτής είναι καλή. Έτσι θα γίνουν τα ακόλουθα βήματα:

1. Παραγωγή δύο τυχαίων δείγματων μεγέθους $n_0 = 300$ και $n_1 = 200$ από την βασική και μεροληπτική κατανομή αντίστοιχα.

2. Βάσει αυτών των δειγμάτων εκτιμώνται οι τιμές των παραμέτρων α_1 και β_1 μία φορά με $h(x) = x^2$ και άλλη μία με $h(x) = x$.

3. Πραγματοποιείται ο έλεγχος καλής προσαρμογής και υπολογίζονται δύο τιμές p, κάθε μία για διαφορετική συνάρτηση h.

4. Επαναλαμβάνονται τα βήματα 1 έως 3 πολλές φορές και υπολογίζεται το ποσοστό των τιμών p του βήματος 3 που ξεπέρασαν την τιμή α, έστω 10%.

Τα αποτελέσματα των προσομοιώσεων έδωσαν ποσοστά 0.34 και 0.04, σύμφωνα με τα οποία φαίνεται ξεκάθαρα ότι η επιλογή της συνάρτησης $h(x) = x^2$ είναι καλή ενώ αντίθετα η επιλογή της συνάρτησης $h(x) = x$ δεν είναι καλή.

4.6.1 Παράδειγμα με πραγματικά δεδομένα (συνέχεια)

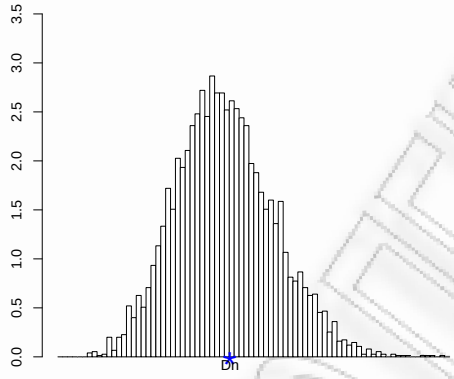
Κάνοντας τον έλεγχο καλής προσαρμογής προκύπτει ο ακόλουθος πίνακας

Πίνακας 4.7: Έλεγχος καλής προσαρμογής

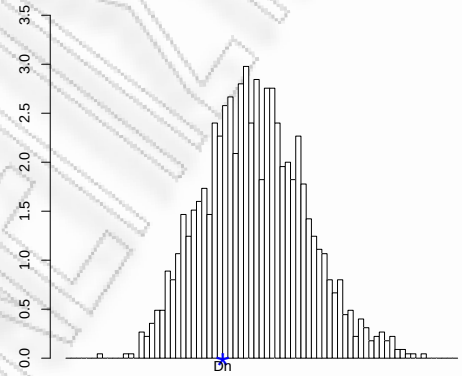
	Δ_n	τιμή p
Μέγιστο πλάτος κρανίου	1.0269	0.44
Ρινικό ύψος κρανίου	0.8693	0.72

Να παρατηρήσουμε ότι και στις δύο περιπτώσεις οι τιμές p είναι πολύ μεγάλες και αυτό υποδεικνύει ότι το μοντέλο (4.14) είναι κατάλληλο για την ανάλυση των δύο μεταβλητών ατομικά.

Τέλος δινονται δύο σχήματα όπου παρουσιάζεται η κατανομή της ποσότητας Δ_n όπως προέκυψε από την διαδικασία του ελέγχου καλής προσαρμογής.



Σχήμα 4.1: Κατανομή της Δ_n για την μεταβλητή BM και παρατηρηθείσα τιμή της.



Σχήμα 4.2: Κατανομή της Δ_n για την μεταβλητή NH και παρατηρηθείσα τιμή της.

Βιβλιογραφία

[1] Arratia, R. and Goldstein, L. 2009. Size bias, sampling, the waiting time paradox, and infnite divisibility: when is the increment independent? arXiv: 1007.3910 (July 2010).

[2] Buckland, S.T., Anderson, D.R., Burnham, K.P. and Laake, J.L. 1993. *Distance Sampling: Estimating Abundance of Biological Populations*. Chapman and Hall, London.

[3] Cox, D. R. 1969. Some sampling problems in technology. In N. L. Johnson & H. Smith (Eds.), New developments in survey sampling (506-527). John Wiley & Sons, New York.

[4] Davidov, O., Iliopoulos, G. 2003. A note on an iterative algorithm for nonparametric estimation in biased sampling models. *Computational Statistics and Data Analysis*, 54, 620 – 624.

[5] Davidov, O. and Iliopoulos, G. 2008. On the existence and uniqueness of the NPMLE in biased sampling models. *Journal of Statistical Planning and Inference*. 139, 176 – 183.

[6] Davidov, O., Fokianos, K., Iliopoulos, G. 2009. Order-Restricted Semiparametric Inference for the Power Bias Model. *Biometrics*, 66, 549 – 557.

[7] Fisher, R. A. 1934. The effect of methods of ascertainment upon the estimation of frequencies. *Annals of Eugenics*, 6, 13 – 25.

[8] Gill, R.D., Vardi, Y., Wellner, J.A. 1988. Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*. 16, 1069 – 1112.

[9] Liang, S. 2005. Generalized linear and additive models with weighted distribution. (Thesis, M.Sc., National University Of Singapore).

[10] Owen, A.B. 2001, *Empirical Likelihood*. Chapman & Hall/CRC.

[11] Patil, G.P., Rao, C.R., Ratnaparkhic, M.V. 1986. On discrete weighted distributions and their use in model choice for observed data. *Communications in Statistics - Theory and Methods*, 15, 907 – 918.

[12] Qin, J. and Zhang, B. 1997, A Goodness of fit test for logistic regression models based on case control data. *Biometrica*, 84, 609 – 618.

[13] Rao, C. 1965. *Linear Statistical Inference and its Applications*. Wiley & Sons, Canada.

[14] Safranyik L. and Linton D.A. 2002. Line transect sampling to estimate the density of lodgepole pine currently attacked by mountain pine beetle. *Pacific Forestry Centre*.

[15] Sen, P.K. 1987. What do the arithmetic, geometric and harmonic means tell us in length-biased sampling. *Statistical and Probability Letters*, 5, 95 – 98.

[16] Shao, J. 2003. *Mathematical Statistics*, 2nd Edition. Springer, New York.

[17] Vardi, Y. 1982 Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10, 616 – 620.

[18] Vardi, Y. 1985. Empirical distributions in selection bias models. *The Annals of Statistics*. 13, 178 – 203.