ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΞΟΜΑΛΥΝΣΗ ΠΙΝΑΚΩΝ ΘΝΗΣΙΜΟΤΗΤΑΣ

Αθανάσιος Π. Σαχλάς

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
Μάιος 2004

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

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΕΞΟΜΑΛΥΝΣΗ ΠΙΝΑΚΩΝ ΘΝΗΣΙΜΟΤΗΤΑΣ

Αθανάσιος Π. Σαχλάς

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
Μάιος 2004

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:

- Παπαϊωάννου Π. (Επιβλέπων)
- Πιτσέλης Γ.
- Φερεντίνος Κ.

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

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

GRADUATION OF MORTALITY TABLES

By

Athanasios P. Sachlas

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
May 2004

Περίληψη

Ένα από τα σημαντικότερα αντικείμενα της αναλογιστικής επιστήμης, είναι η κατασκευή μοντέλων θνησιμότητας τα οποία περιγράφουν το πραγματικό πρότυπο θνησιμότητας ενός πληθυσμού. Για να το επιτύχουμε αυτό, υπολογίζουμε από πρωτογενή δεδομένα τις αρχικές εκτιμήσεις των ποσοστών θνησιμότητας οι οποίες παρουσιάζονται σε μορφή πινάκων και ονομάζονται πίνακες θνησιμότητας. Επειδή οι αρχικές εκτιμήσεις των ποσοστών θνησιμότητας αποτελούν συνήθως μια μη ομαλή – μη κανονική (*irregular*) σειρά, συνηθίζεται να αναθεωρούμε τα αρχικά δεδομένα με σκοπό να πάρουμε ομαλότερες εκτιμήσεις. Η διαδικασία αναθεώρησης των αρχικών εκτιμήσεων, η οποία ονομάζεται *εξομάλυνση (graduation)* και οι μέθοδοι με τις οποίες αυτή επιτυγχάνεται θα παρουσιαστούν σε αυτή τη διπλωματική εργασία. Συγκεκριμένα παρουσιάζεται η εξομάλυνση με χρήση των μοντέλων θνησιμότητας, των γενικευμένων γραμμικών μοντέλων, των συναρτήσεων splines και της παρεμβολής ομαλής σύνδεσης (*smooth – junction interpolation*), που αποτελούν τις παραμετρικές μεθόδους εξομάλυνσης. Επίσης παρουσιάζονται οι μη παραμετρικές μέθοδοι, οι οποίες είναι η γραφική μέθοδος, η εξομάλυνση με αναφορά σε τυπικό πίνακα θνησιμότητας, η εξομάλυνση μέσω κινητών σταθμισμένων μέσων, η εξομάλυνση των Whittaker και Henderson, η μπεϋζιανή εξομάλυνση, η εξομάλυνση μέσω εκτιμητών πυρήνα και η εξομάλυνση με χρήση των εννοιών της θεωρίας πληροφοριών. Για κάθε μια από αυτές παρουσιάζεται η βασική ιδέα και μεθοδολογία για την εφαρμογή της. Επίσης αναφέρονται περιπτώσεις εξομάλυνσης και άλλων ασφαλιστικών ποσοτήτων, όπως είναι η ένταση θνησιμότητας, ο αριθμός των θανάτων ή ο χρόνος έκθεσης στον κίνδυνο του θανάτου. Τέλος δίνεται η εφαρμογή σε δεδομένα της μεθόδου των Whittaker – Henderson, της μεθόδου των splines και της μεθόδου της θεωρίας πληροφοριών, επιχειρείται μια κριτική των μεθόδων και διατυπώνονται κάποια ανοικτά ερωτήματα για την εξομάλυνση.

Abstract

One of the more important tasks of the actuarial science is the construction of mortality models, which describe the actual mortality pattern of a population. In order to achieve this we calculate from raw data the initial estimates of mortality rates that are displayed in tabular form and are called mortality tables. Because the initial estimates of the mortality rates usually form an irregular series, it is common to revise the initial data with the aim of producing smoother estimates. The procedure of revising the initial estimates, which is called *graduation*, and the methods of which graduation can be obtained will be presented in this thesis. Particularly it is presented graduation by means of mortality laws, generalized linear models, splines and smooth – junction interpolation, which are the parametric methods of graduation. It is also presented the non-parametric methods, which are the graphic method, the graduation with reference to a standard mortality table, the moving weighted average graduation, the Whittaker and Henderson graduation, the Bayesian graduation, the graduation by means of kernel estimators and the graduation with use of concepts from information theory. For each of them it is presented the basic idea and methodology for their application. The graduation of other insurance values such as force of mortality, number of deaths or exposure time to the risk of death will also be discussed. An illustration of the method of Whittaker – Henderson, the splines' method and information theory's method is also be given, it is attempted a critique of the methods and some open questions for graduations are also be stated.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

ΕΙΣΑΓΩΓΗ

1.1 Πίνακες θνησιμότητας – Πίνακες επιβίωσης - Ένταση θνησιμότητας

Ένα από τα βασικότερα θέματα της αναλογιστικής επιστήμης είναι η κατασκευή μοντέλων επιβίωσης (ή θνησιμότητας) τα οποία περιγράφουν το πραγματικό πρότυπο θνησιμότητας κάποιου πληθυσμού και πάνω σε αυτά στηρίζεται ο υπολογισμός διαφόρων ασφαλιστικών ποσοτήτων όπως ασφάλιστρα, αποθεματικά κ.α. Για την περιγραφή του προτύπου θνησιμότητας του πληθυσμού που εξετάζουμε, υπολογίζονται τα ποσοστά θνησιμότητας, τα οποία συνήθως παρουσιάζονται σε μορφή πινάκων που ονομάζονται πίνακες θνησιμότητας ή επιβίωσης.

Πίνακας θνησιμότητας (*mortality table*) είναι μια σειρά από ετήσιες συνήθως πιθανότητες θανάτου από μια ελάχιστη ηλικία a και πέρα, δηλαδή

$$q_a\ ,\ q_{a+1},\ldots,\ q_{w-1}\ ,$$

όπου με w συμβολίζουμε την οριακή ηλικία και εννοούμε την ηλικία πέρα από την οποία κανένα άτομο δεν επιζεί. Με άλλα λόγια, q_x είναι η πιθανότητα ένα άτομο ηλικίας x να πεθάνει στο διάστημα $[x, x+1)$ για $x = a, a+1,..., w-1$. Οι συμπληρωματικές πιθανότητες των παραπάνω πιθανοτήτων είναι οι

$$p_a\ ,\ p_{a+1},\ldots,\ p_{w-1}\ ,$$

που δηλώνουν την πιθανότητα ένα άτομο ηλικίας x να επιβιώσει μέχρι την ηλικία $x+1$. Σημειώνουμε ότι μεταξύ των πιθανοτήτων θανάτου και επιβίωσης ισχύει η σχέση $p_x + q_x = 1$, $x = a, a+1,..., w-1$.

Όσον αφορά τον *πίνακα επιβίωσης* (*life table*), ο ορισμός του δεν είναι ξεκάθαρος στην βιβλιογραφία. Ορισμένοι συγγραφείς, όπως ο London (1985), αναφέρουν τον όρο πίνακα επιβίωσης ως εναλλακτική ονομασία του πίνακα θνησιμότητας. Άλλοι συγγραφείς, από την πλευρά της ανάλυσης επιβίωσης, ορίζουν ως πίνακα επιβίωσης τον πίνακα στον οποίο

εμφανίζονται οι τιμές της συνάρτησης επιβίωσης, δηλαδή την πιθανότητα ένα άτομο ηλικίας 0 να φθάσει την ηλικία x, $x = a, a+1,..., w-1$ (Johnson and Johnson, 1980).

Υποτυπώδεις πίνακες θνησιμότητας άρχισαν να κατασκευάζονται από τα μέσα του $17^{ου}$ αιώνα. Πραγματικοί όμως πίνακες κατασκευάστηκαν από τον J. Graunt το 1662 και τον E. Halley ένα χρόνο αργότερα. Όμως ως πρωτοπόρος στην κατασκευή πινάκων θνησιμότητας θεωρείται ο Halley, επειδή χρησιμοποίησε στατιστικά δεδομένα για την κατανομή των θανάτων ανάλογα με την ηλικία.

Οι πίνακες θνησιμότητας και επιβίωσης έχουν αρκετές πρακτικές εφαρμογές. Οι Cox and Oakes (1984) αναφέρουν ότι οι πίνακες επιβίωσης χρησιμοποιούνται συνήθως για τη μη παραμετρική εκτίμηση της συνάρτησης επιβίωσης όταν έχουμε λογοκριμένα δεδομένα. Οι πίνακες θνησιμότητας χρησιμοποιούνται στην ανάλυση της θνησιμότητας του πληθυσμού στον οποίο αναφέρονται, καθώς και στην εκτίμηση της επιβίωσης του ίδιου πληθυσμού, αφού θνησιμότητα και επιβίωση συνδέονται άμεσα. Επίσης μπορούν να χρησιμοποιηθούν για τη σύγκριση της θνησιμότητας ή επιβίωσης μεταξύ δυο διαφορετικών πληθυσμών. Στη βιοστατιστική και την ιατρική επιστήμη χρησιμοποιούνται για την περιγραφή της κατάστασης της υγείας του πληθυσμού και των βιολογικών ορίων της ανθρώπινης ζωής. Τέλος στην αναλογιστική επιστήμη χρησιμοποιούνται ευρύτατα για τον υπολογισμό ασφαλίστρων, ραντών, μαθηματικών και τεχνικών αποθεμάτων κ.α. ασφαλιστικών ποσοτήτων (Pagano and Gauvreau, 2000).

Υπάρχουν δυο βασικοί τύποι πινάκων θνησιμότητας, οι *πλήρεις* και οι *συμπυκνωμένοι (abridged)*. Στους πρώτους, η πινακοποίηση των δεδομένων γίνεται για κάθε ηλικία x, $x = 1, 2,..., w$ ενώ στους δεύτερους, η πινακοποίηση γίνεται για κάθε n έτη, έχουμε δηλαδή διαστήματα ηλικιών, συνήθως πενταετούς διάρκειας.

Ένας πίνακας θνησιμότητας, είτε είναι πλήρης είτε συμπυκνωμένος αποτελείται από ορισμένες βασικές συναρτήσεις (Johnson and Johnson, 1980). Αυτές είναι οι l_x, d_x, q_x και p_x, $x = 1, 2,..., w$.

Με l_x συμβολίζουμε το πλήθος των ατόμων που βρίσκονται στη ζωή στην ηλικία x. Συνήθως ξεκινάμε από την ηλικία 0 και παίρνουμε $l_0 = 100.000$ (ή οποιαδήποτε άλλη δύναμη του 10) και καταλήγουμε στην οριακή ηλικία w. Ο αριθμός των θανάτων στο διάστημα $[x, x+1)$ συμβολίζεται με d_x και ισούται με

$$d_x = l_x - l_{x+1}.$$

Με q_x συμβολίζεται η δεσμευμένη πιθανότητα θανάτου στο ίδιο διάστημα, δηλαδή η πιθανότητα ένα άτομο ηλικίας x να πεθάνει στο διάστημα $[x, x+1)$ και δίνεται από τη σχέση

$$q_x = \frac{d_x}{l_x} = \frac{l_x - l_{x+1}}{l_x}.$$

Η παραπάνω ποσότητα, συχνά συμβολίζεται και με $\overset{0}{q}_x$, το οποίο αναφέρεται στη βιβλιογραφία ως *αδρός δείκτης θνησιμότητας*.

Η δεσμευμένη πιθανότητα επιβίωσης p_x, δηλαδή η πιθανότητα ένα άτομο ηλικίας x να μην πεθάνει στο διάστημα $[x, x+1)$ δίνεται από τη σχέση

$$p_x = 1 - q_x = \frac{l_{x+1}}{l_x},\ x = 0, 1, ..., w-1.$$

Είναι εύκολο να δειχθεί με αναγωγικό τρόπο ότι ισχύει

$$l_x = l_0 p_0 p_1 ... p_{x-1}.$$

Το ποσοστό των επιζώντων ατόμων στην ηλικία x υπολογίζεται ως

$$P_x = \frac{l_x}{l_0} = \prod_{i=0}^{x-1} p_i$$

και ισχύει $P_0 = 1$.

Η πιθανότητα ένα άτομο ηλικίας x να επιζήσει για n επιπλέον έτη, δίνεται από τη σχέση ${}_n p_x = \frac{l_{x+n}}{l_x}$.

Επιπλέον συναρτήσεις που συνήθως παρουσιάζονται σε πίνακες θνησιμότητας είναι οι παρακάτω: Η L_x που συμβολίζει το συνολικό χρόνο σε έτη που έχουν ζήσει στο διάστημα $[x, x+1)$, τα l_x άτομα που ήταν στη ζωή στην αρχή του διαστήματος. Κάθε άτομο που επιβιώνει στο διάστημα αυτό συμμετέχει στο L_x με ένα ολόκληρο έτος ενώ κάθε άτομο που αποβιώνει συνεισφέρει ένα μόνο ποσοστό του έτους. Για τη συνάρτηση L_x ισχύει, αν θεωρήσουμε το l_x ως συνεχή συνάρτηση,

$$L_x = \int_x^{x+1} l_y dy.$$

Με T_x συμβολίζουμε τον αναμενόμενο συνολικό αριθμό των ετών που έχουν επιζήσει l_x άτομα ηλικίας x μετά από αυτή την ηλικία. Ισχύει

$$T_x = L_x + L_{x+1} + ... + L_w = \sum_{i=0}^{w-x-1} L_{x+i} \ .$$

Τέλος εμφανίζεται η συνάρτηση $\overset{0}{e}_x$ που συμβολίζει τον αριθμό των ετών που αναμένεται να ζήσει ένα άτομο ηλικίας x. Η τιμή της συνάρτησης αυτής δίνεται από τον τύπο

$$\overset{0}{e}_x = \frac{T_x}{l_x}.$$

Σε περίπτωση που έχουμε συμπυκνωμένο πίνακα, με διαστήματα ηλικιών $[x, x+n)$, όλες οι παραπάνω συναρτήσεις συμβολίζονται όπως πριν αλλά με το δείκτη n κάτω αριστερά. Για παράδειγμα η πιθανότητα θανάτου στο διάστημα $[x, x+n)$, για ένα άτομο ηλικίας x, συμβολίζεται με ${}_n q_x$.

Υπάρχουν διάφοροι τρόποι κατασκευής πινάκων θνησιμότητας. Στη συνέχεια θα αναφερθούμε σε δυο από αυτούς: ο ένας χρησιμοποιεί τους πίνακες επιβίωσης ενώ ο άλλος είναι στατιστικός τρόπος και είναι γνωστός ως μέθοδος του Halley.

i) Πίνακες επιβίωσης

Στην περίπτωση αυτή υποθέτουμε ότι έχουμε μια ομάδα 100.000 βρεφών που γεννήθηκαν την ίδια ημέρα, όπου για ευκολία θεωρούμε ότι γεννήθηκαν την πρώτη ημέρα του έτους. Επιπλέον, υποθέτουμε ότι η ομάδα αυτή είναι κλειστή, με την έννοια ότι η ομάδα δεν υφίσταται καμία αλλαγή παρά μόνο τη μείωση των μελών της που οφείλεται αποκλειστικά στο θάνατο.

Συμβολίζουμε με l_0 τον αρχικό πληθυσμό (δηλαδή 100.000 άτομα) της ομάδας, με l_1 το σύνολο των ατόμων που συμπληρώνουν το πρώτο έτος της ζωής τους και γενικά με l_x, όπως ήδη έχουμε δει, συμβολίζουμε τα άτομα που έφθασαν στην ηλικία x. Με l_w συμβολίζουμε το σύνολο των ατόμων που έφθασαν την οριακή ηλικία w. Παίρνουμε δηλαδή τη σειρά:

$$l_0 \ , l_1 ,..., \ l_x ,..., \ l_w \ .$$

Θεωρητικά, αυτή η μέθοδος κατασκευής πινάκων είναι πολύ απλή αφού το μόνο που χρειάζεται να γίνει είναι να παρακολουθήσουμε μια γενιά ατόμων από τη στιγμή της γέννησής τους μέχρι την ηλικία w και να καταγράψουμε τον αριθμό των ατόμων σε κάθε ηλικία. Στη συνέχεια, και αφού υπολογίσουμε τον αριθμό των θανάτων σε κάθε ηλικία x, ως $d_x = l_x - l_{x+1}$, υπολογίζουμε τα ποσοστά θνησιμότητας από τη σχέση $q_x = \frac{d_x}{l_x}$, $x = 1, 2,..., w$,

τα οποία μπορούν να συμβολισθούν και με $\overset{0}{q}_x$. Αντιλαμβανόμαστε όμως ότι η παραπάνω μέθοδος είναι πρακτικά δύσκολη έως και αδύνατη, αφού απαιτείται η συλλογή στοιχείων για περισσότερα από 100 χρόνια. Επιπλέον, τα στοιχεία αυτά δεν θα έχουν πρακτικό ενδιαφέρον αφού τα δημογραφικά χαρακτηριστικά του πληθυσμού θα έχουν αλλάξει και συνεπώς αυτά θα αναφέρονται στη θνησιμότητα της προηγούμενης εκατονταετίας.

ii) Μέθοδος του Halley

Ο αστρονόμος Halley το 1663 υπολόγισε την πιθανότητα θανάτου βάσει στατιστικών στοιχείων και με την υπόθεση ότι ο πληθυσμός είναι στάσιμος. Δηλαδή ο Halley υπέθεσε ότι ο αριθμός των γεννήσεων ενός έτους είναι ίσος με τον αριθμό των θανάτων που συμβαίνουν το ίδιο έτος.

Με την επιπλέον υπόθεση του κλειστού πληθυσμού, είναι εύκολο να βρεθεί η ακολουθία l_x των ατόμων που επιζούν σε διάφορες ηλικίες αφού είναι γνωστός ο αριθμός των θανάτων d_x στις διάφορες ηλικίες. Να σημειώσουμε ότι στους πίνακες επιβίωσης που είδαμε πριν, γνωρίζαμε τον αριθμό των επιζώντων σε κάθε ηλικία και βάσει αυτών υπολογίζαμε τον αριθμό των θανάτων ενώ εδώ συμβαίνει το αντίθετο. Πιο συγκεκριμένα, αν $l_0 = m$ και d_0 ο αριθμός των θανάτων στην ηλικία 0, τότε θα ισχύει $l_1 = l_0 - d_0$. Με τον ίδιο τρόπο υπολογίζεται η αντίστοιχη ποσότητα και για τις υπόλοιπες ηλικίες. Γενικά θα ισχύει

$$l_x = l_{x-1} - d_{x-1},\ x = 1, 2, ..., w\ .$$

Στη συνέχεια η πιθανότητα θανάτου στην ηλικία x υπολογίζεται από τον τύπο

$$q_x = \frac{d_x}{l_x},$$

η οποία επίσης μπορεί να συμβολισθεί με $\overset{0}{q}_x$.

Η παραπάνω μέθοδος, θα είχε πρακτική αξία αν όντως υπήρχε σταθερός νόμος θνησιμότητας, αν ο πληθυσμός ήταν πραγματικά κλειστός και αν ο αριθμός των γεννήσεων παρέμενε σταθερός.

1.1.1 Πίνακες ανάλυσης επιβίωσης

Μέχρι τώρα υποθέσαμε ότι παρατηρούμε όλα τα άτομα που έχουμε στη διάθεσή μας μέχρι τη στιγμή του θανάτου τους. Αυτό δεν είναι και τόσο ρεαλιστικό σενάριο καθώς κάποια

άτομα μπορεί να αποχωρήσουν από την έρευνα είτε λόγω μετανάστευσης είτε γιατί δεν θέλουν άλλο να συμμετέχουν σε αυτή. Υπεισέρχεται λοιπόν η έννοια της *λογοκρισίας*.

Έστω ότι έχουμε ένα τυχαίο δείγμα χρόνων ζωής μεγέθους n από έναν πληθυσμό, το οποίο περιέχει πλήρη και λογοκριμένα δεδομένα. Συνήθως οι πίνακες επιβίωσης και θνησιμότητας περιέχουν ετήσιες πιθανότητες επιβίωσης και θνησιμότητας αντίστοιχα. Στην περίπτωση όμως της ανάλυσης επιβίωσης το μέγεθος του τυχαίου δείγματος n είναι σχετικά μικρό, πράγμα που καθιστά αδύνατο τον υπολογισμό ετήσιων πιθανοτήτων.

Συνεπώς τα δεδομένα παρουσιάζονται σε «πίνακες συχνοτήτων» με $k+1$ διαστήματα ηλικιών της μορφής $I_j = [x_{j-1}, x_j)$, $j = 1, 2, ..., k+1$ με $x_0 = 0$, $x_k = w$, όπου w είναι η οριακή ηλικία και $x_{k+1} = \infty$. Για το I_j διάστημα έχουμε τα εξής στοιχεία:

- d_j είναι ο αριθμός των θανάτων που συμβαίνουν στο διάστημα $I_j = [x_{j-1}, x_j)$. Με άλλα λόγια είναι το σύνολο των πλήρων χρόνων ζωής.
- w_j είναι ο αριθμός των διαφυγών, δηλαδή των ατόμων που χάθηκαν από την έρευνα στο διάστημα I_j για λόγο διαφορετικό του θανάτου. Στην ουσία το w_j παριστάνει τους λογοκριμένους χρόνους ζωής.
- l_j είναι ο αριθμός των ατόμων που βρίσκονται σε κίνδυνο στην αρχή του διαστήματος I_j. Είναι δηλαδή τα άτομα των οποίων οι χρόνοι ζωής είναι δυνατόν να αποτύχουν (το άτομο πεθαίνει) ή να λογοκριθούν το χρόνο x_{j-1} ή μεταγενέστερο. Βέβαια ο πραγματικός αριθμός των ατόμων που βρίσκονται στη ζωή το χρόνο x_{j-1} είναι άγνωστος και μπορεί να είναι και μεγαλύτερος του l_j αφού στον πραγματικό αριθμό των ατόμων μπορεί να συμπεριλαμβάνονται και άτομα των οποίων οι χρόνοι ζωής λογοκρίθηκαν πριν τη χρονική στιγμή x_{j-1}. Αξίζει να σημειωθεί ότι το τελευταίο διάστημα I_{k+1} περιέχει μόνο πλήρη δεδομένα αφού πέρα από την οριακή ηλικία κανένα άτομο δεν επιζεί.

Σύμφωνα με τα παραπάνω έχουμε ότι $l_1 = n$ και $l_j = l_{j-1} - d_{j-1} - w_{j-1}$, $j = 1, 2, ..., k+1$.

Στην περίπτωση που δεν έχουμε λογοκριμένα δεδομένα, η πιθανότητα θανάτου στην ηλικία j, ενός ατόμου που έχει επιβιώσει μέχρι την ηλικία $j-1$, υπολογίζεται από τη σχέση

$$q_j = \frac{d_j}{l_j}.$$

Εάν όμως έχουμε και διαφυγές (λογοκριμένα δεδομένα), η παραπάνω εκτίμηση θα υποεκτιμά την πραγματική δεσμευμένη πιθανότητα θανάτου αφού οι χρόνοι ζωής κάποιων

ατόμων που έχουν διαφύγει θα μπορούσαν να έχουν πεθάνει εντός του διαστήματος I_j. Το πρόβλημα αυτό μπορεί να λυθεί, χρησιμοποιώντας την *αναλογιστική υπόθεση*. Η υπόθεση αυτή αναφέρει ότι τη χρονική στιγμή x_{j-1}, σε κίνδυνο βρίσκονται μόνο οι μισές από τις διαφυγές του διαστήματος I_j. Έτσι ο «πραγματικός» αριθμός ατόμων σε κίνδυνο τη στιγμή x_{j-1} είναι

$$l'_j = l_j - \frac{w_j}{2}.$$

Με βάση τα παραπάνω, η δεσμευμένη πιθανότητα θανάτου q_j δίνεται από τη σχέση

$$q_j = \frac{d_j}{l'_j} = \frac{d_j}{l_j - \frac{1}{2}w_j}$$

και η δεσμευμένη πιθανότητα επιβίωσης δίνεται από τον τύπο

$$p_j = 1 - q_j = \frac{l'_j - d_j}{l'_j}.$$

Ο παραπάνω εκτιμητής, ο οποίος ονομάζεται *αναλογιστικός εκτιμητής*, χρησιμοποιείται κυρίως σε μελέτες ανάλυσης επιβίωσης όπου έχουν μικρή διάρκεια και όχι αρκετά μεγάλο μέγεθος δείγματος. Στη συνέχεια, η συνάρτηση επιβίωσης, που θα δούμε καλύτερα στην επόμενη παράγραφο, υπολογίζεται ως

$$S(x_j) = \prod_j p_j = \prod_j (1 - q_j),\ j = 1, 2, ..., k+1.$$

(Johnson and Johnson, 1980).

1.1.2 Ασφαλιστικοί πίνακες θνησιμότητας

Ενώ οι πίνακες θνησιμότητας περιγράφουν τη θνησιμότητα ενός γενικού πληθυσμού, οι ασφαλιστικοί πίνακες καταρτίζονται με βάση στατιστικά στοιχεία που αναφέρονται σε έναν ασφαλισμένο ή μια ομάδα ασφαλισμένων. Είναι κατανοητό ότι η ομάδα αυτή έχει διαφορετική θνησιμότητα από το γενικό πληθυσμό. Στη χώρα μας, σύμφωνα με υπουργική απόφαση (ΦΕΚ 847), από 31 Δεκεμβρίου 1999, όλες οι ασφαλιστικές εταιρείες, ελληνικές ή μη, υποχρεούνται να χρησιμοποιούν για την τιμολόγηση των κινδύνων και τον υπολογισμό των μαθηματικών αποθεμάτων τους τον πίνακα θνησιμότητας 1990 της Ένωσης Ελλήνων Αναλογιστών ο οποίος προέκυψε με βάση τα στοιχεία επιβίωσης του 1990 της Εθνικής Στατιστικής Υπηρεσίας της Ελλάδος. Για ασφαλίσεις που συνάπτονται σε χώρες εκτός

ΟΟΣΑ, μπορούν να χρησιμοποιηθούν οι επίσημοι ή ειδικά προσαρμοσμένοι πίνακες των χωρών αυτών. Για τα ομαδικά συνταξιοδοτικά σχήματα ορισμένων παροχών χρησιμοποιείται ο Ελβετικός πίνακας EVK 90 ή ο ελληνικός πίνακας θνησιμότητας 1990.

Οι ασφαλιστικοί πίνακες θνησιμότητας, διακρίνονται σε τρεις κατηγορίες: (i) *γενικοί πίνακες*, (ii) *ειδικοί πίνακες* και (iii) *μικτοί πίνακες*. Οι γενικοί πίνακες θνησιμότητας συντάσσονται με βάση μόνο την ηλικία των ασφαλισμένων. Οι ειδικοί πίνακες, συντάσσονται βάσει διαφόρων χαρακτηριστικών όπως η ηλικία ή ο χρόνος παραμονής των ασφαλισμένων στο πρόγραμμα ασφάλισης, ενώ οι μικτοί πίνακες λαμβάνουν υπόψη τους την ηλικία και το χρόνο παραμονής των ασφαλισμένων στην ασφάλιση, ο οποίος κυμαίνεται μεταξύ 5 και 10 ετών και διαφέρει από εταιρεία σε εταιρεία ανάλογα με το πόση σημασία δίνει η κάθε μια στις ιατρικές εξετάσεις.

Κλείνοντας το θέμα της κατασκευής των πινάκων θνησιμότητας, μπορούμε να πούμε ότι υπάρχουν τρία διαφορετικά δειγματοληπτικά πλαίσια για την κατασκευή τους.

Το πρώτο είναι να επιλέξουμε μια κοορτή, δηλαδή μια γενιά ατόμων που έχουν γεννηθεί την ίδια χρονιά και να καταγράψουμε τον αριθμό των θανάτων και επιζώντων σε κάθε ηλικία, μέχρι να πεθάνει και το τελευταίο μέλος της κοορτής. Στη συνέχεια, γνωρίζοντας τον αριθμό των επιζώντων και θανάτων σε κάθε ηλικία μπορούμε να υπολογίσουμε τα αντίστοιχα ποσοστά θνησιμότητας. Να τονίσουμε όμως ότι η διαδικασία αυτή απαιτεί εκατό και πλέον χρόνια για να ολοκληρωθεί.

Το δεύτερο πλαίσιο είναι να βασιστούμε στα δεδομένα που συλλέγονται εντός ενός χρονικού διαστήματος, που για παράδειγμα στην Αγγλία είναι τέσσερα χρόνια. Στην περίπτωση αυτή αντί μιας κοορτής, παρακολουθούμε μια σειρά από κοορτές. Αυτό βέβαια μπορεί να σημαίνει ότι δεν δειγματοληπτούμε από την ίδια κατανομή, γι' αυτό μπορεί να εισαχθούν υποθέσεις όπως για παράδειγμα ότι οι πιθανότητες επιβίωσης είναι σταθερές στη διάρκεια κάθε έτους. Να τονίσουμε ότι στο δειγματοληπτικό αυτό σχέδιο υπεισέρχεται η έννοια της λογοκρισίας. Με το σχέδιο αυτό υποθέτουμε ότι έχουμε n άτομα ηλικίας x έως $x+1$ ετών. Για το i άτομο, θέτουμε $x+a_i$ την ηλικία που αυτό αρχίζει να παρατηρείται και $x+b_i$ την ηλικία που η έρευνα πρέπει να σταματήσει αν το άτομο βρίσκεται στη ζωή. Τότε $b_i = \min\{1, a_i + c_i - e_i\}$ όπου e_i και c_i είναι οι ημερομηνίες εισόδου και εξόδου από την έρευνα. Τα a_i, e_i, c_i και b_i είναι γνωστά εκ των προτέρων. Ορίζουμε στη συνέχεια τις τυχαίες μεταβλητές

$$D_i = \begin{cases} 1, \text{ αν το } i \text{ άτομο πεθάνει} \\ \\ 0, \text{ αν το } i \text{ άτομο δεν πεθάνει} \end{cases}$$

και T_i έτσι ώστε $x + T_i$ είναι η ηλικία στην οποία σταματά η παρακολούθηση του i ατόμου και $W_i = T_i - a_i$ η οποία συμβολίζει το χρόνο υπό παρακολούθηση του i ατόμου. Μπορούμε να γράψουμε

$$D_i = \begin{cases} 0, \text{ αν και μόνο αν } W_i = b_i - a_i \\ \\ 1, \text{ αν και μόνο αν } 0 < W_i < b_i - a_i . \end{cases}$$

Αν ορίσουμε $w = \sum_{i=1}^{n} w_i$ και $d = \sum_{i=1}^{n} d_i$, όπου w_i και d_i είναι οι τιμές των τυχαίων μεταβλητών W_i και D_i αντίστοιχα, έχουμε το συνολικό χρόνο έκθεσης και το συνολικό αριθμό θανάτων αντίστοιχα για τη συγκεκριμένη ηλικία x. Τότε μπορεί να εκτιμηθεί η ένταση θνησιμότητας ως $m = \frac{d}{w}$ (Haberman, 1998).

Το τρίτο δειγματοληπτικό πλαίσιο είναι να πάρουμε τα δεδομένα από μια επίσημη υπηρεσία όπως είναι η στατιστική υπηρεσία. Τα δεδομένα αυτά αναφέρονται σε ένα έτος. Για παράδειγμα έχουμε ήδη αναφέρει τον πίνακα θνησιμότητας 1990 της Ένωσης Ελλήνων Αναλογιστών. Στην περίπτωση αυτή, η στατιστική υπηρεσία έχει τον αριθμό των θανάτων για κάθε ηλικία που συνέβησαν το συγκεκριμένο έτος καθώς επίσης γνωρίζει και τον πληθυσμό που υπήρχε σε κάθε ηλικία για το ίδιο έτος. Στη συνέχεια, διαιρώντας τις δυο αυτές ποσότητες προκύπτει η πιθανότητα θανάτου σε κάθε ηλικία. Αν ο πίνακας είναι συμπυκνωμένος, η παραπάνω πιθανότητα που αντιστοιχεί σε κάθε έτος του διαστήματος, πολλαπλασιάζεται με το μήκος του διαστήματος για να προκύψει η πιθανότητα θανάτου για ολόκληρο το διάστημα (Pagano and Gauvreau, 2000).

1.2 Βασικές έννοιες ανάλυσης επιβίωσης

Έστω ότι ο χρόνος ζωής ενός ατόμου περιγράφεται από μια μη αρνητική μεταβλητή T, η οποία μπορεί να είναι είτε συνεχής είτε διακριτή και το σύνολο των τιμών της είναι υποσύνολο του $[0,\infty)$. Έστω $F(x)$ η αθροιστική συνάρτηση κατανομής της T, η οποία δίνεται από τον τύπο

$$F(x) = P(T \le x),\ x \ge 0.$$

Η $F(x)$ είναι αύξουσα και συνεχής από δεξιά συνάρτηση. Επιπλέον η $F(x)$ καθορίζει πλήρως τη συμπεριφορά της τυχαίας μεταβλητής T.

Αν η T είναι συνεχής, η συνάρτηση πυκνότητας πιθανότητάς της δίνεται από τη σχέση

$$f(x) = \frac{\partial F(x)}{\partial x} = \lim_{\Delta x \to 0^+} \frac{P(x \le T < x + \Delta x)}{\Delta x},$$

ενώ αν είναι διακριτή, για τη συνάρτηση πιθανότητάς της ισχύει

$$f(x_j) = F(x_j) - F(x_{j-1}),\ j = 1, 2, ...$$

και $F(x_0) = 0$.

Η πιθανότητα του ενδεχομένου $\{T \ge x\},\ x \ge 0$, δηλαδή ένα άτομο να επιζήσει πέραν της ηλικίας x, καλείται *συνάρτηση επιβίωσης* (*survival function*) της T, συμβολίζεται με $S(x)$ και δίνεται από τη σχέση

$$S(x) = P(T \ge x),\ x \ge 0.$$

Η $S(x)$ είναι φθίνουσα και συνεχής συνάρτηση για την οποία ισχύουν

$$S(x) = P(T \ge x) = 1 - F(x),\ S(0) = 1 \text{ και } \lim_{x \to \infty} S(x) = 0.$$

Μια από τις βασικότερες συναρτήσεις τόσο στην ανάλυση επιβίωσης όσο και στη δημογραφία, είναι η *συνάρτηση κινδύνου* (*hazard function*) της συνεχούς τυχαίας μεταβλητής T. Συμβολίζεται με $h(x)$ και δίνεται από τη σχέση

$$h(x) = \lim_{\Delta x \to 0^+} \frac{P(x \le T < x + \Delta x \mid T \ge x)}{\Delta x},\ x \ge 0.$$

Η συνάρτηση κινδύνου μπορεί να ερμηνευθεί ως η στιγμιαία πιθανότητα θανάτου ενός ατόμου τη χρονική στιγμή x δεδομένου ότι αυτό έχει επιζήσει ως τη χρονική στιγμή x. Εναλλακτικά μπορούμε να πούμε ότι η ποσότητα $h(x)\Delta x$ είναι η πιθανότητα θανάτου στο διάστημα $[x, x + \Delta x)$ δοθέντος ότι το άτομο βρίσκεται στη ζωή τη χρονική στιγμή x.

Σημειώνουμε ότι στη δημογραφία και την αναλογιστική επιστήμη, η συνάρτηση $h(x)$ ονομάζεται *ένταση θνησιμότητας* (*force of mortality*) και συμβολίζεται με m_x, συμβολισμό τον οποίο θα χρησιμοποιήσουμε και εμείς στη συνέχεια της διπλωματικής εργασίας.

Τέλος υπάρχει η αθροιστική συνάρτηση κινδύνου, η οποία συμβολίζεται με $H(x)$, δίνεται από τον τύπο

$$H(x) = \int_0^x h(u)du\,,\ x \geq 0$$

και συνδέεται με τη συνάρτηση επιβίωσης μέσω της σχέσης

$$H(x) = -\log S(x)\,.$$

Ας θεωρήσουμε τώρα την τυχαία μεταβλητή T_x που συμβολίζει τον υπολειπόμενο χρόνο ζωής ενός ατόμου, δοθέντος ότι το άτομο έχει επιβιώσει μέχρι την ηλικία x. Δηλαδή ισχύει $T_x = T - x \,|\, T > x$. Η συνάρτηση κατανομής της τυχαίας μεταβλητής T_x ορίζεται ως

$$F_x(t) = P(T_x \leq t) = P(T \leq x + t \,|\, T > x)\,,$$

η οποία στην αναλογιστική επιστήμη συμβολίζεται με ${}_tq_x$. Η συνάρτηση επιβίωσης της T_x είναι

$$S_x(t) = P(T_x > t) = P(T > x + t \,|\, T > x)$$

και συμβολίζεται με ${}_tp_x$. Η σχέση που συνδέει τις συναρτήσεις κατανομής των T και T_x, είναι η

$$F_x(t) = \frac{F(x+t) - F(x)}{1 - F(x)}$$

ενώ η αντίστοιχη σχέση για τις συναρτήσεις επιβίωσης είναι η

$$S_x(t) = \frac{S(x+t)}{S(x)}\,.$$

Η συνάρτηση πυκνότητας πιθανότητας της T_x δίνεται ως

$$f_x(t) = \frac{\partial}{\partial t} F_x(t)$$

και συνδέεται με την ένταση θνησιμότητας μέσω της σχέσης

$$f_x(t) = {}_tp_x m_{x+t}\,,$$

οπότε ισχύει

$$m_{x+t} = \frac{f_x(t)}{{}_tp_x} = \frac{f_x(t)}{S_x(t)}\,.$$

Από την παραπάνω σχέση, προκύπτει η διαφορική εξίσωση για το ${}_t p_x$,

$$\frac{\partial}{\partial t}{}_t p_x = -{}_t p_x m_{x+t},$$

την οποία αν ολοκληρώσουμε με τη συνθήκη ${}_0 p_x = 1$, προκύπτει η πολύ χρήσιμη για την αναλογιστική επιστήμη σχέση που συνδέει τη δεσμευμένη πιθανότητα επιβίωσης με την ένταση θνησιμότητας. Ισχύει δηλαδή

$${}_t p_x = \exp\left\{-\int_0^t m_{x+u}\,du\right\}$$

ή ισοδύναμα

$${}_t q_x = 1 - \exp\left\{-\int_0^t m_{x+u}\,du\right\}$$

(Haberman, 1998).

Για περαιτέρω σχέσεις που συνδέουν τις παραπάνω ποσότητες, ο αναγνώστης παραπέμπεται στον (Johnson and Johnson, 1980).

1.3 Τι είναι εξομάλυνση

Ας ξαναγυρίσουμε τώρα στον τρόπο υπολογισμού των ποσοστών θνησιμότητας. Στην ουσία δεν υπολογίζουμε τα πραγματικά ποσοστά θνησιμότητας, αλλά απλά τα εκτιμούμε. Γνωρίζοντας τον αριθμό των θανάτων στην ηλικία x, έστω d_x, και τον αριθμό των ατόμων που βρίσκονται σε κίνδυνο στην ίδια ηλικία, έστω l_x, ο πιο απλός τρόπος εκτίμησης του ποσοστού θνησιμότητας γίνεται μέσω του τύπου

$$\overset{\circ}{q}_x = \frac{d_x}{l_x},\ x = 1, 2, \ldots, n\,.$$

Το ποσοστό $\overset{\circ}{q}_x$ αναφέρεται στη βιβλιογραφία ως *αδρός δείκτης θνησιμότητας* στην ηλικία x. Με άλλα λόγια μπορούμε να πούμε ότι ο αδρός δείκτης θνησιμότητας είναι η ανάλογη έννοια της έντασης θνησιμότητας, στην περίπτωση όπου η τυχαία μεταβλητή T που περιγράφει το χρόνο ζωής ενός ατόμου είναι διακριτή.

Επειδή συνήθως υποθέτουμε ότι το l_x είναι μεγάλο, συγκρινόμενο με το d_x, μπορούμε να υποθέσουμε ότι το d_x ακολουθεί τη διωνυμική κατανομή με παραμέτρους l_x και q_x και συνεπώς ο εκτιμητής $\overset{0}{q}_x$ μπορεί να θεωρηθεί ως διωνυμικό ποσοστό. Έτσι μπορούμε να γράψουμε

$$E\left(\overset{0}{q}_x\right) = q_x \text{ και } \text{var}\left(\overset{0}{q}_x\right) = \frac{q_x(1-q_x)}{l_x},$$

όπου q_x είναι το πραγματικό ποσοστό θνησιμότητας. Βέβαια ο αριθμός των θανάτων δεν ακολουθεί ακριβώς διωνυμική κατανομή, καθώς το l_x δεν αντιστοιχεί ακριβώς σε αριθμό ανεξάρτητων διωνυμικών δοκιμών (London, 1985). Επιπλέον, αν παραστήσουμε γραφικά τα $\overset{0}{q}_x$, $x = 1, 2, ..., n$, τα οποία, υπό την υπόθεση της διωνυμικής κατανομής, είναι και εκτιμητές μέγιστης πιθανοφάνειας, θα πάρουμε ένα διάγραμμα με μεγάλες διακυμάνσεις, καθώς κάποια από αυτά θα είναι μεγαλύτερα και κάποια μικρότερα από τα πραγματικά ποσοστά θνησιμότητας του πληθυσμού στις αντίστοιχες ηλικίες. Το επόμενο λοιπόν βήμα της κατασκευής των μοντέλων θνησιμότητας, είναι η αναθεώρηση των αρχικών εκτιμήσεων $\overset{0}{q}_x$ για να προκύψουν καλύτερες εκτιμήσεις των πραγματικών ποσοστών. Η αναθεώρηση αυτή ονομάζεται *εξομάλυνση* (*graduation*) και με τους τρόπους με τους οποίους αυτή επιτυγχάνεται θα ασχοληθούμε σε αυτή τη διπλωματική εργασία.

Ο Miller (1949) όρισε την εξομάλυνση ως τη διαδικασία να πάρουμε, από μια μη ομαλή – μη κανονική (*irregular*) σειρά παρατηρούμενων τιμών μιας συνεχούς μεταβλητής, μια ομαλή σειρά τιμών οι οποίες να είναι «συνεπείς» με τις παρατηρούμενες τιμές. Ένα ερώτημα που γεννιέται είναι αν η εξομάλυνση μπορεί να γίνει σε οποιαδήποτε σειρά παρατηρούμενων τιμών. Σύμφωνα με τον London (1985), η απάντηση δίνεται από τα ίδια τα δεδομένα και όλες οι σειρές δεν ενδείκνυνται για εξομάλυνση. Συγκεκριμένα αναφέρει ότι «μόνο ορισμένοι τύποι δεδομένων είναι κατάλληλοι για εξομάλυνση, δηλαδή αυτοί για τους οποίους πιστεύεται ότι υπάρχει σχέση μεταξύ των στοιχείων της σειράς των δεδομένων». Ο Haberman (1998), για το ίδιο θέμα, αναφέρει ότι αν ο ερευνητής πιστεύει ότι τα δεδομένα είναι ανεξάρτητα μεταξύ τους, τότε οι αδρές τιμές θα είναι οι τελικές εκτιμήσεις των πραγματικών δεδομένων. Σε αντίθετη περίπτωση, πρέπει να γίνει εξομάλυνση.

Στην αναλογιστική επιστήμη, εξομάλυνση μπορεί να γίνει είτε σε εκτιμήσεις των ποσοστών θνησιμότητας q_x είτε σε εκτιμήσεις της έντασης θνησιμότητας m_x, χωρίς να

αποκλείονται ποσότητες όπως ο αριθμός ατόμων σε κίνδυνο, συνταξιοδοτικά ποσά, ποσοστά αποχωρήσεων και άλλες ασφαλιστικές ποσότητες. Ο Haberman (1998) αναφέρει ότι υπάρχει μια εκ των προτέρων άποψη, ότι κάθε πραγματική τιμή του q_x (ή του m_x) είναι στενά συσχετισμένη με την επόμενή της και η σχέση αυτή εκφράζεται από την άποψη ότι οι τιμές των ποσοτήτων αυτών αυξάνονται ομαλά από ηλικία σε ηλικία. Συνεπώς χρειάζεται να γίνει εξομάλυνση των αρχικών εκτιμήσεων των ποσοστών θνησιμότητας ή της έντασης θνησιμότητας.

Ας δούμε τώρα πώς μπορούμε να εκτιμήσουμε την ένταση θνησιμότητας. Για να το επιτύχουμε αυτό θα στηριχθούμε στην υπόθεση ότι οι θάνατοι d_x είναι ανεξάρτητοι μεταξύ τους και κατανέμονται ως Poisson κατανομή με παράμετρο $r_x^c m_x$, όπου r_x^c είναι ο κεντρικός (*central*) χρόνος έκθεσης στον κίνδυνο του θανάτου. Η υπόθεση της κατανομής Poisson, δικαιολογείται με το ακόλουθο σκεπτικό: Υποθέτουμε ότι ένα άτομο μπορεί να πεθάνει ανά πάσα στιγμή μέσα στο χρονικό διάστημα $\left(0, r_x^c\right)$, το οποίο διαιρούμε σε $n = \dfrac{r_x^c}{\Delta x}$ υποδιαστήματα μήκους Δx. Υποθέτουμε ότι η πιθανότητα p να πεθάνει ένα άτομο σε οποιοδήποτε υποδιάστημα $(x, x+\Delta x)$, $0 \le x \le r_x^c$ είναι $m_x \Delta x$, η οποία είναι σταθερή σε κάθε υποδιάστημα. Αν X είναι ο αριθμός των «επιτυχιών», δηλαδή των θανάτων στο $\left(0, r_x^c\right)$ και υποθέσουμε ότι οι θάνατοι είναι ανεξάρτητοι από υποδιάστημα σε υποδιάστημα και ότι η πιθανότητα δυο ή περισσότερων θανάτων στο $(x, x+\Delta x)$ είναι αμελητέα λόγω του μικρού Δx, τότε πρακτικά μπορούμε να θεωρήσουμε ότι η τυχαία μεταβλητή X ακολουθεί τη διωνυμική κατανομή με παραμέτρους n και p, όπου $n = \dfrac{r_x^c}{\Delta x}$ και $p = m_x \Delta x$.

Αν $\Delta x \to 0$ τότε $n \to +\infty$ και $p \to 0$. Αλλά για σταθερό $np = m_x r_x^c$, έχουμε ότι

$$
\begin{aligned}
P(X = x) &= \lim_{n\to\infty} \binom{n}{x} p^x q^{n-x} \\
&= \lim_{n\to\infty} \frac{n(n-1)\ldots(n-x-1)}{x!} \left(\frac{np}{n}\right)^x \left(1 - \frac{np}{n}\right)^{n-x} \\
&= \frac{(np)^x}{x!} \lim_{n\to\infty} \frac{n(n-1)\ldots(n-x-1)}{n \cdot n \cdot \ldots \cdot n} \left(1 - \frac{np}{n}\right)^{n-x} \\
&= \frac{(np)^x}{x!} e^{-np}
\end{aligned}
$$

λαμβάνοντας υπόψη ότι

$$\lim_{n\to\infty}\left(1-\frac{np}{n}\right)^{-x}=1 \text{ και } \lim_{n\to\infty}\left(1-\frac{np}{n}\right)^{n}=e^{-np}.$$

Να σημειώσουμε ότι η παραπάνω σύγκλιση είναι ομοιόμορφη ως προς x. Οπότε μπορούμε να θεωρήσουμε ότι η X ακολουθεί κατανομή Poisson με παράμετρο $np=m_x r_x^c$. (Παπαϊωάννου, 2000). Χρησιμοποιώντας τη μέθοδο της μέγιστης πιθανοφάνειας, προκύπτει ότι η εκτίμηση της έντασης θνησιμότητας δίνεται από τη σχέση

$$m_x=\frac{d_x}{r_x^c},$$

η οποία μπορεί να συμβολισθεί και με $\overset{0}{m}_x$ κατ' αντιστοιχία με το συμβολισμό του αδρού δείκτη θνησιμότητας $\overset{0}{q}_x$.

Παρατηρούμε ότι για την εκτίμηση της έντασης θνησιμότητας m_x, πρέπει να γνωρίζουμε το χρόνο έκθεσης στον κίνδυνο του θανάτου. Ο Hatzopoulos (1997) ότι ο χρόνος έκθεσης διακρίνεται σε *κεντρικό* (*central*) και *αρχικό* (*initial*) χρόνο. Επίσης παρατηρεί ότι κατά τη συλλογή των δεδομένων, αυτά θα πρέπει να προέρχονται από *ομοιογενή* πληθυσμό που σημαίνει ότι τα άτομα έχουν παρόμοια ένταση θνησιμότητας. Αυτό απαιτείται για την επίτευξη ακριβών αποτελεσμάτων.

Ένας τρόπος για τον υπολογισμό του κεντρικού χρόνου είναι να προσθέσουμε το χρόνο παραμονής στην έρευνα κάθε ατόμου που συμμετέχει σε αυτή. Για παράδειγμα, αν το i - οστό άτομο αρχίσει να παρατηρείται στην ηλικία $x+a_i$ και φύγει από την έρευνα είτε λόγω θανάτου είτε για οποιονδήποτε άλλο λόγο στην ηλικία $x+b_i$, τότε ο χρόνος παραμονής του στην έρευνα είναι b_i-a_i. Συνεπώς ο κεντρικός χρόνος έκθεσης στον κίνδυνο του θανάτου, ο οποίος πρόκειται για τον ακριβή χρόνο έκθεσης, είναι

$$r_x^c=\sum_i (b_i-a_i).$$

Η μέθοδος αυτή ονομάζεται *ευθεία* (*direct*) και χρησιμοποιείται όταν επιλέγουμε δείγμα από τον πληθυσμό.

Μια εναλλακτική μέθοδος υπολογισμού είναι αυτή της *απογραφής* (*census*). Σύμφωνα με αυτή ισχύει

$$r_x^c = \int_0^T P_x(t)dt\,,$$

όπου T είναι η διάρκεια της έρευνας και $P_x(t)$ είναι ο πληθυσμός ατόμων ηλικίας x που βρίσκονται στη ζωή τη χρονική στιγμή t. Για τον υπολογισμό του παραπάνω ολοκληρώματος μπορεί να χρησιμοποιηθεί για παράδειγμα ο κανόνας του τραπεζίου οπότε έχουμε

$$r_x^c = \frac{T}{2}\big(P_x(0) + P_x(T)\big).$$

Διαιρώντας τον αριθμό των θανάτων d_x με το χρόνο έκθεσης στον κίνδυνο r_x^c, παίρνουμε την εκτίμηση της έντασης θνησιμότητας, όπως είδαμε παραπάνω.

Σε περίπτωση που ο χρόνος έκθεσης στον κίνδυνο συνεχίζει να μετράται μέχρι τη στιγμή που το άτομο θα έφευγε κανονικά από την έρευνα, ακόμη και αν αυτό έχει πεθάνει, και ο χρόνος αυτός προστίθεται στο r_x^c, τότε μιλάμε για τον αρχικό χρόνο έκθεσης. Μια ικανοποιητική προσέγγιση αυτού δίνεται από τη σχέση

$$r_x^i = r_x^c + \frac{d_x}{2}.$$

Διαιρώντας τον αριθμό των θανάτων d_x με τον χρόνο έκθεσης στον κίνδυνο r_x^i, υπολογίζουμε τον *αρχικό αδρό δείκτη θνησιμότητας* $\overset{\circ}{q}_x$.

Ας δούμε τώρα αν η συνάρτηση επιβίωσης είναι άγνωστη, πώς μπορούμε προσεγγιστικά να υπολογίσουμε την ένταση θνησιμότητας m_x από έναν πίνακα θνησιμότητας. Από τη σχέση ${}_nq_x = 1 - \exp\left\{-\int_0^n m_{x+t}dt\right\}$, που είδαμε στην παράγραφο 1.2, και για $n=1$ παίρνουμε $q_x = 1 - \exp\left\{-\int_0^1 m_{x+t}dt\right\}$ από όπου καταλήγουμε στη σχέση $\int_0^1 m_{x+t}dt = -\ln p_x$. Αν η ένταση θνησιμότητας δεν διαφέρει αρκετά από μια γραμμική συνάρτηση στο πεδίο ορισμού της, δηλαδή ισχύει $m_x = a + bx$, τότε το ορισμένο ολοκλήρωμα παριστά τη μέση τιμή της m_x μεταξύ των ηλικιών x και $x+1$. Προσεγγίζοντας τη μέση τιμή με $m_{x+1/2}$, προκύπτει ότι $m_{x+1/2} \approx -\ln p_x$. Ολοκληρώνοντας τη m_{x+t} από $t=-1$ έως $t=1$, παίρνουμε

$$\int_{-1}^1 m_{x+t}dt = \int_{-1}^0 m_{x+t}dt + \int_0^1 m_{x+t}dt = -\ln p_{x-1} - \ln p_x\,.$$

Υπό την υπόθεση της γραμμικότητας, έχουμε

$$\int_{-1}^{1} m_{x+t}\,dt \approx 2m_x .$$

Άρα παίρνουμε

$$2m_x \approx -\ln p_{x-1} - \ln p_x$$

οπότε

$$m_x \approx -\frac{1}{2}\left(\ln p_{x-1} + \ln p_x\right)$$

ή

$$m_x \approx -\frac{1}{2}\left(\ln(1-q_{x-1}) + \ln(1-q_x)\right).$$

Αν τώρα υποθέσουμε ότι η συνάρτηση επιβίωσης S_x είναι πολυώνυμο $2^{ου}$ βαθμού, αναπτύσσοντας τη συνάρτηση σε σειρά Taylor, έχουμε ότι

$$S_{x+h} = S_x + hS'_x + \frac{h^2}{2}S''_x ,$$

όπου ο τόνος συμβολίζει την παραγώγιση ως προς x.

Άρα ισχύουν οι σχέσεις

$$S_{x-1} = S_x - S'_x + \frac{1}{2}S''_x \quad \text{και} \quad S_{x+1} = S_x + S'_x + \frac{1}{2}S''_x,$$

οπότε ισχύει

$$S_{x-1} - S_{x+1} = -2S'_x .$$

Από τη σχέση $m_x = -\frac{\partial}{\partial x}\ln S_x = -\frac{1}{S_x}S'_x$ και λαμβάνοντας υπόψη ότι $S_x = kl_x$, όπου k είναι σταθερά (συνήθως 100.000) και l_x ο αριθμός των ατόμων σε κίνδυνο στην ηλικία x, παίρνουμε προσεγγιστικά ότι

$$m_x \approx \frac{1}{2l_x}\left(l_{x-1} - l_{x+1}\right) = \frac{1}{2l_x}\left(d_{x-1} + d_x\right).$$

Για περισσότερες πληροφορίες, παραπέμπουμε στον Μπλέσιο (1998).

Για κλασματικές ηλικίες, ο London (1985) προτείνει την προσέγγιση

$$m_{x+1/2} = \frac{q_x}{1 - \frac{1}{2}q_x},$$

υπό την προϋπόθεση της ομοιόμορφης κατανομής των θανάτων και $m_{x+1/2}$ είναι η προσεγγιστική μέση τιμή της έντασης θνησιμότητας.

Γνωρίζουμε ότι ένας εκτιμητής, μπορεί να γραφεί ως το άθροισμα της πραγματικής τιμής και ενός θετικού ή αρνητικού σφάλματος. Χρησιμοποιώντας το συμβολισμό του London (1985), μπορούμε να γράψουμε ότι

$$u_x = t_x + e_x,$$

όπου u_x είναι οι αρχικές εκτιμήσεις των ποσοστών θνησιμότητας ή της έντασης θνησιμότητας, t_x είναι οι πραγματικές τιμές των παραπάνω ποσοτήτων, e_x είναι τα σφάλματα και $x = 1, 2, ..., n$ είναι η ηλικία, όπου το n μπορεί να συμπίπτει με την οριακή ηλικία w ή όχι. Δηλαδή μπορεί να εξομαλύνουμε έναν ολόκληρο πίνακα θνησιμότητας ή ένα μόνο κομμάτι του. Θεωρώντας τυχαίες μεταβλητές, έχουμε

$$U_x = t_x + E_x.$$

Σημειώνουμε ότι οι πραγματικές τιμές t_x δεν είναι τυχαίες μεταβλητές.

Στο σημείο αυτό να σημειώσουμε, ότι στη συνέχεια της διπλωματικής εργασίας, θα γίνεται εναλλαγή στον τρόπο συμβολισμού των ποσοστών θνησιμότητας και της έντασης θνησιμότητας. Πιο συγκεκριμένα, όταν μια μέθοδος εξομάλυνσης αναφέρεται αποκλειστικά στα ποσοστά θνησιμότητας, τότε οι πραγματικές τιμές αυτών θα συμβολίζονται με q_x, $x = 1,2,...,n$ και οι αρχικές εκτιμήσεις τους με $\overset{0}{q}_x$. Αντίστοιχα, στις μεθόδους που αναφέρονται αποκλειστικά στην ένταση θνησιμότητας, οι πραγματικές τιμές της θα συμβολίζονται με m_x, ενώ οι αρχικές εκτιμήσεις της με $\overset{0}{m}_x$. Αντίθετα, στις μεθόδους που μπορούν να εφαρμοσθούν τόσο στα ποσοστά όσο και στην ένταση θνησιμότητας, οι πραγματικές τιμές των παραπάνω ποσοτήτων θα συμβολίζονται με t_x, οι αρχικές εκτιμήσεις τους με u_x και οι εξομαλυμένες τιμές τους με v_x.

Σύμφωνα με τον Miller (1949), ο αναλογιστής επιθυμεί τα ποσοστά ή η ένταση θνησιμότητας να μεταβάλλονται ομαλά, γιατί έτσι εξασφαλίζει ότι θα μεταβάλλονται ομαλά και οι διάφορες ασφαλιστικές ποσότητες που υπολογίζονται με βάση τις τιμές αυτές.

1.4 Χαρακτηριστικά της εξομάλυνσης

Δυο είναι τα βασικά χαρακτηριστικά της εξομάλυνσης, η *ομαλότητα* (*smoothness*) και η *καλή προσαρμογή* (*goodness of fit*). Τα δυο αυτά στοιχεία είναι αντικρουόμενα και η επίτευξη του ενός απαιτεί τη θυσία του άλλου (Haberman, 1998 και Greville, 1983). Ας δούμε τώρα τί εννοούμε με τους δυο παραπάνω όρους.

i) Ομαλότητα

Η ομαλότητα είναι μια μαθηματική έννοια, της οποίας όμως ο ορισμός δεν είναι ξεκάθαρος. Ο Weber (1976) ορίζει μια συνάρτηση $f(x)$ ως *ομαλή*, αν τόσο η συνάρτηση όσο και η πρώτη της παράγωγος $f'(x)$ είναι συνεχείς. Ο Mac Lane (1986) αναφέρει ως ομαλή, μια συνάρτηση με όσο το δυνατόν περισσότερες συνεχείς παραγώγους.

Η ομαλότητα μπορεί να ελεγχθεί γραφικά εξετάζοντας αν η γραφική παράσταση των αρχικών δεδομένων δεν έχει σημαντικές διακυμάνσεις. Εναλλακτικά, υπολογίζεται ένα αριθμητικό μέτρο της ομαλότητας το οποίο βασίζεται στις διαφορές τρίτης ή τέταρτης τάξης των εξομαλυμένων τιμών. Ως μέτρο της ομαλότητας μπορεί να χρησιμοποιηθεί το άθροισμα των τετραγώνων ή των απόλυτων τιμών αυτών των διαφορών (London, 1985 και Haberman, 1998). Συνήθως χρησιμοποιείται το

$$S = \sum_x \left(\Delta^z v_x\right)^2 \ , \ x = 1, 2, ..., n - z \ ,$$

όπου v_x είναι οι εξομαλυμένες τιμές, $\Delta^z v_x$ είναι οι διαφορές των εξομαλυμένων τιμών και συνήθως $z =$ 3 ή 4. Σημειώνουμε ότι οι διαφορές $1^{ης}$, $2^{ης}$, $3^{ης}$ και $4^{ης}$ τάξης ορίζονται ως

$$\Delta v_x = v_{x+1} - v_x \ , \ \ \Delta^2 v_x = v_x - 2v_{x+1} + v_{x+2} \ ,$$

$$\Delta^3 v_x = -v_x + 3v_{x+1} - 3v_{x+2} + v_{x+3} \text{ και } \Delta^4 v_x = v_x - 4v_{x+1} + 6v_{x+2} - 4v_{x+3} + v_{x+4}$$

αντίστοιχα.

Εάν η τιμή του μέτρου S είναι σχετικά μικρή (κοντά στο μηδέν), συμπεραίνουμε ότι η σειρά των νέων εκτιμήσεων είναι ομαλή και συνεπώς καλύτερη από τη σειρά των αρχικών εκτιμήσεων. Βέβαια το μέτρο S από μόνο του δεν έχει καμία πρακτική σημασία. Απλά μπορεί να χρησιμοποιηθεί ως μέτρο σύγκρισης της ομαλότητας που προκύπτει από διαφορετικές εξομαλύνσεις των ίδιων δεδομένων. Σημειώνεται ότι το μέτρο S μπορεί να υπολογισθεί και για τις αρχικές τιμές u_x.

ii) Καλή προσαρμογή

Είναι λογικό να θέλουμε οι εξομαλυμένες τιμές να μην απέχουν πολύ από τις αρχικές εκτιμήσεις. Άλλωστε η απόκλιση από τις αρχικές τιμές υπάρχει για την επίτευξη του επιθυμητού βαθμού ομαλότητας. Ο London (1985) προτείνει ως μέτρα προσαρμογής τα παρακάτω:

$$F_1 = \sum_x w_x (u_x - v_x)$$

$$F_2 = \sum_x w_x (u_x - v_x)^2$$

$$F_3 = \sum_x x w_x (u_x - v_x)$$

όπου v_x είναι οι εξομαλυμένες τιμές, u_x είναι οι αρχικές εκτιμήσεις και w_x είναι βάρη για $x = 1, 2, ..., n$.

Όσο μικρότερη είναι η τιμή του μέτρου τόσο καλύτερη προσαρμογή υπάρχει. Όσον αφορά το μέτρο F_1, αυτό είναι ανεπαρκές καθώς μεγάλες θετικές ή αρνητικές διαφορές μπορούν τυχαία να αναιρεθούν και έτσι να έχουμε μικρή τιμή για το F_1 αλλά κακή προσαρμογή. Για την αποφυγή αυτού του προβλήματος, συνήθως χρησιμοποιείται το μέτρο F_2. Αν στο μέτρο F_3 χρησιμοποιήσουμε ως βάρη w_x τον αριθμό των ατόμων σε κίνδυνο l_x, τότε μια μικρή τιμή του μέτρου αυτού, σημαίνει ότι ο αριθμός των θανάτων θα είναι περίπου ίδιος τόσο για τα παρατηρούμενα όσο και για τα εξομαλυμένα δεδομένα.

1.5 Στατιστικά τεστ για την εξομάλυνση

Μετά από κάθε εξομάλυνση ποσοστών θνησιμότητας, θα πρέπει να ελέγχεται κατά πόσο οι εξομαλυμένες τιμές βρίσκονται κοντά στις αρχικές εκτιμήσεις και ικανοποιούν τους επιπλέον περιορισμούς που θέτουμε όπως για παράδειγμα μονοτονία, κυρτότητα κ.α. ανάλογα με τη φύση του προβλήματος. Για παράδειγμα, ο αναλογιστής επιθυμεί τα ποσοστά ή η ένταση θνησιμότητας να αναπαρίστανται από μια μονότονη και πιο συγκεκριμένα αύξουσα συνάρτηση, δηλαδή οι τιμές αυτές να αυξάνονται από ηλικία σε ηλικία. Επίσης προτιμά η συνάρτηση να είναι κυρτή, δηλαδή οι τιμές να αυξάνονται πιο απότομα στις μεγάλες ηλικίες.

Οι Benjamin and Pollard (1980), παρουσιάζουν ορισμένα τεστ για τον έλεγχο της υπόθεσης ότι τα παρατηρούμενα ποσοστά θνησιμότητας προέρχονται από έναν πληθυσμό με δεδομένα ποσοστά q_{x0}, $x = 1, 2, ..., n$, που για παράδειγμα παίρνουμε από κάποιον δημοσιευμένο πίνακα θνησιμότητας. Πέρα όμως από την προσαρμογή στις δεδομένες εκτιμήσεις θα πρέπει να ελέγχονται οι αποκλίσεις για την πιθανή ύπαρξη των παρακάτω προβλημάτων:

α) αριθμός αρκετά μεγάλων αποκλίσεων οι οποίες μπορεί να αντισταθμίζονται από άλλες μικρές αποκλίσεις,

β) μεγάλη αθροιστική απόκλιση σε μέρος ή ολόκληρο το εύρος των ηλικιών,

γ) πολλές θετικές ή αρνητικές αποκλίσεις (δηλαδή συσσώρευση αποκλίσεων του ίδιου προσήμου) σε όλο το εύρος των ηλικιών.

Έτσι παίρνουμε ως μηδενική υπόθεση του ελέγχου, την υπόθεση ότι τα πραγματικά ποσοστά θνησιμότητας δεν διαφέρουν σημαντικά από τα αντίστοιχα ποσοστά κάποιου πίνακα θνησιμότητας, δηλαδή $H_0 : q_x = q_{x0}$ έναντι $H_a : q_x \neq q_{x0}$.

Όλα τα τεστ που χρησιμοποιούνται για τον έλεγχο της παραπάνω υπόθεσης είναι γνωστά από τη θεωρία ελέγχου υποθέσεων.

$\mathbf{X^2}$ τεστ καλής προσαρμογής

Για τον *έλεγχο καλής προσαρμογής* χρησιμοποιείται το γνωστό X^2 *τεστ*. Ορίζουμε ως l_x τον αριθμό των ατόμων σε κίνδυνο ηλικίας x και ως d_x τον αριθμό των θανάτων στην ίδια ηλικία. Υπό τη μηδενική υπόθεση, θεωρούμε ότι το πλήθος των θανάτων d_x ακολουθεί τη διωνυμική κατανομή με παραμέτρους l_x και q_{x0}, δηλαδή

$$d_x \sim Bi(l_x, q_{x0}),\ x = 1, 2, ..., n.$$

Σε περίπτωση που ο αναμενόμενος αριθμός θανάτων $l_x q_{x0}$ είναι μεγαλύτερος του 5, τότε ασυμπτωτικά το d_x θα ακολουθεί τη κανονική κατανομή με μέσο $l_x q_{x0}$ και διακύμανση $l_x q_{x0}(1 - q_{x0})$. Τότε υπό τη μηδενική υπόθεση, η ποσότητα

$$X^2 = \sum_{x=1}^{n} \frac{(d_x - l_x q_{x0})^2}{l_x q_{x0}(1 - q_{x0})}$$

ακολουθεί τη c^2 κατανομή με n βαθμούς ελευθερίας – υπό την προϋπόθεση ότι τα d_x είναι ανεξάρτητα μεταξύ τους – όπου n είναι ο αριθμός των ηλικιών ή των διαστημάτων ηλικιών. Η μηδενική υπόθεση απορρίπτεται σε επίπεδο σημαντικότητας a αν $X^2 > c^2_{n,1-a}$ όπου το $c^2_{n,1-a}$ βρίσκεται από τους πίνακες της c^2 κατανομής και ορίζεται από τη σχέση $P(X^2 \geq c^2_{n,1-a}) = 1-a$.

Σημειώνουμε επίσης ότι το X^2 τεστ αδυνατεί να ελέγξει τα προβλήματα (α) – (γ), για τον έλεγχο των οποίων χρησιμοποιούνται τα παρακάτω τεστ.

Έλεγχος μεμονωμένων τυποποιημένων αποκλίσεων

Για τον *έλεγχο του προβλήματος (α)* εφαρμόζεται το τεστ των *μεμονωμένων τυποποιημένων αποκλίσεων*. Υπό τη μηδενική υπόθεση, ισχύει ότι ο αριθμός των θανάτων στην ηλικία x, $x = 1, 2, ..., n$, d_x, ακολουθεί τη διωνυμική κατανομή η οποία μπορεί να προσεγγιστεί από τη $N\left(l_x q_{x0}, l_x q_{x0}(1-q_{x0})\right)$. Θεωρώντας ότι οι θάνατοι στις διάφορες ηλικίες είναι ανεξάρτητοι μεταξύ τους και ότι δεν υπάρχει ετερογένεια σε κάθε ηλικία, οι τυποποιημένες αποκλίσεις

$$\frac{d_x - l_x q_{x0}}{\sqrt{l_x q_{x0}(1-q_{x0})}}$$

θα πρέπει να συμπεριφέρονται ως ανεξάρτητες παρατηρήσεις που προέρχονται από την τυπική κανονική κατανομή. Σε επίπεδο σημαντικότητας, έστω $a = 0.05$, απορρίπτεται η μηδενική υπόθεση όταν περισσότερο από το 5% των τυποποιημένων αποκλίσεων υπερβαίνει κατ' απόλυτο τιμή την τιμή 1.96.

Έλεγχος απόλυτων τυποποιημένων αποκλίσεων

Για το *ίδιο πρόβλημα* μπορούν επίσης να χρησιμοποιηθούν οι *απόλυτες τυποποιημένες αποκλίσεις*

$$\left| \frac{d_x - l_x q_{x0}}{\sqrt{l_x q_{x0}(1-q_{x0})}} \right|, \; x = 1, 2, ..., n .$$

Επειδή $P\left(|Z| \le \frac{2}{3}\right) \approx 0.5$, $Z \sim N(0,1)$, όπου Z είναι οι παραπάνω τυποποιημένες αποκλίσεις, τότε υπό τη μηδενική υπόθεση, ο αριθμός, έστω R, των απόλυτων τυποποιημένων αποκλίσεων που υπερβαίνουν την τιμή $\frac{2}{3}$ θα κατανέμεται ως διωνυμική κατανομή με παραμέτρους n και $\frac{1}{2}$. Η μηδενική υπόθεση θα απορρίπτεται αν ο αριθμός R ανήκει στην άνω a περιοχή της διωνυμικής κατανομής. Αν ο αριθμός των ηλικιών n είναι μεγαλύτερος του 20, τότε προσεγγιστικά το R θα ακολουθεί τη κανονική κατανομή με μέσο $\frac{n}{2}$ και διακύμανση $\frac{n}{4}$ και η μηδενική υπόθεση θα απορρίπτεται αν η ποσότητα

$$T = \frac{2R - n}{\sqrt{n}}$$

ανήκει στην άνω a περιοχή της κατανομής $N(0,1)$.

Έλεγχος αθροιστικών (συσσωρευμένων) αποκλίσεων

Για τον *έλεγχο του προβλήματος (β)* μπορούν να χρησιμοποιηθούν οι *αθροιστικές αποκλίσεις*. Υποθέτοντας ότι οι θάνατοι στις γειτονικές ηλικίες είναι ανεξάρτητοι μεταξύ τους, τότε υπό τη μηδενική υπόθεση, η αθροιστική απόκλιση ενός τυχαία επιλεγμένου διαστήματος ηλικιών, έστω $[x_1, x_2]$, η οποία ισούται με

$$\sum_{x=x_1}^{x_2} (d_x - l_x q_{x0}),$$

θα είναι μια κανονική τυχαία μεταβλητή με μέσο 0 και διακύμανση $\sum_{x=x_1}^{x_2} l_x q_{x0}(1 - q_{x0})$. Η μηδενική υπόθεση απορρίπτεται αν η απόλυτη τιμή της αθροιστικής απόκλισης σε όλο το εύρος των ηλικιών είναι μεγαλύτερη από το διπλάσιο της τετραγωνικής ρίζας της αντίστοιχης διακύμανσης ή αν οι αθροιστικές αποκλίσεις σε τμήματα του εύρους των ηλικιών είναι πολύ μεγάλες κατ' απόλυτη τιμή. Σημειώνεται, ότι τα διαστήματα ηλικιών δεν πρέπει να αλληλοκαλύπτονται, διαφορετικά οι αποκλίσεις θα είναι θετικά συσχετισμένες και το αναμενόμενο ποσοστό των τεστ που δίνουν σημαντικό αποτέλεσμα δεν θα ξεπερνά το 5%.

Προσημικό τεστ

Το *πρόβλημα (γ)* μπορεί να ελεγχθεί μέσω του *προσημικού τεστ*. Υπό τη μηδενική υπόθεση, τα πρόσημα των μεμονωμένων αποκλίσεων είναι ανεξάρτητα και περίπου ισοκατανεμημένα σε θετικά και αρνητικά. Συνεπώς ο αριθμός K των θετικών (ή αρνητικών) προσήμων σε μια σειρά $n+1$ αποκλίσεων θα ακολουθεί τη $Bi\left(n,\frac{1}{2}\right)$ και η μηδενική υπόθεση απορρίπτεται αν ο αριθμός K βρίσκεται στην κάτω ή άνω a περιοχή της $Bi\left(n,\frac{1}{2}\right)$. Αν $n \geq 20$, το K προσεγγίζεται από τη $N\left(\frac{n}{2},\frac{n}{4}\right)$ και η μηδενική υπόθεση απορρίπτεται αν η ποσότητα

$$L = \frac{2K - n}{\sqrt{n}},$$

ανήκει στην κάτω ή άνω a περιοχή της τυπικής κανονικής κατανομής.

Εναλλακτικό τεστ για το ίδιο πρόβλημα είναι το *τεστ του Steven* για τις ομαδοποιήσεις των προσήμων, που βασίζεται στην υπεργεωμετρική κατανομή και το *διωνυμικό τεστ* για την αλλαγή των προσήμων. Για περισσότερες πληροφορίες για τα παραπάνω τεστ παραπέμπουμε στους Benjamin and Pollard (1980). Οι ίδιοι συγγραφείς αναφέρουν, ότι καθώς όλα τα παραπάνω τεστ είναι ευκολότερα στην εφαρμογή τους από το X^2 τεστ, είναι προτιμότερο να γίνονται πριν από το X^2, ο υπολογισμός του οποίου θα πρέπει να αποφεύγεται αν τα προηγούμενα τεστ δείχνουν ανεπαρκή ή ισχυρή προσαρμογή.

Τα παραπάνω τεστ μπορούν να χρησιμοποιηθούν και για τον έλεγχο της εξομάλυνσης. Στην περίπτωση αυτή, ως μηδενική υπόθεση θεωρούμε την υπόθεση ότι η μέθοδος εξομάλυνσης που χρησιμοποιείται δίνει ικανοποιητικά αποτελέσματα. Δηλαδή ότι τα αρχικά δεδομένα δεν διαφέρουν από τα εξομαλυμένα. Έτσι στους τύπους των τεστ αντικαθιστούμε το q_{x0} με $\hat{q}_x$, όπου $\hat{q}_x$ είναι τα εξομαλυμένα ποσοστά. Όμως επειδή στις μεθόδους αυτές εκτιμούμε κάποιες παραμέτρους, οι βαθμοί ελευθερίας του X^2 τεστ δεν είναι πλέον n αλλά αφαιρούμε από αυτούς τόσους όσες είναι οι παράμετροι που εκτιμούμε. Θα δούμε στη συνέχεια, κατά την παρουσίαση κάθε μεθόδου εξομάλυνσης, πόσοι είναι οι τελικοί βαθμοί

ελευθερίας. Επίσης να σημειώσουμε ότι το στατιστικό X^2 είναι ίδιο με το μέτρο F_2 της ενότητας 1.4 με βάρη $w_x = \frac{l_x}{v_x(1-v_x)}$, όπου v_x είναι οι εξομαλυμένες τιμές.

Τα τεστ που παρουσιάσαμε, μπορούν να χρησιμοποιηθούν και για την εξομάλυνση που αναφέρεται στην ένταση θνησιμότητας m_x. Υποθέτοντας ότι $d_x \sim P(r_x^c m_x)$, μπορούμε να θεωρήσουμε ότι ασυμπτωτικά το d_x θα ακολουθεί την κανονική κατανομή με μέσο και διακύμανση ίσα με $r_x^c m_x$, όπου r_x^c είναι ο κεντρικός χρόνος έκθεσης στον κίνδυνο.

Αν το X^2 τεστ υποδηλώνει ισχυρή προσαρμογή, πρέπει να απορρίψουμε την εξομάλυνση καθώς τότε η ομαλότητα θα είναι πάρα πολύ μικρή, αν φυσικά η εξομάλυνση έχει γίνει με μια από τις μη παραμετρικές μεθόδους που θα δούμε παρακάτω. Η εξομάλυνση επίσης θα πρέπει να απορρίπτεται αν το παραπάνω τεστ δείχνει κακή προσαρμογή γιατί στην περίπτωση αυτή η ομαλότητα θα είναι πάρα πολύ μεγάλη. Όπως έχουμε ήδη αναφέρει, η ομαλότητα και η καλή προσαρμογή είναι δυο αντικρουόμενα χαρακτηριστικά της εξομάλυνσης και για να είναι αυτή ικανοποιητική, θα πρέπει και τα δυο αυτά χαρακτηριστικά να υπάρχουν σε κάποιο συγκεκριμένο βαθμό.

Τα υπόλοιπα τεστ χρησιμοποιούνται ως έχουν. Βέβαια οι Benjamin and Pollard (1980) αναφέρουν ότι ο Beard το 1951 πρότεινε την τροποποίηση αυτών των ελέγχων, αντικαθιστώντας τη διακύμανση $l_x\hat{q}_x(1-\hat{q}_x)$ με τη $\frac{n-k}{n} l_x\hat{q}_x(1-\hat{q}_x)$, όπου $\hat{q}_x$ είναι τα εξομαλυμένα ποσοστά θνησιμότητας και k ο αριθμός των παραμέτρων που εκτιμώνται, χωρίς η τροποποίηση αυτή να επιφέρει σημαντικές αλλαγές στο αποτέλεσμα των ελέγχων.

1.6 Κατηγορίες μεθόδων εξομάλυνσης

Για την κατασκευή ενός μοντέλου θνησιμότητας το οποίο θα αναπαριστά, δηλαδή θα εκτιμά, το πραγματικό πρότυπο θνησιμότητας, υπάρχουν δυο τρόποι. Ο πρώτος τρόπος είναι να κατασκευάσουμε έναν πίνακα θνησιμότητας και ο δεύτερος είναι να υιοθετήσουμε την άποψη ότι το πρότυπο θνησιμότητας περιγράφεται από μια συνάρτηση. Οι δυο αυτοί τρόποι μας καθοδηγούν στο να θεωρήσουμε δυο κατηγορίες για τις μεθόδους εξομάλυνσης: τις *μη παραμετρικές* και τις *παραμετρικές* (London, 1985 και Haberman, 1998).

Στις παραμετρικές μεθόδους εξομάλυνσης, σκοπός μας είναι να προσαρμόσουμε μια μαθηματική συνάρτηση στις αρχικές εκτιμήσεις των ποσοστών θνησιμότητας $\overset{0}{q}_x$ ή της έντασης θνησιμότητας $\overset{0}{m}_x$, έτσι ώστε να εκφράσουμε τη σχέση που υπάρχει μεταξύ των γειτονικών τιμών. Εξομάλυνση μπορεί να γίνει είτε προσαρμόζοντας μια συνάρτηση σε όλο το εύρος των ηλικιών είτε περισσότερες συναρτήσεις σε υποδιαστήματα των δεδομένων. Οι παράμετροι των συναρτήσεων εκτιμώνται μέσω γνωστών μεθόδων όπως της μέγιστης πιθανοφάνειας, των ελαχίστων τετραγώνων κ.α. Παρόλο που οι παραμετρικές μέθοδοι δίνουν επαρκείς εκτιμητές, περιέχουν κάποια μεροληψία, καθώς δεν υπάρχει καμία συνάρτηση που να αναπαριστά επακριβώς τις πραγματικές τιμές των πιθανοτήτων θνησιμότητας q_x ή της έντασης θνησιμότητας m_x.

Οι μη παραμετρικές μέθοδοι εξομάλυνσης, εφαρμόζονται κυρίως σε πινακοποιημένα δεδομένα συνδυάζοντας δεδομένα σε διάφορες τιμές της ηλικίας x, $x = 1, 2, ..., n$ χωρίς να προϋποθέτουν καμία συναρτησιακή μορφή για τα q_x ή m_x. Και αυτές οι μέθοδοι περιέχουν μεροληψία αλλά αυτή μπορεί να ελεγχθεί μειώνοντας τη δειγματική διακύμανση. Όσον αφορά το βαθμό ομαλότητας, αυτός μπορεί να ελεγχθεί με την επιλογή της κατάλληλης τιμής κάποιων παραμέτρων, όπως θα δούμε στη συνέχεια, ενώ στην παραμετρική εξομάλυνση, η ομαλότητα θεωρείται δεδομένη και δεν μπορεί να αλλάξει εξαιτίας των συναρτησιακών μορφών που χρησιμοποιούνται.

Σύμφωνα με τον Greville (1983), η κατηγοριοποίηση των μεθόδων εξομάλυνσης μπορεί να γίνει και με άλλους τρόπους. Έτσι υπάρχουν οι *διακριτές* μέθοδοι, στις οποίες κάθε παρατηρούμενη τιμή ή αρχική εκτίμηση αντικαθίσταται με μια εξομαλυμένη τιμή και οι *συνεχείς* μέθοδοι, όπου στα παρατηρούμενα δεδομένα προσαρμόζεται μια ομαλή καμπύλη. Ένας άλλος διαχωρισμός είναι σε *τοπικές* (*local*) και *ολικές* (*global*) μεθόδους. Στις τοπικές μεθόδους, η εξομαλυμένη τιμή εξαρτάται από ορισμένες αρχικές εκτιμήσεις ενώ στις ολικές για κάθε εξομαλυμένη τιμή χρησιμοποιούνται όλες οι παρατηρούμενες. Επιπλέον υπάρχει και ο συνδυασμός των τεσσάρων παραπάνω κατηγοριών.

Σε αυτή τη διπλωματική εργασία θα χρησιμοποιήσουμε την κατηγοριοποίηση των London (1985) και Haberman (1998).

Το 2^{ο} κεφάλαιο της εργασίας καλύπτει τις παραμετρικές μεθόδους εξομάλυνσης. Πιο συγκεκριμένα παρουσιάζεται, για την εξομάλυνση των πινάκων θνησιμότητας, η χρήση των

μοντέλων θνησιμότητας, των γενικευμένων γραμμικών μοντέλων και των συναρτήσεων splines. Επίσης παρουσιάζεται η μέθοδος της παρεμβολής ομαλής σύνδεσης (*smooth – junction interpolation*). Το κεφάλαιο 3 έχει αφιερωθεί στις μη παραμετρικές μεθόδους οι οποίες είναι η γραφική μέθοδος, η εξομάλυνση με αναφορά σε τυπικό πίνακα θνησιμότητας, η εξομάλυνση μέσω κινητών σταθμισμένων μέσων, η εξομάλυνση των Whittaker και Henderson, η μπεϋζιανή εξομάλυνση και η εξομάλυνση μέσω εκτιμητών πυρήνα. Στο 4^{ο} κεφάλαιο παρουσιάζεται η εξομάλυνση με χρήση των εννοιών της θεωρίας πληροφοριών, η οποία είναι επίσης μη παραμετρική μέθοδος. Στο κεφάλαιο 5, παρουσιάζεται η εφαρμογή σε δεδομένα ορισμένων από τις παραπάνω μεθόδους ενώ στο 6^{ο} και τελευταίο κεφάλαιο γίνεται μια κριτική των μεθόδων και δίνονται κάποια ανοικτά ερωτήματα που αφορούν την διαδικασία της εξομάλυνσης.

Κεφάλαιο 2

ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΕΞΟΜΑΛΥΝΣΗΣ

Οι παραμετρικές μέθοδοι εξομάλυνσης, όπως υπαινίσσεται και το όνομά τους, προσπαθούν να προσαρμόσουν μια ή περισσότερες μαθηματικές συναρτήσεις στις αρχικές εκτιμήσεις των ποσοστών θνησιμότητας ή της έντασης θνησιμότητας. Στην ουσία δηλαδή, προσαρμόζουμε μια συνάρτηση στα αρχικά δεδομένα, οπότε παίρνουμε τις εξομαλυμένες τιμές. Χρησιμοποιούνται ομαλές συναρτήσεις έτσι ώστε μετά την εξομάλυνση δεν απαιτείται ο υπολογισμός του μέτρου της ομαλότητας, η οποία θεωρείται πλέον δεδομένη. Επειδή καμία συνάρτηση δεν μπορεί να εκτιμήσει με ακρίβεια τις πραγματικές τιμές, οι παραμετρικές μέθοδοι εξομάλυνσης έχουν το μειονέκτημα της μεροληψίας.

Οι μέθοδοι εξομάλυνσης που εντάσσονται στην κατηγορία των παραμετρικών μεθόδων, είναι η εφαρμογή κάποιου από τα μοντέλα θνησιμότητας, η εξομάλυνση με χρήση των γενικευμένων γραμμικών μοντέλων, η εξομάλυνση μέσω των συναρτήσεων splines και η μέθοδος της παρεμβολής ομαλής σύνδεσης. Να σημειώσουμε ότι οι δυο τελευταίες μέθοδοι, σε αντίθεση με τις δυο πρώτες, προσαρμόζουν περισσότερες από μια συναρτήσεις στις αρχικές εκτιμήσεις.

Επίσης να τονίσουμε ότι κατά την παρουσίαση των παραπάνω μεθόδων, όταν αυτές θα αναφέρονται αποκλειστικά στα ποσοστά θνησιμότητας, οι πραγματικές τιμές αυτών θα συμβολίζονται με q_x, $x = 1,2,...,n$ ενώ οι αρχικές εκτιμήσεις τους θα συμβολίζονται με $\overset{0}{q}_x$. Αν αναφέρονται μόνο στην εξομάλυνση της έντασης θνησιμότητας, οι πραγματικές και οι αρχικές τιμές της, θα συμβολίζονται με m_x και $\overset{0}{m}_x$ αντίστοιχα. Αν όμως η μέθοδος μπορεί να χρησιμοποιηθεί τόσο για την εξομάλυνση των ποσοστών όσο και της έντασης θνησιμότητας, τότε οι πραγματικές τιμές θα συμβολίζονται με t_x και οι αρχικές εκτιμήσεις με u_x.

2.1 Εξομάλυνση μέσω μοντέλων θνησιμότητας

Ένας από τους στόχους της αναλογιστικής επιστήμης είναι η περιγραφή του μηχανισμού που περιγράφει τη θνησιμότητα. Για το λόγο αυτό από πολύ παλιά, προτείνονται μαθηματικά μοντέλα. Τα μοντέλα θνησιμότητας μπορεί να αναφέρονται σε διάφορα αναλογιστικά μέτρα με βασικότερες επιλογές την ένταση θνησιμότητας m_x και την πιθανότητα θανάτου q_x. Αν και έχουν προταθεί πάρα πολλά μοντέλα, θεωρείται ότι δύσκολα κάποιο από αυτά μπορεί να περιγράψει τη θνησιμότητα σε όλο ή μεγάλο εύρος ηλικιών. Τα μοντέλα αυτά πρέπει να είναι απλά και να επιτρέπουν τροποποιήσεις, αφού ο χρόνος παρεμβαίνει καταλυτικά στην ανάλυση της θνησιμότητας (Μπλέσιος, 1998).

Σύμφωνα με τον Hatzopoulos (1997) ένα καλό μοντέλο θα πρέπει να έχει σωστό θεωρητικό υπόβαθρο έτσι ώστε να δίνει καλύτερες ερμηνείες. Επίσης θα πρέπει να ισχύει η αρχή της γενικότητας, δηλαδή το μοντέλο να μην περιορίζεται στην περιγραφή μόνο της ανθρώπινης θνησιμότητας, αλλά και άλλων οργανισμών. Επιπλέον, είναι φανερό ότι ένα μοντέλο θα πρέπει να δίνει τα καλύτερα αποτελέσματα με τις ελάχιστες παραμέτρους. Δηλαδή οι παράμετροι θα πρέπει να περιγράφουν τα σημαντικότερα μόνο χαρακτηριστικά που επηρεάζουν τη θνησιμότητα. Τέλος, ένα μοντέλο μπορεί να χρησιμοποιείται ικανοποιητικά μόνο για κάποιο περιορισμένο διάστημα ηλικιών, χωρίς αυτό να σημαίνει ότι δεν είναι καλό. Αυτό ίσως να σημαίνει ότι αποτελεί κομμάτι ενός άλλου γενικότερου αλλά άγνωστου μοντέλου.

Τα *μαθηματικά μοντέλα* ή *νόμοι θνησιμότητας* όπως αλλιώς λέγονται, μπορούν να χρησιμοποιηθούν ως μια εναλλακτική μέθοδος εξομάλυνσης των πινάκων θνησιμότητας.

Βέβαια σε περίπτωση που το μοντέλο περιγράφει την ένταση θνησιμότητας ενώ τα δεδομένα προς εξομάλυνση είναι πιθανότητες θανάτου, μπορούμε να χρησιμοποιήσουμε τον τύπο μετατροπής για κλασματικές ηλικίες

$$m_{x+t} = \frac{q_x}{1 - tq_x},\ x = 1, 2, ..., n,\ 0 < t < 1$$

υπό την υπόθεση της ομοιόμορφης κατανομής των θανάτων (London, 1985 και Μπλέσιος, 1998)

2.1.1 Βασικότερα μοντέλα θνησιμότητας

Ο πρώτος που επιχείρησε να προτείνει ένα μοντέλο θνησιμότητας ήταν ο *DeMoivre* το 1725. Συγκεκριμένα πρότεινε τη μοντελοποίηση της έντασης θνησιμότητας ως συνάρτησης της ηλικίας,

$$m_x = \frac{1}{w-x},\ x = 1, 2, ..., n,$$

όπου w είναι η οριακή ηλικία. Ο ίδιος ο DeMoivre παρατήρησε ότι το μοντέλο αυτό περιγράφει καλύτερα τη θνησιμότητα στις μεγάλες ηλικίες.

Ο πιο γνωστός νόμος κατασκευάστηκε εκατό χρόνια αργότερα από τον *Gompertz*. Η συνάρτηση πυκνότητας πιθανότητας της κατανομής Gompertz με παραμέτρους l και a, όπου $l, a > 0$, είναι η

$$f(x) = l e^{ax} \exp\left\{\frac{l}{a}(1 - e^{ax})\right\},\ x \geq 0.$$

Σύμφωνα με το νόμο αυτό, η ένταση θνησιμότητας m_x αυξάνεται με γεωμετρική πρόοδο, ισχύει δηλαδή

$$m_x = BC^x.$$

Τα B και C, όπου $B = l$ και $C = e^a$ είναι θετικές παράμετροι ενώ η Lytrokapi (1998) αναφέρει ότι η παράμετρος C παίρνει τιμές γύρω στο 1.09. Ο Hatzopoulos (1997) αναφέρει ότι ο λογάριθμος της έντασης θνησιμότητας είναι ίδιος με το λογαριθμικό σύνδεσμο (*log link*) των Γενικευμένων Γραμμικών Μοντέλων, που θα δούμε στη συνέχεια.

Το 1860 ο *Makeham* επέκτεινε το νόμο του Gompertz προσθέτοντας μια σταθερά παράμετρο, έτσι ώστε να ισχύει

$$m_x = A + BC^x,$$

θεωρώντας ότι ο θάνατος μπορεί να είναι τυχαίος ή να οφείλεται στην επιδείνωση της υγείας. Ο νόμος αυτός δεν καταφέρνει να περιγράψει ικανοποιητικά τη θνησιμότητα καθ' όλη τη διάρκεια της ζωής, παρά μόνο για τη μέση ηλικία. Δίνει απογοητευτικά αποτελέσματα για τη νεανική θνησιμότητα ενώ την υπερεκτιμά στις μεγάλες ηλικίες.

Ο *Oppermann* το 1870 πρότεινε το μαθηματικό μοντέλο

$$m_x = \frac{a}{\sqrt{x}} + b + c\sqrt{x},$$

όπου τα a, b και c είναι παράμετροι. Έχει αποδειχθεί ότι ο νόμος αυτός είναι αρκετά ικανοποιητικός για την εξομάλυνση της παιδικής και νεανικής θνησιμότητας. Αντίθετα δεν δίνει καλά αποτελέσματα για τις μεγαλύτερες ηλικίες.

Ένα πρωτοποριακό μοντέλο για τη θνησιμότητα σε όλες τις ηλικίες έδωσαν οι *Thiele and Steffenson* το 1872. Σύμφωνα με αυτούς ισχύει

$$m(x) = m_1(x) + m_2(x) + m_3(x),$$

όπου

$$m_1(x) = a_1 \exp(-b_1 x),$$

$$m_2(x) = a_2 \exp\left(-\frac{1}{2} b_2^2 (x - C)^2\right),$$

$$m_3(x) = a_3 \exp(b_3 x)$$

είναι οι εντάσεις θνησιμότητας στην παιδική ηλικία, στη μέση ηλικία και στα γηρατειά αντίστοιχα. Τα a_1, b_1, a_2, b_2, C, a_3 και b_3 είναι θετικές παράμετροι. Παρατηρούμε ότι τα $m_1(x)$ και $m_3(x)$ μοντελοποιούνται σύμφωνα με το νόμο του Gompertz ενώ το $m_2(x)$ μοιάζει με τη συνάρτηση πυκνότητας πιθανότητας μιας κανονικής κατανομής. Πρόκειται για πολύπλοκο για ευρεία χρήση νόμο και έτσι σήμερα είναι απλά ιστορικού ενδιαφέροντος.

Το 1883 ο γερμανός *Wittstein*, μοντελοποίησε την πιθανότητα θανάτου ως

$$q_x = \frac{1}{m} a^{-(m-x)^n} + a^{-(w-x)^n},$$

όπου w είναι η οριακή ηλικία και τα a, m, n είναι άγνωστες παράμετροι. Ο νόμος αυτός κάνει τη διάκριση της θνησιμότητας σε παιδική και ενήλικη.

Ο *Makeham* το 1889 παρουσίασε ένα νέο μοντέλο, γνωστό ως «*2ος νόμος του Makeham*», που δεν είναι τίποτα παραπάνω από το προηγούμενό του με έναν επιπλέον όρο, δηλαδή

$$m_x = A + Hx + BC^x,$$

όπου το H είναι άγνωστη παράμετρος.

Ο *Perk* το 1932 πρότεινε τους τύπους

$$m_x = \frac{A + BC^x}{1 + DC^x}$$

και

$$m_x = \frac{A + BC^x}{KC^{-x} + 1 + DC^x},$$

οι οποίοι καλύπτουν τη θνησιμότητα όλων των ηλικιών και τα A, B, C, K, D είναι παράμετροι.

Ο *Beard* το 1951 πρότεινε το μοντέλο για την πιθανότητα θανάτου

$$q_x = \frac{A + BC^x}{EC^{-2x} + 1 + DC^x},$$

σύμφωνα με τη Lytrokapi (1998), και το μοντέλο για την ένταση θνησιμότητας

$$m_x = \frac{BC^x}{1 + DC^x},$$

σύμφωνα με τον Hatzopoulos (1997), όπου τα A, B, C, E, D είναι παράμετροι. Ο νόμος αυτός δεν αναπαράγει τη μείωση των ποσοστών θνησιμότητας που παρατηρείται στις μικρές ηλικίες.

Το 1951, επίσης, προτάθηκε η χρήση, για το χρόνο ζωής T, της κατανομής *Weibull* με παραμέτρους l και a, όπου $l, a > 0$, η συνάρτηση πυκνότητας πιθανότητας της οποίας είναι

$$f(x) = lax^{a-1} \exp\{-lx^a\},\ x \geq 0.$$

Η ένταση θνησιμότητας της κατανομής αυτής δίνεται από τη σχέση

$$m_x = Bx^C,$$

όπου $B = la$ και $C = a - 1$.

Οι *Gavrilov and Gavrilova* (1991) πρότειναν το *γενικευμένο νόμο του Weibull*, προσθέτοντας μια επιπλέον παράμετρο, την A, με βάση τον οποίο έχουμε

$$m_x = A + Bx^C.$$

Το 1974 ο *Barnett*, για την εξομάλυνση της θνησιμότητας των ασφαλισμένων στη Βρετανία για τα έτη 1967 – 1970, χρησιμοποίησε τον τύπο

$$\frac{q_x}{1 - q_x} = A - Hx + BC^x$$

όπου τα A, H, B και C είναι άγνωστες παράμετροι.

Η πιο γνωστή ίσως σχέση για εξομάλυνση των ποσοστών θνησιμότητας, είναι αυτή των *Heligman and Pollard* (1980). Σύμφωνα με αυτή ισχύει

$$\frac{q_x}{p_x} = A^{(x+B)^C} + D\exp\{-E(\ln x - \ln F)^2\} + GH^x.$$

Σύμφωνα με τον Hatzopoulos (1997) η καμπύλη αναπαράγει τρία διακριτά χαρακτηριστικά: τη θνησιμότητα ενός παιδιού που προσαρμόζεται στο φυσικό περιβάλλον, τη θνησιμότητα

που σχετίζεται με τη γήρανση του σώματος και τη θνησιμότητα που οφείλεται σε βίαιους λόγους.

Όσον αφορά τις παραμέτρους του μοντέλου έχουμε τις παρακάτω ερμηνείες:

Το A είναι περίπου ίδιο με την πιθανότητα θανάτου πριν την ηλικία του ενός έτους, δηλαδή την q_1. Το B αναπαριστά τη διαφορά μεταξύ των ποσοστών q_0 και q_1 ενώ το C μετρά το ρυθμό μείωσης της βρεφικής ή παιδικής θνησιμότητας. Η παράμετρος D παριστάνει την ένταση του όγκου ή καμπούρας (*hump*) ατυχημάτων, η E την εξάπλωση (*spread*) ενώ η F τη θέση (*location*) του. Τέλος το G αναφέρεται στο επίπεδο της γηράσκουσας θνησιμότητας ενώ το H μετρά το ρυθμό αύξησής της.

Ο τύπος αυτός περιγράφει πολύ ικανοποιητικά τη θνησιμότητα σε ολόκληρο το εύρος των ηλικιών, δημιουργεί όμως κάποιες συστηματικές αποκλίσεις από τα εμπειρικά δεδομένα καθώς ο όγκος των ατυχημάτων τοποθετείται σε μεγαλύτερη ηλικία.

Ορισμένες παραλλαγές της παραπάνω έκφρασης είναι οι εξής:

$$q_x = A^{(x+B)^C} + D\exp\left\{-E(\ln x - \ln F)^2\right\} + \frac{GH^x}{1+GH^x},$$

$$q_x = A^{(x+B)^C} + D\exp\left\{-E(\ln x - \ln F)^2\right\} + \frac{GH^x}{1+KGH^x},$$

$$q_x = A^{(x+B)^C} + D\exp\left\{-E(\ln x - \ln F)^2\right\} + \frac{GH^{x^K}}{1+GH^{x^K}},$$

όπου η παράμετρος K εξασφαλίζει κυρτότητα στις μεγάλες ηλικίες.

Οι *Mode and Bushy* το 1982 μοντελοποίησαν τη συνάρτηση επιβίωσης $S(x)$ ως εξής:

$$S(x) = \begin{cases} S_0(x), & 0 \le x \le d_0 \\ S_0(d_0)S_1(x-d_0), & d_0 \le x \le d_1 \\ S_0(d_0)S_1(d_1-d_2)S_2(x-d_1), & x \ge d_1 \end{cases}$$

όπου

$$S_0(x) = \exp\{-a_0(\exp\{-b_0 x\}-1)\},$$

$$S_1(x) = \exp\left\{\frac{b_1 g_1^3}{3} - a_1 x + \frac{b_1}{3}(x-g_1)^3\right\},$$

$$S_2(x) = \exp\{-a_2 x - b_2(\exp\{g_2 x\}-1)\}.$$

Το $S_0(x)$ είναι η πιθανότητα το άτομο να είναι εν ζωή στην ηλικία $x \le d_0$, το $S_1(x-d_0)$ είναι η πιθανότητα το άτομο να είναι εν ζωή στην ηλικία x $(d_0 \le x \le d_1)$ δοθέντος ότι έχει επιζήσει μέχρι την ηλικία d_0 και το $S_2(x-d_1)$ είναι η πιθανότητα το άτομο να βρίσκεται στη ζωή στην ηλικία $x \ge d_1$ δοθέντος ότι έχει επιζήσει μέχρι την ηλικία d_1. Οι τιμές d_0 και d_1 επιλέγονται αυθαίρετα, πράγμα που αποτελεί μειονέκτημα για το μοντέλο. Τα a_0, a_1, a_2, b_0, b_1, b_2, g_1 και g_2 είναι άγνωστες παράμετροι.

Η *Kostaki* (1992), στην προσπάθειά της να μειώσει τα συστηματικά σφάλματα που παράγει το μοντέλο των Heligman – Pollard, πρότεινε μια παραλλαγή του μοντέλου προσθέτοντας έναν επιπλέον όρο. Συγκεκριμένα έχουμε

$$\frac{q_x}{p_x} = \begin{cases} A^{(x+B)^C} + D\exp\left\{-E_1(\ln x - \ln F)^2\right\} + GH^x, & x \le F \\ A^{(x+B)^C} + D\exp\left\{-E_2(\ln x - \ln F)^2\right\} + GH^x, & x > F, \end{cases}$$

όπου οι παράμετροι E_1 και E_2 αναπαριστούν την αριστερή και δεξιά αντίστοιχα περιοχή της κορυφής του όγκου των ατυχημάτων.

Παρατηρούμε ότι υπάρχει μια πληθώρα μοντέλων, κάποια από τα οποία αναφέρονται στην ένταση θνησιμότητας m_x και κάποια άλλα στα ποσοστά θνησιμότητας q_x. Επίσης ορισμένα μοντέλα είναι πλέον ιστορικού ενδιαφέροντος ενώ άλλα χρησιμοποιούνται ευρύτατα στις μέρες μας. Ελάχιστα μοντέλα έχουν αποδειχθεί να περιγράφουν τη θνησιμότητα σε όλο το εύρος των ηλικιών. Συνεπώς ο ερευνητής, ανάλογα με τις ανάγκες του, θα πρέπει να επιλέγει και το ανάλογο μοντέλο. Πάντως, ένα ικανοποιητικό μοντέλο είναι αυτό των Heligman and Pollard καθώς και η παραλλαγή αυτού που πρότεινε η Kostaki (1992).

2.1.2 Προσαρμογή των μοντέλων

Ας δούμε τώρα πώς μπορούμε να εξομαλύνουμε τα αρχικά δεδομένα με βάση τα μοντέλα θνησιμότητας. Έστω ότι τα μοντέλα αναφέρονται σε πιθανότητες θανάτου q_x. Υποθέτουμε ή έχουμε κάποια ένδειξη ότι τα αδρά ποσοστά θνησιμότητας που έχουμε στη διάθεσή μας, περιγράφονται ικανοποιητικά από κάποιο από τα μοντέλα. Στη συνέχεια, προσαρμόζουμε το

μοντέλο στα αρχικά δεδομένα και εκτιμούμε τις άγνωστες παραμέτρους με κάποια από τις γνωστές μεθόδους, όπως είναι η μέθοδος μέγιστης πιθανοφάνειας, τα ελάχιστα τετράγωνα και το ελάχιστο X^2. Οι προσαρμοσμένες τιμές που προκύπτουν, αποτελούν τις εξομαλυμένες τιμές.

i) Μέγιστη πιθανοφάνεια

Υποθέτουμε ότι ο αριθμός των θανάτων στην ηλικία x, d_x, ακολουθεί τη διωνυμική κατανομή με παραμέτρους l_x και q_x, όπου l_x είναι ο αριθμός των ατόμων σε κίνδυνο στην ηλικία x και q_x είναι το αντίστοιχο πραγματικό ποσοστό θνησιμότητας. Επίσης υποθέτουμε ότι οι θάνατοι στις διάφορες ηλικίες είναι ανεξάρτητοι μεταξύ τους. Η πιθανοφάνεια είναι

$$L = \prod_{x=1}^{n} \binom{l_x}{d_x} q_x^{d_x} (1-q_x)^{l_x - d_x} .$$

Σημειώνουμε ότι τα μόνα άγνωστα στοιχεία της παραπάνω πιθανοφάνειας είναι τα q_x, τα οποία αντικαθιστούμε με κάποιο από τα μοντέλα.

Οι εκτιμητές μέγιστης πιθανοφάνειας για τις άγνωστες παραμέτρους του μοντέλου είναι οι τιμές που μεγιστοποιούν την ποσότητα L ή αντίστοιχα την

$$\log L = \sum_{x=1}^{n} \left[\log \binom{l_x}{d_x} + d_x \log q_x + (l_x - d_x) \log(1-q_x) \right].$$

Στη συνέχεια, εξισώνοντας με μηδέν τις πρώτες παραγώγους, ως προς τις άγνωστες παραμέτρους που εμφανίζονται στους τύπους των q_x, του

$$\Lambda = \sum_{x=1}^{n} \left[d_x \log q_x + (l_x - d_x) \log(1-q_x) \right]$$

παίρνουμε τις ταυτόχρονες εξισώσεις για τον υπολογισμό των εκτιμητών μέγιστης πιθανοφάνειας για τις παραμέτρους.

ii) Ελάχιστα τετράγωνα

Υποθέτουμε ότι $\overset{0}{q}_x$ είναι η αρχική εκτίμηση του ποσοστού θνησιμότητας στην ηλικία x και $q_x = m(x)$ είναι η τιμή που θέλουμε να προσαρμόσουμε σε αυτή την ηλικία. Η επιλογή των παραμέτρων του μοντέλου $m(x)$ πρέπει να γίνει έτσι ώστε οι προσαρμοσμένες τιμές να είναι όσο το δυνατόν πιο κοντά στις παρατηρούμενες τιμές $\left\{ \overset{0}{q}_x \right\}$. Ένας τρόπος για να το

πετύχουμε αυτό είναι η ελαχιστοποίηση του αθροίσματος των τετραγώνων των διαφορών μεταξύ των προσαρμοσμένων και των παρατηρούμενων τιμών, δηλαδή της ποσότητας

$$\sum_{x=1}^{n}\left[\overset{\circ}{q}_x - m(x)\right]^2 .$$

Η ποσότητα αυτή αξίζει να χρησιμοποιείται αν όλα τα ποσοστά θνησιμότητας έχουν την ίδια διακύμανση, πράγμα που δεν ισχύει. Τότε, εφαρμόζεται η μέθοδος των σταθμισμένων ελαχίστων τετραγώνων, οπότε ελαχιστοποιούμε τη σχέση

$$\sum_{x=1}^{n} w_x\left[\overset{\circ}{q}_x - m(x)\right]^2 . \qquad (2.1)$$

Σημειώνουμε ότι τα βάρη είναι ανάλογα της διακύμανσης των ποσοστών θνησιμότητας και όσο πιο μικρά είναι αυτά τόσο πιο καλή είναι η προσαρμογή.

Στην περίπτωσή μας η διακύμανση των πραγματικών ποσοστών θνησιμότητας q_x είναι ίση με

$$s_x^2 = \frac{q_x(1-q_x)}{l_x}$$

ή προσεγγιστικά

$$s_x^2 = \frac{q_x}{l_x},$$

αν $(1-q_x)\approx 1$. Επειδή τα πραγματικά ποσοστά q_x είναι άγνωστα τα προσεγγίζουμε είτε με τα αντίστοιχα ποσοστά q_x^s ενός τυπικού πίνακα θνησιμότητας είτε με τα αδρά ποσοστά $\overset{\circ}{q}_x$.

Για την προσαρμογή ενός μη γραμμικού μοντέλου $q_x = m(x)$ χρησιμοποιούνται είτε τα βάρη $\left\{\frac{l_x}{q_x^s}\right\}$ είτε τα $\left\{\frac{l_x}{\overset{\circ}{q}_x}\right\}$ για τον υπολογισμό των αρχικών ποσοστών $\left\{q_x^{(1)}\right\}$. Στη συνέχεια οι εξομαλυμένες τιμές υπολογίζονται χρησιμοποιώντας ως βάρη τις τιμές $\left\{\frac{l_x}{q_x^{(1)}}\right\}$.

Για τον υπολογισμό των άγνωστων παραμέτρων εξισώνουμε με μηδέν τις πρώτες μερικές παραγώγους της σχέσης (2.1).

iii) Ελάχιστο $\mathbf{X^2}$

Θεωρώντας ότι $d_x \sim Bi(l_x, q_x)$ και ότι ο αναμενόμενος αριθμός θανάτων $l_x q_x$ δεν είναι μικρότερος του πέντε, τότε η ασυμπτωτική κατανομή του d_x είναι η κανονική με μέσο $l_x q_x$ και διακύμανση $l_x q_x (1 - q_x)$. Τότε η ποσότητα

$$X^2 = \sum_{x=1}^{n} \frac{(d_x - l_x q_x)^2}{l_x q_x (1 - q_x)}$$

ακολουθεί τη c^2 κατανομή με $n - k$ βαθμούς ελευθερίας, όπου n είναι ο αριθμός των ηλικιών και k ο αριθμός των παραμέτρων που εκτιμώνται. Οι παράμετροι του μοντέλου $q_x = m(x)$ επιλέγονται έτσι ώστε να ελαχιστοποιείται το X^2.

Για εναλλακτικές μεθόδους εκτίμησης των παραμέτρων των μοντέλων θνησιμότητας ο αναγνώστης παραπέμπεται στην Lytrokapi (1998).

Στην περίπτωση που θέλουμε να προσαρμόσουμε ένα μοντέλο θνησιμότητας που αναφέρεται στην ένταση θνησιμότητας m_x, έχουμε τα παρακάτω:

Στη μέθοδο της μέγιστης πιθανοφάνειας, η πιθανοφάνεια είναι

$$L = \prod_{x=1}^{n} \frac{\left(r_x^c m_x\right)^{d_x} \exp\left\{- r_x^c m_x\right\}}{d_x!}.$$

Στη μέθοδο ελαχίστων τετραγώνων η μόνη διαφορά βρίσκεται στη διακύμανση που εμπλέκεται στα βάρη ή οποία δίνεται ως $s_x^2 = \frac{m_x}{r_x^c}$, ενώ στη μέθοδο του ελάχιστου X^2, βασιζόμαστε στο ότι $d_x \sim P(r_x^c m_x)$, οπότε μπορούμε να θεωρήσουμε ότι ασυμπτωτικά το d_x θα ακολουθεί την κανονική κατανομή με μέσο και διακύμανση ίσα με $r_x^c m_x$, όπου r_x^c είναι ο κεντρικός χρόνος έκθεσης στον κίνδυνο.

Σημειώνουμε ότι με όποιο μοντέλο και αν κάνουμε την εξομάλυνση των ασφαλιστικών ποσοτήτων, θα πρέπει στο X^2 τεστ να αφαιρέσουμε από τους n βαθμούς ελευθερίας και k επιπλέον βαθμούς όπου k ο αριθμός των παραμέτρων του μοντέλου που εκτιμώνται.

2.2 Εξομάλυνση μέσω Γενικευμένων Γραμμικών Μοντέλων

Μια δεύτερη παραμετρική μέθοδος εξομάλυνσης πινάκων θνησιμότητας είναι η χρησιμοποίηση των *Γενικευμένων Γραμμικών Μοντέλων* (*Generalized Linear Models*, *GLM*) (Hatzopoulos, 1997).

Πριν προχωρήσουμε στην περιγραφή της μεθόδου αυτής, ας δούμε πρώτα συνοπτικά το θεωρητικό υπόβαθρο των GLM.

2.2.1 Γενικευμένα Γραμμικά Μοντέλα

Γνωρίζουμε ότι τα κλασσικά γραμμικά μοντέλα είναι της μορφής

$$\mathbf{Y} = \mathbf{Xb} + \boldsymbol{\varepsilon} \tag{2.2}$$

όπου $\mathbf{Y}$ είναι το $n \times 1$ διάνυσμα των εξαρτημένων μεταβλητών οι τιμές των οποίων ονομάζονται αποκρίσεις, $\mathbf{b}$ είναι το $k \times 1$ διάνυσμα των παραμέτρων, $\mathbf{X}$ είναι ο $n \times k$ πίνακας των ανεξάρτητων ή ερμηνευτικών μεταβλητών και $\boldsymbol{\varepsilon}$ είναι το $n \times 1$ διάνυσμα των σφαλμάτων τα οποία θεωρούνται ότι είναι ανεξάρτητα μεταξύ τους και κατανέμονται ως $N(0, s^2)$ (Agresti, 2002).

Η γενίκευση των μοντέλων (2.2) οδηγεί στα γενικευμένα γραμμικά μοντέλα. Η γενίκευση αυτή μπορεί να γίνει με δυο τρόπους (Haberman and Pitacco, 1999). Ο πρώτος είναι μέσω της εισαγωγής μιας μεγαλύτερης οικογένειας κατανομών, της *εκθετικής διασποράς (exponential dispersion family)*. Αυτό γίνεται επειδή οι εξαρτημένες μεταβλητές είναι κυρίως κατηγορικές ή περιορισμένες σε ένα τμήμα των τιμών, οπότε δεν μπορούν να προέρχονται από την κανονική κατανομή.

Στην εκθετικής διασποράς οικογένεια κατανομών ανήκουν οι κατανομές που μπορούν να γραφούν στη μορφή

$$f_{Y_i}(y_i, q_i, f) = \exp\left\{\frac{y_i q_i - b(q_i)}{a(f)} + c(y_i, f)\right\},$$

για κάποιες συναρτήσεις $a(\cdot)$, $b(\cdot)$ και $c(\cdot)$, η οποία είναι μια ειδική περίπτωση της εκθετικής οικογένειας κατανομών (Agresti, 2002). Το q_i ονομάζεται *φυσική ή κανονική* παράμετρος ενώ το $f > 0$ καλείται *παράμετρος κλίμακας ή διασποράς* (*scale or dispersion parameter*). Στην εκθετικής διασποράς οικογένεια, ανήκουν η κανονική κατανομή, η Poisson,

η Διωνυμική, η Gamma κ.α. κατανομές. Εάν χρησιμοποιήσουμε το λογάριθμο της $f_{Y_i}(y_i, q_i, f)$ παίρνουμε ότι

$$E(Y_i) = m_i = \frac{\partial b(q_i)}{\partial q_i} \text{ και } \text{var}(Y_i) = \frac{\partial^2 b(q_i)}{\partial q_i^2} a(f).$$

Ο δεύτερος τρόπος της γενίκευσης είναι ενώνοντας τη *γραμμική πρόβλεψη*

$$h_i = \mathbf{x}_i' \mathbf{b}$$

με το μέσο m_i, μέσω μιας μονότονης διαφορίσιμης συνάρτησης g, η οποία λέγεται *συνάρτηση σύνδεσης* (*link function*). Ισχύει δηλαδή

$$h_i = g(m_i)$$

έτσι ώστε

$$m_i = g^{-1}(h_i) = g^{-1}(\mathbf{x}_i' \mathbf{b}).$$

Να σημειώσουμε ότι τα $\mathbf{x}_i'$ είναι οι γραμμές του $n \times k$ πίνακα των ανεξάρτητων μεταβλητών $\mathbf{X}$.

Συνεπώς παρατηρούμε ότι ένα GLM αποτελείται από τρία μέρη:

1. Τη τυχαία μεταβλητή απόκρισης Y_i (*τυχαία συνιστώσα*), η κατανομή της οποίας ανήκει στην οικογένεια εκθετικής διασποράς,
2. Τη *συστηματική συνιστώσα* για την οποία οι συμμεταβλητές X_i παράγουν μια γραμμική πρόβλεψη για κάθε παρατήρηση, δηλαδή $h_i = \mathbf{x}_i' \mathbf{b}$ και
3. Τη συνάρτηση σύνδεσης που συνδέει τη τυχαία με τη συστηματική συνιστώσα, δηλαδή την $h_i = g(m_i)$.

Άρα η αναμενόμενη τιμή του Y_i, m_i, συνδέεται με το h_i μέσω της συνάρτησης σύνδεσης και η διακύμανσή του είναι $\text{var}(Y_i) = \frac{f V(m_i)}{w_i}$, όπου τα w_i είναι *βάρη* (*prior weights*).

Στα κλασσικά γραμμικά μοντέλα ισχύει $E(Y_i) = m_i = h_i$, οπότε έχουμε *ταυτοτική* (*identity*) σύνδεση, ενώ αν ισχύει $q_i = h_i$ έχουμε τον *φυσιολογικό* (*natural*) σύνδεσμο.

Για τις βασικές κατανομές της οικογένειας εκθετικής διασποράς, ο Gerber (1997) προτείνει τις παρακάτω συναρτήσεις σύνδεσης:

Κατανομή	**Κανονική Συνάρτηση Σύνδεσης**	**Παράμετροι**
Κανονική	Identity	$q = m$
Poisson	Log	$q = \log m$
Διωνυμική	Logit	$q = \log \frac{1}{1-m}$
Gamma	Reciprocal	$q = -\frac{1}{m}$

Πίνακας 2.1: Συναρτήσεις σύνδεσης για τα μέλη της εκθετικής διασποράς οικογένειας κατανομών

Για την εκτίμηση των αγνώστων παραμέτρων του μοντέλου, b_j, μεγιστοποιούμε τη λογαριθμοπιθανοφάνεια, η οποία δίνεται από τη σχέση

$$l(\mathbf{y}, \mathbf{m}) = \sum_{i=1}^{n} \left[\frac{y_i m_i - b(m_i)}{a(f)} + c(y_i, f) \right]. \tag{2.3}$$

Οι εξισώσεις που οδηγούν στις εκτιμήσεις των b_j είναι οι

$$\frac{\partial l(\mathbf{y}, \mathbf{m})}{\partial b_j} = 0 \Rightarrow \sum_{i=1}^{n} \frac{(y_i - m_i) x_{ij}}{\text{var}(Y_i)} \frac{\partial m_i}{\partial h_i} = 0 ,$$

για κάθε j, $j = 1, 2, ..., k$, οι οποίες ονομάζονται εξισώσεις εκτίμησης πιθανοφάνειας και λύνονται με επαναληπτικές μεθόδους όπως η μέθοδος των Newton – Raphson. Παρατηρούμε ότι τα b_j εμπλέκονται στη σχέση (2.3) μέσω του m_i και της γραμμικής πρόβλεψης.

Αν γνωρίζουμε εκ των προτέρων προσθετικούς όρους της γραμμικής πρόβλεψης, τότε η συμμετοχή τους σε αυτή καλείται *αντισταθμιστικός παράγοντας* (*offset*) έτσι ώστε

$$h_i = \sum_{j=1}^{k} X_{ij} b_j .$$

Σκοπός μας είναι να προσαρμόσουμε ένα μοντέλο που θα δίνει τιμές $\hat{\mathbf{m}}$ όσο το δυνατόν πιο κοντά στις πραγματικές τιμές $\mathbf{m}$, χρησιμοποιώντας τις λιγότερες παραμέτρους. Ένα μέτρο για τον έλεγχο της καλής προσαρμογής του μοντέλου είναι η *κλιμακωτή ψευδοαπόκλιση* (*scaled quasi deviance*)

$$S_{c,f}(\mathbf{y}, \hat{\mathbf{m}}) = \frac{D_{c,f}(\mathbf{y}, \hat{\mathbf{m}})}{f} ,$$

όπου

$$D_{c,f}(\mathbf{y},\hat{\mathbf{m}}) = -2fQ(\mathbf{y},\hat{\mathbf{m}}) = \sum_{i=1}^{n} d_i = 2\sum_{i=1}^{n} w_i\left[\, y_i(\tilde{q}_i - \hat{q}_i) - b(\tilde{q}_i) + b(\hat{q}_i)\right] \qquad (2.4)$$

η *μη κλιμακωτή ψευδοαπόκλιση* (*unscaled quasi deviance*), f το πλήρες μοντέλο (με τις αρχικές εκτιμήσεις των δεδομένων), c το μοντέλο που ελέγχουμε – το οποίο έχει λιγότερες παραμέτρους από το f – και $\hat{\mathbf{m}}$ οι προσαρμοσμένες τιμές του μοντέλου c. Επίσης έχουμε $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}(\hat{\mathbf{m}})$ και $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}(\mathbf{y})$.

Για τις κατανομές της οικογένειας εκθετικής διαποράς, η κλιμακωτή ψευδοαπόκλιση ισούται με το

$$-2\log\frac{l_c}{l_f},$$

όπου l_c και l_f είναι οι τιμές της πιθανοφάνειας κάτω από τα μοντέλα c και f αντίστοιχα. Η ποσότητα

$$\hat{f} = \frac{D_{c,f}(\mathbf{y},\hat{\mathbf{m}})}{n-k}$$

αποτελεί ασυμπτωτική και αμερόληπτη εκτιμήτρια για την παράμετρο κλίμακας f και ονομάζεται *εκτιμημένη διασπορά* (*dispersion*).

Αποδεικνύεται ότι $S_{c,f}(\mathbf{y},\hat{\mathbf{m}}) \sim c^2_{n-k}$ και αν $S_{c,f}(\mathbf{y},\hat{\mathbf{m}}) > c^2_{n-k,1-a}$ απορρίπτεται σε επίπεδο σημαντικότητας a, η υπόθεση ότι το μοντέλο έχει k παραμέτρους.

Σύμφωνα με τον Hatzopoulos (1997), δυο είναι τα βασικά είδη σφαλμάτων στα GLM: τα *σφάλματα του Pearson* και τα *σφάλματα απόκλισης*.

Τα σφάλματα του Pearson ορίζονται ως

$$r_i^P = \frac{y_i - \hat{m}_i}{\sqrt{\frac{V(\hat{m}_i)}{w_i}}}$$

ενώ τα σφάλματα απόκλισης υπολογίζονται ως

$$r_i^D = sign(y_i - \hat{m}_i)\sqrt{d_i}\,,$$

όπου το d_i έχει οριστεί στη σχέση (2.4). Και οι δυο τύποι τυποποιούνται, οπότε παίρνουμε τα αντίστοιχα τυποποιημένα σφάλματα, αν διαιρεθούν με την ποσότητα $\sqrt{\hat{f}(1-h_i)}$, όπου h_i είναι το (i,i) στοιχείο του πίνακα $\mathbf{H} = \mathbf{W}^{-\frac{1}{2}}\mathbf{X}(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}^{-\frac{1}{2}}$, όπου το T δηλώνει την

αντιστροφή του πίνακα, και ονομάζεται *μόχλευση* (*leverage*). Ο πίνακας $\mathbf{W}$ είναι διαγώνιος με στοιχεία τα βάρη.

Τα σφάλματα και κυρίως αυτά της απόκλισης, χρησιμοποιούνται για τον έλεγχο της εξομάλυνσης όπως θα δούμε παρακάτω.

2.2.2 Εξομάλυνση πινάκων θνησιμότητας και GLM

Εδώ πρέπει να τονίσουμε ότι τα δεδομένα που έχουμε στη διάθεση μας, είναι συνήθως της μορφής (d_x, r_x^j), όπου d_x είναι ο αριθμός των θανάτων στην ηλικία x, $x = 1, 2, ..., n$, r_x^j με $j = c, i$ είναι οι κεντρικοί και αρχικοί χρόνοι έκθεσης στον κίνδυνο αντίστοιχα, που είδαμε στην ενότητα 1.3 και στην περίπτωση αυτή ο αδρός δείκτης θνησιμότητας υπολογίζεται ως $\overset{\circ}{q}_x = \frac{d_x}{r_x^i}$ και η ένταση θνησιμότητας ως $\overset{\circ}{m}_x = \frac{d_x}{r_x^c}$.

Για την εξομάλυνση της έντασης θνησιμότητας m_x, υποθέτουμε ότι οι θάνατοι σε κάθε ηλικία είναι ανεξάρτητοι μεταξύ τους και ακολουθούν την κατανομή Poisson. Ο Gerber (1997) παρατηρεί ότι η ίδια έκφραση για την πιθανοφάνεια, με την προηγούμενη περίπτωση, προκύπτει αν υποτεθεί ότι οι χρόνοι έκθεσης στον κίνδυνο του θανάτου μοντελοποιούνται σαν Gamma τυχαίες μεταβλητές δοθέντος βέβαια του αριθμού των θανάτων σε κάθε ηλικία x. Στην περίπτωση όμως που μοντελοποιούμε το χρόνο έκθεσης στον κίνδυνο ως Gamma τυχαία μεταβλητή, υποκείμενο εξομάλυνσης είναι η *ένταση ζωτικότητας* (*vitality*) $\frac{1}{m_x}$. Επίσης υποθέτουμε ότι η ένταση θνησιμότητας είναι σταθερή στο διάστημα ηλικιών $(x, x+1)$, οπότε θα τη συμβολίζουμε με $m_{x+1/2}$ (Haberman and Pittacco, 1999 και Renshaw et al., 1996α).

Η εξομάλυνση των ποσοστών θνησιμότητας, μέσω των GLM, βασίζεται στην υπόθεση ότι η τυχαία μεταβλητή απόκρισης D_x ακολουθεί τη διωνυμική κατανομή (Haberman, 1998).

Στη συνέχεια θα δούμε πώς εφαρμόζονται τα GLM στις τρεις αυτές περιπτώσεις (Hatzopoulos, 1997).

2.2.3 GLM και Poisson κατανομή

Έστω ότι έχουμε μια ομάδα ατόμων ηλικίας x, $x = 1, 2, ..., n$. Συμβολίζουμε με D_x την τυχαία μεταβλητή που παριστάνει τον αριθμό των θανάτων και με r_x^c τον κεντρικό χρόνο

έκθεσης στον κίνδυνο. Υποθέτουμε ότι η D_x ακολουθεί την κατανομή Poisson με μέσο και διακύμανση ίσα με $r_x^c m_{x+1/2}$, δηλαδή ισχύει

$$D_x \sim P(r_x^c m_{x+1/2}).$$

Σε περίπτωση που χρησιμοποιούμε ως ανεξάρτητες τυχαίες μεταβλητές απόκρισης τη σειρά $\{D_x\}$, για να ορίσουμε το κατάλληλο GLM παίρνουμε

$$E(D_x) = m_x = r_x^c m_{x+1/2} \text{ και } \text{var}(D_x) = \frac{fV(m_x)}{w_x} = m_x,$$

όπου $V(m_x) = m_x$, $f = 1$ και $w_x = 1$.

Η κλιμακωτή απόκλιση υπολογίζεται ως

$$S(c, f) = 2\sum_{x=1}^{n}\left\{y_x \log\frac{y_x}{\hat{m}_x} - (y_x - \hat{m}_x)\right\},$$

όπου τα y_x ταυτίζονται με τα παρατηρούμενα d_x και $\hat{m}_x$ είναι οι εξομαλυμένες τιμές $r_x^c \hat{m}_{x+1/2}$. Έτσι η παραπάνω έκφραση μπορεί να γραφεί ως

$$S(c, f) = 2\sum_{x=1}^{n}\left\{y_x \log\frac{y_x}{r_x^c \hat{m}_{x+1/2}} - (y_x - r_x^c \hat{m}_{x+1/2})\right\}. \quad (2.5)$$

Για την εφαρμογή αυτού του GLM χρησιμοποιούμε τη λογαριθμική γραμμική πρόβλεψη,

$$h_x = \log m_x = \log(r_x^c m_{x+1/2}) = \log r_x^c + \log m_{x+1/2} = \log r_x^c + \sum_{j=1}^{k} b_j x^j,$$

όπου το $\log r_x^c$ συμπεριφέρεται ως αντισταθμιστικός παράγοντας. Ο Hatzopoulos (1997) παρατηρεί ότι βάσει της σχέσης $\log m_{x+1/2} = \sum_{j=1}^{k} b_j x^j$, η ένταση θνησιμότητας μοντελοποιείται σαν ένα εκθετικό πολυώνυμο στην ηλικία x.

Μια εναλλακτική περίπτωση είναι να χρησιμοποιήσουμε ως τυχαίες μεταβλητές απόκρισης, τη σειρά $\left\{\frac{D_x}{r_x^c}\right\}$. Στην περίπτωση αυτή για τον ορισμό του κατάλληλου GLM παίρνουμε

$$E\left(\frac{D_x}{r_x^c}\right) = m_x = \frac{1}{r_x^c}E(D_x) = m_{x+1/2} \text{ και } \text{var}\left(\frac{D_x}{r_x^c}\right) = \frac{fV(m_x)}{w_x} = \frac{1}{(r_x^c)^2}\text{var}(D_x) = \frac{m_x}{r_x^c},$$

όπου $f = 1$ και $w_x = r_x^c$.

Η σχέση (2.5) ισχύει και σε αυτή την περίπτωση (Renshaw et al., 1996α). Ο μετασχηματισμός αυτός γίνεται όταν δεν θέλουμε να χρησιμοποιήσουμε αντισταθμιστικό παράγοντα ή/και όταν θέλουμε να χρησιμοποιήσουμε άλλο σύνδεσμο όπως για παράδειγμα είναι ο σύνδεσμος δύναμης (*power link*).

2.2.4 GLM και Gamma κατανομή

Υποθέτοντας ότι $D_x \sim P(r_x^c m_{x+1/2})$, $x = 1, 2, ..., n$ κάνουμε την επιπλέον υπόθεση ότι οι χρόνοι έκθεσης στον κίνδυνο είναι ανεξάρτητες τυχαίες μεταβλητές οι οποίες ακολουθούν την κατανομή Gamma με μέσο d_x και διακύμανση $m_{x+1/2}$.

Παρατηρούμε ότι στην περίπτωση αυτή άγνωστη είναι μόνο η ένταση θνησιμότητας.

Για τον ορισμό του κατάλληλου GLM χρησιμοποιώντας ως μεταβλητές απόκρισης τη σειρά $\{R_x^c\}$ παίρνουμε

$$E(R_x^c) = m_x = \frac{d_x}{m_{x+1/2}} \quad \text{και} \quad \text{var}(R_x^c) = \frac{fV(m_x)}{w_x} = \frac{m_x^2}{d_x},$$

όπου $f = 1$, $w_x = d_x$ και $V(m_x) = m_x^2$.

Στην περίπτωση αυτή η κλιμακωτή απόκλιση δίνεται από τη σχέση

$$S(c, f) = -2\sum_{x=1}^{n} d_x \left\{ \log \frac{R_x^c}{\hat{m}_x} - \frac{R_x^c - \hat{m}_x}{\hat{m}_x} \right\}$$

όπου $\hat{m}_x = \dfrac{d_x}{\hat{m}_{x+1/2}}$ είναι οι εξομαλυμένες τιμές. Είναι εύκολο να δειχθεί ότι η παραπάνω σχέση είναι ίδια με τη σχέση (2.5).

Για την εφαρμογή του παραπάνω GLM, στοχεύοντας στην εξομάλυνση της έντασης ζωτικότητας $\dfrac{1}{m_{x+1/2}}$, χρησιμοποιούμε πάλι το λογαριθμικό σύνδεσμο

$$h_x = \log m_x = \log d_x + \log \frac{1}{m_{x+1/2}} = \log d_x + \sum_{j=1}^{k} b_j x^j,$$

όπου το $\log d_x$ είναι ο αντισταθμιστικός παράγοντας και πάλι η ένταση θνησιμότητας μοντελοποιείται σαν εκθετικό πολυώνυμο στην ηλικία x.

Οι Renshaw et al. (1996α) αποδεικνύουν ότι, δοθέντος ότι τα βάρη $w_x = d_x$ είναι μη μηδενικά, η μέθοδος αυτή δίνει ακριβώς τα ίδια εξομαλυμένα αποτελέσματα όπως η μέθοδος με την Poisson κατανομή.

Και εδώ υπάρχει η εναλλακτική λύση της χρησιμοποίησης ως τυχαίων μεταβλητών απόκρισης της σειράς $\left\{\frac{R_x^c}{d_x}\right\}$ οπότε έχουμε

$$E\left(\frac{R_x^c}{d_x}\right)=m_x=\frac{1}{m_{x+1/2}} \quad \text{και} \quad \text{var}(R_x^c)=\frac{fV(m_x)}{w_x}=\frac{m_x^2}{d_x},$$

όπου $f=1$, $w_x=d_x$, $V(m_x)=m_x^2$ και επιπλέον ισχύει η σχέση (2.5) για την κλιμακωτή απόκλιση.

Οι Renshaw et al. (1996α) σημειώνουν ότι οι δυο μέθοδοι δίνουν τις ίδιες εκτιμήσεις για τις άγνωστες παραμέτρους b_j, $j=1,2,...,k$, αλλά με αντίθετο πρόσημο, όταν ως μεταβλητές απόκρισης χρησιμοποιούνται οι σειρές $\{D_x\}$ ή $\{R_x^c\}$. Ταύτιση τόσο στις τιμές όσο και στο πρόσημο έχουμε στην περίπτωση όπου ως αποκρίσεις χρησιμοποιούνται οι σειρές $\left\{\frac{D_x}{r_x^c}\right\}$ ή $\left\{\frac{R_x^c}{d_x}\right\}$.

Ως διαγνωστικά μέτρα για την εξομάλυνση, όταν χρησιμοποιείται η κατανομή Poisson, οι Renshaw et al. (1996α) προτείνουν τα

$$dev_x=d_x-e_x,\ v_x=\sqrt{e_x},\ z_x=\frac{dev_x}{\sqrt{v_x}} \quad \text{και} \quad 100\frac{d_x}{e_x},$$

όπου $e_x=r_x^c\hat{m}_{x+1/2}$ είναι ο αναμενόμενος αριθμός θανάτων στην ηλικία x, $x=1,2,...,n$. Παρατηρούμε ότι z_x είναι τα σφάλματα του Pearson που όπως έχουμε δει, υπολογίζονται μέσω της σχέσης $z_x=\frac{y_x-\hat{m}_x}{\sqrt{\frac{V(\hat{m}_x)}{w_x}}}$. Έτσι το $\sum_{x=1}^{n}z_x^2$ μπορεί να χρησιμοποιηθεί για τον έλεγχο του κατά πόσο είναι καλή η εξομάλυνση.

Υπό τη δυϊκή μεθοδολογία, όταν δηλαδή χρησιμοποιείται η κατανομή Gamma, προτείνουν τα μέτρα

$$d\bar{e}v_x=r_x^c-\bar{e}_x,\ \bar{v}_x=\sqrt{\frac{\bar{e}_x}{d_x}},\ \bar{z}_x=\frac{d\bar{e}v_x}{\sqrt{\bar{v}_x}} \quad \text{και} \quad 100\frac{r_x^c}{\bar{e}_x},$$

όπου $\bar{e}_x = \frac{d_x}{\hat{m}_{x+1/2}}$ είναι ο αναμενόμενος χρόνος έκθεσης στον κίνδυνο στην ηλικία x και z_x είναι πάλι τα σφάλματα του Pearson.

Οι δυο μέθοδοι δίνουν σφάλματα με αντίθετα πρόσημα και συγκεκριμένα ισχύει $d\bar{e}v_x = -\frac{dev_x}{\hat{m}_{x+1/2}}$. Τα σφάλματα z_x και $\bar{z}_x$ διαφέρουν στις τιμές και έχουν αντίθετα πρόσημα. Αντίθετα τα σφάλματα απόκλισης, $sign(dev_x)\sqrt{d_x}$ και $sign(d\bar{e}v_x)\sqrt{d_x}$, εξαιτίας της ισότητας των στοιχείων της απόκλισης, έχουν το ίδιο μέγεθος αλλά και αυτά έχουν αντίθετο πρόσημο.

Ο Hatzopoulos (1997), για τον έλεγχο της μηδενικής υπόθεσης ότι οι εξομαλυμένες τιμές δεν διαφέρουν σημαντικά από τις αρχικές εκτιμήσεις, προτείνει να υπολογίζονται τα τυποποιημένα σφάλματα απόκλισης

$$z_x^D = \frac{sign(dev_x)\sqrt{d_x}}{\sqrt{\hat{f}(1-h_x)}}.$$

Τότε

$$X^2 = \sum_{x=1}^{n} (z_x^D)^2 \sim c_{n-k}^2,$$

όπου k είναι ο αριθμός των παραμέτρων που εκτιμώνται και απορρίπτουμε τη μηδενική υπόθεση αν το

$$p - value = 1 - F_{n-k}(X^2) < a.$$

Επίσης ο Hatzopoulos (1997) προτείνει τον έλεγχο της εξομάλυνσης μέσω των *διαγραμμάτων των σφαλμάτων* (*residual plots*). Τα διαγράμματα αυτά δημιουργούνται σχεδιάζοντας τα τυποποιημένα σφάλματα απόκλισης έναντι της γραμμικής πρόβλεψης ή των εξομαλυμένων τιμών μετασχηματισμένων στη *σταθερή κλίμακα πληροφορίας* (*constant information scale, CIS*), η οποία ορίζεται ως

$$\int \frac{d\hat{m}}{V^{1/2}(\hat{m})}.$$

Τέτοια διαγράμματα μπορούν να αποκαλύψουν απομακρυσμένα σημεία ή γενική καμπυλότητα (*general curvature*) που δείχνει μη ικανοποιητική κλίμακα συμμεταβλητών ή συνάρτηση σύνδεσης. Επίσης μπορεί να αποκαλύψει μια τάση στη διασπορά που σημαίνει μη ικανοποιητική συνάρτηση διακύμανσης.

Πέραν από αυτές τις περιπτώσεις ο Hatzopoulos (1997) αναφέρει τη μοντελοποίηση των ασφαλιστηρίων συμβολαίων ως σύνθετη κατανομή Poisson, χρησιμοποιώντας τον κεντρικό χρόνο έκθεσης στον κίνδυνο. Επίσης παρουσιάζει τη μοντελοποίηση της ανθεκτικότητας (*resistivity*) στον θάνατο $\frac{1}{\hat{m}_{x+1/2}}$, βασισμένος στα ασφαλιστήρια συμβόλαια, ως Gamma κατανομή και την υπόθεση ότι ο λογάριθμος του $\frac{1}{\hat{m}_{x+1/2}}$ ακολουθεί την κανονική κατανομή.

2.2.5 GLM και διωνυμική κατανομή

Υποθέτουμε ότι έχουμε έναν κλειστό πληθυσμό ατόμων ηλικίας x, $x = 1, 2, ..., n$, καθένα από τα οποία συνεισφέρει στο χρόνο έκθεσης στον κίνδυνο με έναν ολόκληρο χρόνο κατά την αρχή της έρευνας. Με άλλα λόγια χρησιμοποιούμε τον αρχικό χρόνο έκθεσης στον κίνδυνο r_x^i. Υπό αυτές τις υποθέσεις μπορούμε να πούμε ότι κάθε άτομο αποτελεί μια δοκιμή Bernoulli με επιτυχία το θάνατό του. Η πιθανότητα επιτυχίας είναι ένα μέτρο της θνησιμότητας αφού στην ουσία είναι το πραγματικό ποσοστό θνησιμότητας q_x (Hatzopoulos, 1997).

Η τυχαία μεταβλητή D_x που είναι το άθροισμα όλων των επιτυχιών οι οποίες είναι ανεξάρτητες, κατανέμεται ως διωνυμική κατανομή με μέση τιμή $r_x^i q_x$ και διακύμανση $r_x^i q_x (1 - q_x)$, δηλαδή ισχύει

$$D_x \sim Bi(r_x^i, q_x).$$

Για να ορίσουμε το κατάλληλο GLM, με τυχαία μεταβλητή απόκρισης την σειρά $\{D_x\}$, παίρνουμε

$$E(D_x) = m_x = r_x^i q_x \text{ και } \text{var}(D_x) = \frac{fV(m_x)}{w_x} = m_x\left(1 - \frac{m_x}{r_x^i}\right)$$

όπου $f = 1$, $w_x = 1$ και $V(m_x) = m_x\left(1 - \frac{m_x}{r_x^i}\right)$.

Η κλιμακωτή απόκλιση δίνεται από τη σχέση

$$S(c, f) = 2\sum_{x=1}^{n}\left\{d_x \log\frac{d_x}{\hat{m}_x} + (r_x^i - d_x)\log\frac{r_x^i - d_x}{r_x^i - \hat{m}_x}\right\},$$

όπου $\hat{m}_x = r_x^i \hat{q}_x$.

Η εφαρμογή του παραπάνω GLM μπορεί να γίνει χρησιμοποιώντας ως συνδέσμους τα παρακάτω:

α) Τον complementary log – log σύνδεσμο, όπου $q_x = 1 - \exp\{-\exp\{h_x\}\}$.

β) Τον logit σύνδεσμο, όπου $q_x = \dfrac{\exp\{h_x\}}{1+\exp\{h_x\}}$.

γ) Τον probit σύνδεσμο, όπου $q_x = \Phi(h_x)$ και Φ είναι η αθροιστική συνάρτηση κατανομής της τυπικής κανονικής κατανομής $N(0,1)$.

Ο Hatzopoulos (1997) βασιζόμενος στον αρχικό χρόνο έκθεσης στον κίνδυνο μοντελοποιεί τα ασφαλιστήρια συμβόλαια ως σύνθετη διωνυμική κατανομή. Επίσης κάνει μια παρουσίαση των μοντέλων και των συνδέσμων που μπορούν να χρησιμοποιηθούν για τη διαχρονική εξομάλυνση των ποσοστών θνησιμότητας στην ηλικία x μέσω των GLM.

Ο λογαριθμικός σύνδεσμος $h_{x_t} = \log m_{x_t}$ αποτελεί μια ικανοποιητική επιλογή. Όταν χρησιμοποιείται πολυωνυμικού τύπου πρόβλεψη, δηλαδή όταν ισχύει

$$\log m_{x_t} = \sum_{j=0}^{k} b_j L_j(x') + \sum_{i=1}^{r} a_i t'^i ,$$

δίνει τη μικρότερη κλιμακωτή απόκλιση. Τα x', t' είναι οι μετασχηματισμοί των x και t στο διάστημα $[x, x+1]$. Συγκεκριμένα ισχύει $x' = \dfrac{x - c_x}{w_x}$ και $t' = \dfrac{x - c_t}{w_t}$ όπου $c_i = \dfrac{i_{\min} + i_{\max}}{2}$ και $w_i = \dfrac{i_{\max} - i_{\min}}{2}$, για $i = x, t$. $L_j(s)$ είναι τα Legendre πολυώνυμα j βαθμού για τα οποία ισχύει

$$L_0(s) = 1,\ L_1(s) = s,\ (n+1)L_{j+1}(s) = (2j+1)sL_j(s) - jL_{j-1}(s),$$

όπου $s \geq 1$, ακέραιος (Renshaw et al., 1996β).

Σε συνδυασμό με τετραγωνικής μορφής spline ως πρόβλεψη για τις επιδράσεις της ηλικίας, δηλαδή

$$\log m_x = \begin{cases} a_1 + b_1 x + g_1 x^2, & x < k \\ a_2 + b_2 x + g_2 x^2, & x \geq k \end{cases}$$

για συγκεκριμένο χρόνο t και για παραμέτρους a_1, a_2, b_1, b_2, $g_1 > 0$, $g_2 < 0$ και κλασματικά πολυώνυμα της μορφής $a + bt^k$ για τις επιδράσεις του χρόνου παράγεται ένα μοντέλο με τις ελάχιστες παραμέτρους. Περισσότερα για τις συναρτήσεις splines θα δούμε στην επόμενη ενότητα.

Ο δεσμός δύναμης (*power link*), $h_{x_t} = m_{x_t}^k$, όπου $k \neq 0$, δίνει το μοντέλο με τις λιγότερες παραμέτρους αλλά με τη μεγαλύτερη απόκλιση όταν χρησιμοποιηθεί πολυωνυμική πρόβλεψη για τις επιδράσεις της ηλικίας και κλασματική πολυωνυμική πρόβλεψη για τις επιδράσεις του χρόνου. Επίσης αν χρησιμοποιηθεί πρόβλεψη τύπου τετραγωνικού πολυωνύμου τόσο για τις επιδράσεις της ηλικίας όσο και του χρόνου, παράγεται ένα ακόμη μοντέλο με λίγες παραμέτρους για κάθε έτος. Σημειώνεται ότι στις περιπτώσεις αυτές αντικείμενο εξομάλυνσης αποτελεί το $\frac{1}{m_{x+1/2}}$.

Τέλος τα προσθετικά (*additive*) μοντέλα δίνουν καλά αποτελέσματα όταν χρησιμοποιούνται τρίτου βαθμού splines για τις επιδράσεις της ηλικίας και κλασματικά πολυώνυμα για τις επιδράσεις του χρόνου. Και σε αυτή την περίπτωση στοχεύουμε στην εξομάλυνση του $\frac{1}{m_{x+1/2}}$.

Ο Hatzopoulos (1997) επιπλέον περιγράφει τη μοντελοποίηση των συνταξιούχων ως διωνυμική κατανομή με complementary log – log σύνδεσμο, πολυωνυμική πρόβλεψη για την επίδραση του χρόνου και αντίστροφη πολυωνυμική πρόβλεψη για την επίδραση της ηλικίας. Οι Renshaw and Hatzopoulos (1996) περιγράφουν την εξομάλυνση των ποσών των συντάξεων με σκοπό την πρόβλεψη της πιθανότητας θανάτου, βασιζόμενοι στην εμπειρία των συνταξιοδοτικών ποσών.

Κλείνοντας την περιγραφή της μεθόδου παρατηρούμε ότι τα GLM παρέχουν ένα περιεκτικό και επαρκώς θεωρητικό πλαίσιο για τη μοντελοποίηση και εξομάλυνση των διαφόρων ασφαλιστικών ποσοτήτων, καθώς και διάφορα στατιστικά τεστ για τον έλεγχο της εξομάλυνσης.

2.3 Εξομάλυνση μέσω συναρτήσεων splines

Σε όλες τις περιπτώσεις που έχουμε δει έως τώρα, για την εξομάλυνση των διαφόρων ασφαλιστικών δεδομένων, όπως τα ποσοστά ή η ένταση θνησιμότητας, χρησιμοποιείται ένα μόνο μαθηματικό μοντέλο το οποίο προσαρμόζεται σε όλο το εύρος των ηλικιών. Όταν όμως το εύρος των ηλικιών είναι μεγάλο, είναι προτιμότερο να χωρίζουμε τις τιμές σε υποδιαστήματα και να προσαρμόζουμε σε καθένα από αυτά διαφορετικό μοντέλο. Ιδιαίτερη προσοχή πρέπει να δίνεται στον τρόπο με τον οποίο κάθε μοντέλο συναντά το επόμενό του. Η μέθοδος αυτή ονομάζεται *εξομάλυνση μέσω splines* (London, 1985).

Το βασικότερο πλεονέκτημα της μεθόδου είναι ότι τα μαθηματικά μοντέλα που χρησιμοποιούνται σε κάθε υποδιάστημα, είναι απλούστερα από το μοντέλο που θα χρησιμοποιόταν σε όλο το εύρος των ηλικιών. Για την εξομάλυνση των ασφαλιστικών δεδομένων, τα μοντέλα που χρησιμοποιούνται είναι της μορφής πολυωνύμων τρίτου βαθμού, τα οποία συνήθως προσαρμόζονται με τη μέθοδο των ελαχίστων τετραγώνων.

Πριν προχωρήσουμε στην περιγραφή της μεθόδου, κρίνουμε σκόπιμο να κάνουμε τις παρακάτω παρατηρήσεις: Splines είναι συναρτήσεις, οι οποίες αποτελούνται από δυο ή περισσότερες συνεχόμενες τρίτου βαθμού (*cubic*) καμπύλες ή τόξα (*arcs*) τα οποία ενώνονται μεταξύ τους με ομαλό τρόπο. Στην περίπτωσή μας, μια συνάρτηση spline είναι η συνάρτηση της ηλικίας x, $x = 1, 2, ..., n$ που δίνει τις εξομαλυμένες τιμές v_x. Δεν απαιτείται να ισχύει $u_x = v_x$ για κάποια τιμή του x καθώς επίσης δεν είναι απαραίτητο να γνωρίζουμε τις τιμές u_x για όλα τα x.

Ας δούμε τώρα ορισμένα είδη συναρτήσεων spline.

2.3.1 Τρίτου βαθμού συνάρτηση ελαχίστων τετραγώνων

Έστω ότι γνωρίζουμε τις αρχικές εκτιμήσεις των τιμών u_x σε ένα διάστημα ηλικιών της μορφής $[a, b]$. Έστω επίσης ότι οι εξομαλυμένες τιμές v_x εκφράζονται ως τρίτου βαθμού συναρτήσεις, δηλαδή $v_x = c_0 + c_1 x + c_2 x^2 + c_3 x^3$, όπου c_0, c_1, c_2 και c_3 είναι άγνωστες παράμετροι.

Ορίζουμε τη στατιστική συνάρτηση

$$SS = \sum_{x=a}^{b} w_x (u_x - v_x)^2 = \sum_{x=a}^{b} w_x (u_x - c_0 - c_1 x - c_2 x^2 - c_3 x^3)^2 , \qquad (2.6)$$

όπου w_x είναι κατάλληλα βάρη. Μια επιλογή για τα βάρη είναι ο αντίστροφος (*reciprocal*) της ασυμπτωτικής διακύμανσης της τυχαίας μεταβλητής U_x.

Εξισώνοντας με μηδέν τις μερικές παραγώγους, ως προς τις παραμέτρους c_0, c_1, c_2 και c_3, της σχέσης (2.6) και λύνοντας τις κανονικές εξισώσεις, παίρνουμε τις εκτιμήσεις ελαχίστων τετραγώνων για τις παραμέτρους του μοντέλου. Σημειώνουμε ότι σε πινακική μορφή οι εξισώσεις γράφονται ως

$$\mathbf{X}^T\mathbf{W}\mathbf{X}\mathbf{c} = \mathbf{X}^T\mathbf{W}\mathbf{u}\,, \tag{2.7}$$

όπου $\mathbf{X}$ είναι ο $m\times 4$ πίνακας

$$\begin{bmatrix} 1 & a & a^2 & a^3 \\ 1 & a+1 & (a+1)^2 & (a+1)^3 \\ \mathrm{M} & \mathrm{M} & \mathrm{M} & \mathrm{M} \\ 1 & b & b^2 & b^3 \end{bmatrix},$$

m είναι ο αριθμός των αρχικών τιμών που γνωρίζουμε στο διάστημα ηλικιών $[a,b]$, $\mathbf{W}$ είναι ο $m\times m$ διαγώνιος πίνακας που περιέχει τα βάρη, $\mathbf{c}$ είναι το διάνυσμα των παραμέτρων, $\mathbf{u}$ είναι το διάνυσμα των m αρχικών εκτιμήσεων u_x και το T δηλώνει την αντιστροφή του πίνακα.

2.3.2 Δίτοξη τρίτου βαθμού spline

Σε περίπτωση που αποδειχθεί ότι η απλή τρίτου βαθμού συνάρτηση δεν αναπαριστά ικανοποιητικά τα δεδομένα μας, μπορούμε να θεωρήσουμε δυο τέτοιες συναρτήσεις $p_0(x)$ και $p_1(x)$, οι οποίες ενώνονται στο σημείο $x = k$. Το σημείο k καλείται *δεσμός* (*knot*) και για την τιμή αυτή δεν πρέπει να υπάρχει η αρχική εκτίμηση u_x.

Συνεπώς οι εξομαλυμένες τιμές v_x δίνονται από τη σχέση

$$v_x = \begin{cases} p_0(x), & a \le x \le k \\ p_1(x), & k \le x \le b \end{cases}$$

και η σχέση (2.6) γράφεται ως εξής:

$$SS = \sum_{x=a}^{b} w_x (u_x - v_x)^2 = \sum_{x=a}^{h} w_x (u_x - p_0(x))^2 + \sum_{x=h+1}^{b} w_x (u_x - p_1(x))^2\,, \tag{2.8}$$

όπου h είναι η μεγαλύτερη τιμή του x, για την οποία γνωρίζουμε την τιμή u_x, και είναι μικρότερη του k.

Για την επίτευξη ομαλής σύνδεσης μεταξύ των $p_0(x)$ και $p_1(x)$, απαιτούμε την ισχύ των παρακάτω σχέσεων:

$$p_0(k) = p_1(k) \qquad (2.9a)$$

$$p_0'(k) = p_1'(k) \qquad (2.9b)$$

$$p_0''(k) = p_1''(k) \qquad (2.9c)$$

όπου ο τόνος δηλώνει παραγώγιση ως προς x, και οι οποίες εξασφαλίζουν ότι η συνάρτηση spline είναι δυο φορές διαφορίσιμη, αφού καθένα από τα δυο κομμάτια της, δηλαδή οι $p_0(x)$ και $p_1(x)$, είναι δυο φορές διαφορίσιμα.

Θεωρούμε τις συναρτήσεις

$$p_0(x) = c_0 + c_1x + c_2x^2 + c_3x^3$$

και

$$p_1(x) = c_0 + c_1x + c_2x^2 + c_3x^3 + c_4(x-k)^3$$

για τις οποίες εύκολα αποδεικνύεται ότι ισχύουν οι σχέσεις (2.9a) - (2.9c). Έτσι η σχέση (2.8) γίνεται

$$SS = \sum_{x=a}^{h} w_x(u_x - c_0 - c_1x - c_2x^2 - c_3x^3)^2$$

$$+ \sum_{x=h+1}^{b} w_x\left(u_x - c_0 - c_1x - c_2x^2 - c_3x^3 - c_4(x-k)^3\right)^2$$

η οποία λύνεται με τον ίδιο τρόπο που λύνεται η σχέση (2.6).

2.3.3 Τρίτοξη τρίτου βαθμού spline

Στην περίπτωση αυτή χρησιμοποιούμε δυο δεσμούς k_1 και k_2 έτσι ώστε το διάστημα $[a,b]$ να χωρίζεται σε τρία υποδιαστήματα, στα οποία προσαρμόζονται οι συναρτήσεις $p_0(x)$, $p_1(x)$ και $p_2(x)$ αντίστοιχα. Για την ομαλή σύνδεσή τους εκτός από τις σχέσεις (2.9a) - (2.9c) όπου αντί για k θέτουμε k_1, απαιτούμε και την ισχύ των σχέσεων:

$$p_1(k_2) = p_2(k_2),$$

$$p_1'(k_2) = p_2'(k_2),$$

$$p_1''(k_2) = p_2''(k_2),$$

όπου ο τόνος δηλώνει παραγώγιση ως προς x.

Ορίζουμε τις συναρτήσεις

$$p_0(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3,$$

$$p_1(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 (x - k_1)^3$$

και

$$p_2(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 (x - k_1)^3 + c_5 (x - k_2)^3$$

και η σχέση (2.6) γίνεται

$$SS = \sum_{x=a}^{b} w_x (u_x - v_x)^2 = \sum_{x=a}^{h_1} w_x \left(u_x - p_0(x)\right)^2 + \sum_{x=h_1+1}^{h_2} w_x \left(u_x - p_1(x)\right)^2 + \sum_{x=h_2+1}^{b} w_x \left(u_x - p_2(x)\right)^2 ,$$

όπου h_1 είναι η μεγαλύτερη τιμή του x, η οποία είναι μικρότερη του k_1 και αντίστοιχα h_2 είναι η μεγαλύτερη τιμή του x η οποία είναι μικρότερη του k_2. Τόσο για την τιμή h_1 όσο και την h_2 πρέπει να γνωρίζουμε τις αντίστοιχες τιμές u_x.

Για την εκτίμηση των παραμέτρων ακολουθούμε τη γνωστή διαδικασία.

Είναι προφανές ότι η μέθοδος μπορεί να γενικευθεί, για την περίπτωση των $r+1$ υποδιαστημάτων, θεωρώντας r δεσμούς, με $r < n$. Για τη γενίκευση των εξισώσεων (2.7) ο ενδιαφερόμενος παραπέμπεται στον London (1985).

Πριν κλείσουμε το θέμα της εξομάλυνσης ποσοστών θνησιμότητας με splines, ας δούμε κάποιες παρατηρήσεις σχετικά με τους δεσμούς (London, 1985). Για την επιλογή του αριθμού και της θέσης των δεσμών δεν υπάρχουν κάποιοι αυστηροί ή δεσμευτικοί κανόνες. Έτσι μπορεί να χρησιμοποιηθεί η προσωπική κρίση του ερευνητή όσο αυθαίρετο και αν είναι αυτό. Για παράδειγμα μπορεί να χρησιμοποιηθεί η γραφική παράσταση των τιμών u_x, για να παρατηρήσουμε αλλαγές στο σχήμα της και να χρησιμοποιήσουμε τα σημεία αλλαγής ως δεσμούς.

Διαισθητικά όσο αυξάνεται ο αριθμός των δεσμών, τόσο καλύτερη γίνεται η προσαρμογή. Βέβαια αν πάρουμε το μέγιστο αριθμό δεσμών, δηλαδή n δεσμούς, θα πάρουμε $u_x = v_x$ οπότε δεν θα έχουμε εξομάλυνση.

Σκοπός μας είναι να επιτύχουμε όσο το δυνατόν καλύτερη εξομάλυνση. Συνεπώς το μέτρο SS, που μπορεί να ερμηνευθεί ως X^2 στατιστικό, μπορεί να χρησιμοποιηθεί ως συγκριτικό μέτρο για την εξομάλυνση. Τέλος, μια ενδιαφέρουσα εναλλακτική λύση είναι να θεωρηθεί το SS σαν συνάρτηση των δεσμών, έτσι ώστε αυτοί να εκτιμώνται ελαχιστοποιώντας το SS, οπότε μειώνονται αντίστοιχα και οι βαθμοί ελευθερίας του X^2 τεστ. Εννοείται ότι οι δεσμοί θα πρέπει να ανήκουν στο διάστημα $[a, b]$.

Τέλος να σημειώσουμε ότι εφόσον εκτιμούμε τις παραμέτρους των συναρτήσεων splines θα πρέπει να αφαιρούμε και τον αντίστοιχο αριθμό βαθμών ελευθερίας κατά την εφαρμογή του X^2 τεστ.

2.4 Εξομάλυνση μέσω παρεμβολής ομαλής σύνδεσης

Μια εναλλακτική μέθοδος εξομάλυνσης, η οποία μοιάζει σε ορισμένα σημεία με τη μέθοδο των splines, είναι η μέθοδος της *παρεμβολής ομαλής σύνδεσης* (*smooth – junction interpolation*) (London, 1985). Για τη μέθοδο αυτή, αρκεί να γνωρίζουμε έναν περιορισμένο αριθμό αρχικών εκτιμήσεων u_x, όπου τα u_x μπορεί να είναι ο αριθμός ατόμων σε κίνδυνο, ο αριθμός θανάτων, τα ποσοστά θνησιμότητας ή η ένταση θνησιμότητας και στη συνέχεια να προσαρμόσουμε ένα διαφορετικό τόξο (*arc*) παρεμβολής σε κάθε υποδιάστημα που δημιουργείται από τις γνωστές τιμές u_x για να πάρουμε εξομαλυμένες τιμές για τα ενδιάμεσα σημεία. Κατ' αντιστοιχία με τη μέθοδο των splines, οι τιμές u_x είναι οι *δεσμοί* της παρεμβολής, και τα γειτονικά τόξα θα πρέπει να ενώνονται μεταξύ τους ομαλά.

Συνοπτικά μπορούμε να αναφέρουμε ότι με τη μέθοδο αυτή, επιθυμούμε να πάρουμε εξομαλυμένες τιμές v_x για τις ηλικίες x, που βρίσκονται ενδιάμεσα σε κάποιες, για τις οποίες γνωρίζουμε τις αρχικές εκτιμήσεις u_x. Ένα τόξο παρεμβολής προσαρμόζεται σε κάθε υποδιάστημα που καθορίζεται από τις αρχικές εκτιμήσεις, οι οποίες λέγονται *κεντρικά σημεία* (*pivotal points*) και τα γειτονικά τόξα θα πρέπει να ενώνονται μεταξύ τους με ίσες τεταγμένες (*ordinates*) και να έχουν τουλάχιστον ίσες πρώτες παραγώγους.

2.4.1 Τύπος της παρεμβολής

Όλα τα μοντέλα παρεμβολής είναι δυνατόν να γραφούν στη γενική μορφή του Everret (Kellison, 1975), η οποία είναι η εξής:

$$v_{x+s} = F(s)u_{x+1} + F(t)u_x, \quad 0 \le s \le 1, \; t = 1 - s, \tag{2.10}$$

όπου

$$F(s) = A(s) + B(s)d^2 + C(s)d^4 + \ldots$$

$x = 1, 2, \ldots, n$, u_x είναι οι αρχικές εκτιμήσεις, v_{x+s} είναι οι εξομαλυμένες τιμές και με d συμβολίζονται οι κεντρικές διαφορές. Δηλαδή για $z = 2$ και 4 έχουμε αντίστοιχα,

$$d^2u_x = u_{x+1} - 2u_x + u_{x-1} \text{ και } d^4u_x = u_{x+2} - 4u_{x+1} + 6u_x - 4u_{x-1} + u_{x-2}.$$

Οι τιμές …, u_{x-1}, u_x, u_{x+1}, … είναι ισαπέχουσες και ύστερα από μετασχηματισμό μπορούν να θεωρηθούν ότι εμφανίζονται σε μοναδιαία διαστήματα. Τα $A(s)$, $B(s)$, $C(s)$, … είναι συναρτήσεις του s ενώ το $F(s)$ δεν είναι συνάρτηση αλλά ένας τελεστής (*operator*)

για τα u_x, ο οποίος σαν συντελεστές των διαφόρων τάξεων των διαφορών d, έχει συναρτήσεις του s.

Εάν το $F(s)$ τελειώνει με όρο που περιέχει το d^{2m}, ο τύπος χρησιμοποιεί $2m+2$ τιμές u_x, δηλαδή τις $u_{x-m}, u_{x-m-1}, \ldots, u_{x+m+1}$, και συνεπώς καλείται *τύπος παρεμβολής* $2m+2$ *σημείων*.

Γράφοντας τον τύπο της παρεμβολής στη μορφή (2.10), υποθέτουμε ότι είναι συμμετρικός, που σημαίνει ότι δίνει τις ίδιες παρεμβλημένες τιμές για κάθε x είτε η παρεμβολή γίνει προς τα εμπρός είτε προς τα πίσω.

Η πιο απλή μορφή για γραμμική παρεμβολή είναι ο τύπος δυο σημείων που προκύπτει στην περίπτωση όπου $F(s) = s$, οπότε ισχύει

$$v_{x+s} = su_{x+1} + tu_x, \quad 0 \le s \le 1, \quad t = 1 - s.$$

Οι συναρτήσεις $A(s)$, $B(s)$, $C(s)$, … μπορεί να είναι οποιεσδήποτε συνεχείς συναρτήσεις του s, οι οποίες όμως θα πρέπει να έχουν τον απαιτούμενο αριθμό παραγώγων. Συνήθως όμως χρησιμοποιούνται πολυώνυμα του s μικρού βαθμού, τα οποία θα πρέπει να ικανοποιούν κάποιες συνθήκες ή ιδιότητες που θα δούμε αμέσως παρακάτω. Οι τύποι αυτοί μπορούν να χρησιμοποιηθούν για παρεμβολή αλλά και για εξομάλυνση τιμών, όπως είναι τα ποσοστά ή η ένταση θνησιμότητας.

2.4.2 Ιδιότητες των τύπων παρεμβολής

Ιδιότητα της ομαλής σύνδεσης

Χρησιμοποιώντας τύπους της μορφής Everret, οι οποίοι θα τελειώνουν σε διαφορές $2^{ης}$ ή $4^{ης}$ τάξης, και επιλέγοντας ως συναρτήσεις $A(s)$, $B(s)$, $C(s)$, … κάποια πολυώνυμα, υπονοείται ότι η τιμή v_{x+s} είναι επίσης πολυώνυμο του s, το οποίο χρησιμοποιείται για τον υπολογισμό των εξομαλυμένων (παρεμβλημένων) τιμών στο διάστημα $[x, x+1]$. Συμβολίζοντας με y το $x+s$ και θεωρώντας το v_y ως συνάρτηση του y, τότε το v_y είναι μια spline και ισχύει

$$v_y = \begin{cases} \text{M} \\ p_{x-1}(y), \quad x-1 \le y \le x \\ \\ p_x(y), \quad x \le y \le x+1 \\ \text{M} \end{cases}$$

Σημειώνουμε ότι συνήθως το πολυώνυμο $p_{x-1}(y)$ είναι διαφορετικό από το $p_x(y)$.

Για να πάρουμε μια ομαλή καμπύλη, αν θεωρήσουμε το διάγραμμα των v_y, θα πρέπει τα γειτονικά τόξα των πολυωνύμων να ενώνονται στις άκρες τους, δηλαδή να ισχύει

$$p_{x-1}(x) = p_x(x),$$

όπου x είναι η ηλικία για την οποία γνωρίζουμε την αρχική εκτίμηση u_x. Για το λόγο αυτό πρέπει να ισχύει η σχέση

$$v_{x-1+s/s=1} = v_{x+s/s=0}.$$

Για τη γενική μορφή των τύπων Everret, ισχύει

$$v_{x-1+s/s=1} = A(1)u_x + B(1)d^2u_x + C(1)d^4u_x + \ldots + A(0)u_{x-1} + B(0)d^2u_{x-1} + C(0)d^4u_{x-1} + \ldots$$

και

$$v_{x+s/s=0} = A(0)u_{x+1} + B(0)d^2u_{x+1} + C(0)d^4u_{x+1} + \ldots + A(1)u_x + B(1)d^2u_x + C(1)d^4u_x + \ldots$$

Εξισώνοντας τις δυο παραπάνω σχέσεις παίρνουμε

$$A(0)u_{x-1} + B(0)d^2u_{x-1} + C(0)d^4u_{x-1} + \ldots = A(0)u_{x+1} + B(0)d^2u_{x+1} + C(0)d^4u_{x+1} + \ldots$$

με την ισότητα να ισχύει για όλες τις τιμές των u_x αν και μόνο αν

$$A(0) = B(0) = C(0) = \ldots = 0. \tag{2.11}$$

Για την επίτευξη γενικής ομαλότητας για τα v_y απαιτείται επιπλέον η ισχύς των σχέσεων

$$p'_{x-1}(x) = p'_x(x)$$

ή/και

$$p''_{x-1}(x) = p''_x(x),$$

όπου ο τόνος δηλώνει παραγώγιση ως προς x.

Σε περίπτωση που ισχύουν οι σχέσεις

$$p_{x-1}(x) = p_x(x) \text{ και } p'_{x-1}(x) = p'_x(x)$$

ο τύπος της παρεμβολής ονομάζεται *εφαπτόμενος* (*tangential*) ενώ αν ισχύει επιπλέον και η σχέση

$$p''_{x-1}(x) = p''_x(x)$$

ο τύπος καλείται *ισχυρά εφαπτόμενος* (*osculatory*).

Για να ισχύει η ισότητα των πρώτων παραγώγων θα πρέπει να ισχύει η σχέση

$$v'_{x-1+s/s=1} = v'_{x+s/s=0} \tag{2.12}$$

με την παραγώγιση ως προς s. Επίσης ισχύει

$$F'(s) = A'(s) + B'(s)d^2 + C'(s)d^4 + \ldots$$

και για $t = 1 - s$ ισχύει

$$\frac{\partial}{\partial s}F(t) = -F'(t).$$

Συνεπώς η σχέση (2.12) δίνει

$$F'(1)u_x - F'(0)u_{x-1} = F'(0)u_{x+1} - F'(1)u_x$$

ή

$$2F'(1)u_x = F'(0)(u_{x+1} + u_{x-1}) = F'(0)(2 + d^2)u_x$$

αφού $d^2u_x = u_{x+1} - 2u_x + u_{x-1}$ είναι η $2^{\eta\varsigma}$ τάξης κεντρική διαφορά. Έτσι ο τύπος της παρεμβολής θα είναι εφαπτόμενος αν και μόνο αν

$$2F'(1) = F'(0)(2 + d^2). \tag{2.13}$$

Εργαζόμενοι με τον ίδιο τρόπο και ξεκινώντας από τη σχέση $v''_{x-1+s/s=1} = v''_{x+s/s=0}$, ο τύπος θα είναι ισχυρά εφαπτόμενος αν και μόνο αν

$$F''(0)(u_{x+1} - u_{x-1}) = 0$$

και επιπλέον ισχύουν οι σχέσεις (2.13) και $F''(0) = 0$.

Ιδιότητα ομαλότητας έναντι αναπαραγωγικότητας

Σε περίπτωση που ο τύπος της παρεμβολής δίνει τιμές $v_x = u_x$, για τις ηλικίες x που γνωρίζουμε τις αρχικές εκτιμήσεις u_x, καλείται *τύπος αναπαραγωγής* ενώ σε αντίθετη περίπτωση ονομάζεται *τύπος ομαλότητας*.

Για να επιτύχουμε μια συνεχή καμπύλη, θα πρέπει να ισχύει η σχέση (2.11), οπότε

$$v_x = v_{x+s/s=0} = A(1)u_x + B(1)d^2u_x + C(1)d^4u_x + \ldots.$$

Για έναν αναπαραγωγικό τύπο η παραπάνω σχέση δείχνει ότι θα ισχύει $v_x = u_x$, αν και μόνο αν ισχύουν οι σχέσεις

$$A(1) = 1,\ B(1) = C(1) = \ldots = 0$$

και (2.11).

Ιδιότητα του βαθμού ακρίβειας

Μπορούμε να πούμε ότι ένας τύπος παρεμβολής είναι *ακριβής* για πολυώνυμο βαθμού z αν και μόνο αν δίνει ακριβή αποτελέσματα όταν χρησιμοποιείται για την παρεμβολή τιμών ενός πολυωνύμου βαθμού μικρότερου ή ίσου του z.

Χρησιμοποιώντας τη σχέση για τις διαφορές

$$\Delta^{2m} u_{y+1} = \Delta^{2m} u_y + \Delta^{2m+1} u_y$$

η σχέση (2.10) μπορεί να γραφεί ως

$$v_{x+s} = \left[A(s) + A(t)\right] u_x + A(s)\Delta u_x + \left[B(s) + B(t)\right]\Delta^2 u_{x-1}$$

$$+ B(s)\Delta^3 u_{x-1} + \left[C(s) + C(t)\right]\Delta^4 u_{x-2} + C(s)\Delta^5 u_{x-2} + \ldots, \qquad (2.14)$$

η οποία είναι της μορφής τύπου Gauss (Kellison, 1975). Συνεπώς το v_{x+s} θα είναι ακριβές για πολυώνυμα βαθμού ίσου ή μικρότερου του z αν και μόνο αν συμφωνεί με τη σχέση (2.14) και περιλαμβάνει τον όρο της διαφοράς z τάξεως.

2.4.3 Βασικοί τύποι παρεμβολής

Συνήθως χρησιμοποιούνται τύποι τεσσάρων ή έξι σημείων και προσδιορίζονται από τις ιδιότητες που θέλουμε αυτοί να έχουν.

Για έναν τύπο της μορφής Everett τεσσάρων σημείων με ένα βαθμό ακρίβειας (δηλαδή να είναι ακριβής για γραμμικές συναρτήσεις), εφαπτόμενο και με μια παράμετρο που να ελέγχει την ιδιότητα της ομαλότητας, επιλέγουμε $F(s) = A(s) + B(s)d^2$, όπου $A(s) = s$ και $B(s) = \left(3L - \frac{1}{2}\right)s^2 + \left(\frac{1}{2} - 2L\right)s^3$, με $L = B(1)$.

Ειδικές περιπτώσεις του παραπάνω τύπου έχουμε, επιλέγοντας $B(s) = \frac{1}{2}s^2(s-1)$, οπότε παίρνουμε τον τύπο Karup – King, $B(s) = \frac{1}{4}s^2$, οπότε ο τύπος είναι ο μοναδικός που είναι

τετραγωνικός και $B(s)=\frac{1}{6}s^3$, όπου δίνει τον μοναδικό ισχυρά εφαπτόμενο τύπο με τέσσερα σημεία.

Για τον προσδιορισμό ενός τύπου έξι σημείων της μορφής Everett με τρεις βαθμούς ακρίβειας, ισχυρά εφαπτόμενο και με έναν παράγοντα που να ελέγχει την ιδιότητα της ομαλότητας, παίρνουμε $F(s)=A(s)+B(s)d^2+C(s)d^4$, με $A(s)=s$, $B(s)=\frac{1}{6}s(s^2-1)$ και $C(s)=\left(4M+\frac{1}{12}\right)s^3-\left(3M+\frac{1}{12}\right)s^4$, όπου $M=C(1)$.

Ειδικές περιπτώσεις του παραπάνω τύπου, παίρνουμε για $C(s)=\frac{1}{12}s^3(1-s)$ που δίνει αναπαραγωγικό τύπο, $C(s)=-\frac{1}{36}s^3$ που δίνει το μόνο τετραγωνικό τύπο και $C(s)=-\frac{1}{48}s^4$.

Για περισσότερες πληροφορίες για την εξαγωγή των παραπάνω τύπων παρεμβολής ο αναγνώστης παραπέμπεται στον London (1985).

Ας δούμε τώρα πώς μπορεί να γίνει η επιλογή δεσμών ή όπως αλλιώς λέγονται των κεντρικών τιμών. Σε περίπτωση που έχουμε τις τιμές u_x για όλα τα έτη x, το πρώτο βήμα είναι να τα ομαδοποιήσουμε. Στη συνέχεια επιλέγουμε τις κεντρικές τιμές βάσει του τύπου του King (Benjamin and Pollard, 1980).

Όταν έχουμε δεδομένα θνησιμότητας, καλό είναι να επιλέγουμε ξεχωριστά κεντρικές τιμές για τους θανάτους και ξεχωριστά για τα άτομα σε κίνδυνο. Στη συνέχεια διαιρώντας αυτά, παίρνουμε κεντρικές τιμές για τα ποσοστά θνησιμότητας τα οποία θα χρησιμοποιήσουμε για την παρεμβολή.

Έστω ότι s_x είναι οι θάνατοι ή τα άτομα σε κίνδυνο στην ηλικία x και συμβολίζουμε με w_x το άθροισμα πέντε συνεχόμενων s_y με κέντρο το x, δηλαδή

$$w_x = \sum_{y=x-2}^{x+2} s_y \ .$$

Ο τύπος του King, διορθωμένος για πέμπτες διαφορές είναι

$$s_x^P = 0.2w_x - 0.008(w_{x-5} - 2w_x + w_{x+5}) + 0.000896(w_{x-10} - 4w_{x-5} + 6w_x - 4w_{x+5} + w_{x+10}) \ .$$

Γενικά χρησιμοποιούνται οι δυο πρώτοι όροι, οπότε η προηγούμενη σχέση γράφεται ως

$$s_x^P = 0.2w_x - 0.008(w_{x-5} - 2w_x + w_{x+5}),$$

οπότε ο τύπος είναι διορθωμένος για τρίτες διαφορές. Επειδή ο τύπος δεν δίνει κεντρικά σημεία για το πρώτο και τελευταίο πενταετές άθροισμα, έστω w_a και w_b, χρησιμοποιούνται αντίστοιχα οι τύποι

$$s_a^P = 0.2w_a - 0.008(w_a - 2w_{a+5} + w_{a+10})$$

και

$$s_b^P = 0.2w_b - 0.008(w_b - 2w_{b-5} + w_{b-10})$$

οι οποίοι είναι διορθωμένοι για δεύτερες διαφορές.

Σημειώνουμε επίσης ότι για κάθε παράμετρο που εκτιμούμε, αφαιρούμε και ένα βαθμό ελευθερίας από την κατανομή του X^2 τεστ.

Κεφάλαιο 3

ΜΗ ΠΑΡΑΜΕΤΡΙΚΕΣ ΜΕΘΟΔΟΙ ΕΞΟΜΑΛΥΝΣΗΣ

Σε αντιδιαστολή με τις παραμετρικές μεθόδους εξομάλυνσης, οι μη παραμετρικές δεν προϋποθέτουν ότι οι αρχικές εκτιμήσεις των ποσοστών θνησιμότητας ή της έντασης θνησιμότητας περιγράφονται από κάποιο παραμετρικό μοντέλο. Απλά συνδυάζουν δεδομένα σε διαφορετικές τιμές της ηλικίας x και με κατάλληλες τεχνικές προκύπτουν οι εξομαλυμένες τιμές. Η ομαλότητα δεν θεωρείται εξασφαλισμένη και η επίτευξη του κατάλληλου βαθμού ομαλότητας επιτυγχάνεται με την κατάλληλη επιλογή των τιμών κάποιων παραμέτρων όπως θα δούμε στη συνέχεια. Η μεροληψία που οι μέθοδοι αυτές περιλαμβάνουν, μπορεί να μειωθεί ελαττώνοντας τη δειγματική διασπορά.

Στην κατηγορία αυτή ανήκει η γραφική μέθοδος, η εξομάλυνση με αναφορά σε έναν τυπικό πίνακα θνησιμότητας, η εξομάλυνση μέσω κινητών σταθμισμένων μέσων, η εξομάλυνση των Whittaker και Henderson, καθώς και η μπεϋζιανή εξομάλυνση. Επίσης μη παραμετρική μέθοδος είναι η εξομάλυνση μέσω εκτιμητών πυρήνα.

3.1 Γραφική μέθοδος

Η *γραφική μέθοδος* είναι μια καθιερωμένη και ευρύτατα χρησιμοποιούμενη μέθοδος για την εξομάλυνση πινάκων θνησιμότητας (Benjamin and Pollard, 1980). Στις μέρες μας, βέβαια χρησιμοποιείται περισσότερο για την εξομάλυνση των συνταξιοδοτικών ποσών και των ποσοστών αποχωρήσεων ή συνταξιοδότησης.

Για την εφαρμογή της μεθόδου στα ποσοστά θνησιμότητας, απαιτείται η γνώση των αρχικών εκτιμήσεων ή αδρών ποσοστών θνησιμότητας

$$\overset{\circ}{q}_x = \frac{d_x}{l_x},\ x = 1, 2, ..., n,$$

τα οποία παρουσιάζονται γραφικά ως σημεία. Στη συνέχεια χαράσσεται μια ομαλή καμπύλη όσο το δυνατόν γίνεται πιο κοντά στα σημεία και οι εξομαλυμένες τιμές των ποσοστών θνησιμότητας διαβάζονται για κάθε ηλικία από την καμπύλη.

Επιπρόσθετα για την καλύτερη χάραξη της καμπύλης, μπορούμε να χρησιμοποιήσουμε την παρακάτω διαδικασία:

Σε περίπτωση που ο πραγματικός αριθμός των θανάτων d_x σε κάθε ηλικία υπερβαίνει τους 10, οι Benjamin and Pollard (1980) υπολογίζουν ένα προσεγγιστικό 95% διάστημα εμπιστοσύνης για κάθε ηλικία ως εξής:

$$\overset{0}{q}_x \pm \frac{2\sqrt{d_x}}{l_x}. \tag{3.1}$$

Το διάστημα αυτό δικαιολογείται με το ακόλουθο σκεπτικό:
Θεωρώντας ότι ο αριθμός των θανάτων ακολουθεί τη διωνυμική κατανομή με παραμέτρους l_x και q_x και για μεγάλο μέγεθος δείγματος $(l_x \geq 30)$ ισχύει

$$\frac{d_x - l_x q_x}{\sqrt{l_x q_x (1-q_x)}} \sim N(0,1).$$

Συνεπώς το 95% διάστημα εμπιστοσύνης για το ποσοστό $\overset{0}{q}_x$ είναι

$$\overset{0}{q}_x \pm z_{a/2}\sqrt{\frac{\overset{0}{q}_x(1-\overset{0}{q}_x)}{l_x}}. \tag{3.2}$$

Με την προϋπόθεση ότι $1-\overset{0}{q}_x \approx 1$ και για $z_{a/2} = 1.96 \approx 2$, προκύπτει ότι προσεγγιστικά το 95% διάστημα εμπιστοσύνης δίνεται από τη σχέση (3.1). Βέβαια εφόσον γνωρίζουμε όλες τις ποσότητες της σχέσης (3.2), μπορούμε να χρησιμοποιήσουμε αυτή για να κατασκευάσουμε το διάστημα εμπιστοσύνης.

Στη συνέχεια τα διαστήματα εμπιστοσύνης χαράσσονται στη γραφική παράσταση και ενώνονται τα σημεία έτσι ώστε να έχουμε μια ζώνη εμπιστοσύνης για την κατανομή των ποσοστών, η οποία λειτουργεί ως οδηγός για τη χάραξη της καμπύλης. Σύμφωνα με τους Benjamin and Pollard, (1980) «η καμπύλη δεν πρέπει να βγαίνει εκτός των ορίων της ζώνης εμπιστοσύνης περισσότερο της μιας φοράς κάθε είκοσι σημεία».

Εάν τα δεδομένα είναι σποραδικά θα πρέπει να γίνει ομαδοποίηση για την αποφυγή δειγματικών σφαλμάτων. Καλό είναι η επιλογή των διαστημάτων να γίνεται έτσι ώστε σε κάθε διάστημα τα ποσοστά να προχωρούν ομαλά. Στην περίπτωση που τα διαστήματα είναι ομοιόμορφα, υπολογίζουμε τα διαστήματα εμπιστοσύνης με τον τύπο (3.1) αλλά χρησιμοποιώντας τα $\overset{0}{q}$, d και l για κάθε διάστημα.

Σε περίπτωση εξομάλυνσης της έντασης θνησιμότητας, ακολουθούμε ακριβώς την ίδια διαδικασία, θεωρώντας ότι ο αριθμός των θανάτων d_x ακολουθεί κατανομή Poisson με παράμετρο $r_x^c m_{x+1/2}$ η οποία προσεγγίζεται από τη $N(r_x^c m_{x+1/2}, r_x^c m_{x+1/2})$. Έτσι το ασυμπτωτικό 95% διάστημα εμπιστοσύνης για κάθε ηλικία υπολογίζεται ως

$$\overset{o}{m}_{x+1/2} \pm \frac{2\sqrt{d_x}}{r_x^c},\ x = 1, 2, ..., n,$$

όπου r_x^c είναι ο κεντρικός χρόνος έκθεσης στον κίνδυνο.

Το *πλεονέκτημα* της γραφικής μεθόδου είναι ότι μπορεί να χρησιμοποιηθεί και όταν τα δεδομένα είναι λιγοστά, οπότε δεν μπορούν να χρησιμοποιηθούν άλλες μέθοδοι εξομάλυνσης.

Τα *μειονεκτήματα* της μεθόδου είναι τα εξής:

α) απαιτείται ιδιαίτερη εμπειρία και περισσή υπομονή για τη χάραξη μιας ικανοποιητικά ομαλής καμπύλης,

β) η μέθοδος αυτή δεν επιτυγχάνει μεγάλο βαθμό ομαλότητας αφού είναι δύσκολο να διαβαστούν πολλά δεκαδικά ψηφία από ένα διάγραμμα και

γ) εμπλέκεται η προσωπική κρίση του ερευνητή πράγμα που οδηγεί σε μεροληψία.

Όσον αφορά το X^2 τεστ, οι Benjamin and Pollard, (1980), αναφέρουν ότι «για κάθε δέκα περίπου ηλικίες, θα πρέπει να αφαιρούμε 2 ή 3 βαθμούς ελευθερίας».

3.2 Εξομάλυνση με αναφορά σε τυπικό πίνακα θνησιμότητας

Μια άλλη μέθοδος εξομάλυνσης είναι αυτή *με αναφορά σε έναν τυπικό πίνακα θνησιμότητας* (Benjamin and Pollard, 1980). Εφαρμόζεται συνήθως όταν ο αναλογιστής έχει υποψίες ότι τα παρατηρούμενα δεδομένα, προέρχονται από μια κατανομή που είναι παρόμοια με αυτή ενός άλλου εξομαλυμένου και δημοσιευμένου πίνακα, ο οποίος ονομάζεται τυπικός, ή όταν τα δεδομένα είναι ανεπαρκή.

Ο πιο απλός τρόπος εφαρμογής της μεθόδου, είναι η εξομάλυνση των λόγων των αδρών τιμών u_x προς τις αντίστοιχες τιμές u_x^s του τυπικού πίνακα θνησιμότητας, δηλαδή των $\frac{u_x}{u_x^s}$, όπου u_x είναι τα ποσοστά ή η ένταση θνησιμότητας για $x = 1, 2, ..., n$. Σε περίπτωση που έχει γίνει σωστή επιλογή του τυπικού πίνακα, οι τιμές των παραπάνω λόγων θα πρέπει να βρίσκονται κοντά στη μονάδα. Μετά την εξομάλυνση αυτών των λόγων, οι τελικές εκτιμήσεις για τις πραγματικές τιμές t_x, υπολογίζονται πολλαπλασιάζοντας τους εξομαλυμένους λόγους με τις τιμές u_x^s.

Ο Lidstone το 1892 πρότεινε για την εξομάλυνση των ποσοστών θνησιμότητας, τη χρησιμοποίηση των ποσοτήτων $\log\frac{p_x^s}{\overset{0}{p}_x}$, $p_x^s = 1 - q_x^s$ και $\overset{0}{p}_x = 1 - \overset{0}{q}_x$, οι οποίες δίνουν ομαλότερες και μικρότερες τιμές από τους λόγους $\frac{\overset{0}{q}_x}{q_x^s}$. Το p_x^s είναι το ποσοστό επιβίωσης στην ηλικία x που παίρνουμε από τον τυπικό πίνακα θνησιμότητας και $\overset{0}{p}_x$ είναι το αντίστοιχο αδρό ποσοστό. Ο Lidstone χρησιμοποίησε τη γραφική μέθοδο εξομάλυνσης, μπορούν όμως να εφαρμοστούν και μαθηματικές συναρτήσεις όπως οι παρακάτω:

$$q_x = aq_x^s + b \tag{3.3}$$

$$m_x = am_x^s + b \tag{3.4}$$

$$q_x = q_x^s(ax + b) \tag{3.5}$$

$$q_x = aq_x^{(1)} + bq_x^{(2)} \tag{3.6}$$

$$m_x = m^s_{x+n} + K \tag{3.7}$$

$$Y_x = a + bY^s_x \tag{3.8}$$

όπου τα a, b, K και n είναι παράμετροι, το $q^{(1)}_x$ αναφέρεται στα ποσοστά θνησιμότητας σε κάθε ηλικία ενός τυπικού πίνακα ενώ το $q^{(2)}_x$ στα αντίστοιχα ποσοστά ενός δεύτερου τυπικού πίνακα. Είναι φανερό ότι οι σχέσεις (3.4) και (3.7) αναφέρονται στην εξομάλυνση των τιμών της έντασης θνησιμότητας. Οι τιμές Y_x και Y^s_x είναι οι logit μετασχηματισμοί των παρατηρούμενων και τυπικών τιμών επιβίωσης S_x και S^s_x αντίστοιχα. Δηλαδή έχουμε

$$Y_x = \log it(S_x) = \frac{1}{2}\ln\frac{1-S_x}{S_x}$$

και

$$Y^s_x = \log it(S^s_x) = \frac{1}{2}\ln\frac{1-S^s_x}{S^s_x}.$$

Για περισσότερες πληροφορίες και λεπτομέρειες για την προσαρμογή των συναρτήσεων (3.3) – (3.8), ο αναγνώστης παραπέμπεται στην Lytrokapi (1998) και στους Benjamin and Pollard (1980).

Το *πλεονέκτημα* της μεθόδου αυτής είναι ότι μπορεί να χρησιμοποιηθεί, δίνοντας ικανοποιητικά αποτελέσματα, όταν τα δεδομένα είναι τόσο λίγα που ακόμη και η γραφική μέθοδος εξομάλυνσης δεν μπορεί να εφαρμοστεί. Επιπλέον, εφόσον ο τυπικός πίνακας είναι εξομαλυμένος, εξασφαλίζεται και η ομαλότητα των δεδομένων. Αντίθετα, το *μειονέκτημα* της μεθόδου έγκειται στη δυσκολία επιλογής του καταλληλότερου τυπικού πίνακα θνησιμότητας με αποτέλεσμα η εξομάλυνση να μην είναι ικανοποιητική.

Η εκτίμηση των παραμέτρων των συναρτήσεων (3.3) – (3.8), επιβάλλει την αντίστοιχη μείωση των βαθμών ελευθερίας της c^2 κατανομής του X^2 τεστ. Επιπλέον, η επιλογή τυπικού πίνακα θνησιμότητας, θέτει κάποιους επιπλέον περιορισμούς σχετικά με τις εξομαλυμένες τιμές και συνεπώς απαιτείται μια περαιτέρω ελάττωση των βαθμών ελευθερίας. Σημειώνουμε όμως ότι η ελάττωση αυτή είναι άγνωστη (Benjamin and Pollard, 1980).

3.3 Εξομάλυνση μέσω κινητών σταθμισμένων μέσων

Μια από τις πρώτες μεθόδους εξομάλυνσης είτε των ποσοστών θνησιμότητας είτε της έντασης θνησιμότητας είναι η *μέθοδος των κινητών σταθμισμένων μέσων* (London, 1985). Υπάρχουν δυο βασικές κατηγορίες αυτής της μεθόδου: οι *γραμμικώς σύνθετοι τύποι*, που αναπτύχθηκαν κυρίως από τον DeForest γύρω στο 1870 και οι *τύποι άθροισης (summation)*, που αναπτύχθηκαν στη Μεγάλη Βρετανία. Και οι δυο κατηγορίες αρχικά ονομάζονταν *προσαρμοσμένοι μέσοι*.

3.3.1 Γραμμικώς σύνθετοι τύποι

Η μέθοδος των γραμμικώς σύνθετων τύπων είναι απλή, τόσο στην κατανόηση όσο και στην εφαρμογή της και γι' αυτό θα ασχοληθούμε μόνο με αυτή. Για περισσότερες πληροφορίες για τους τύπους άθροισης, οι οποίοι δεν χρησιμοποιούνται ιδιαίτερα στις μέρες μας, παραπέμπουμε στους Benjamin and Pollard (1980).

Ο βασικός τύπος της μεθόδου των γραμμικώς σύνθετων τύπων, είναι ο

$$v_x = \sum_{r=-m}^{m} a_r u_{x+r} \ , \ x = 1, 2, ..., n \ , \ n > m \, , \tag{3.9}$$

δηλαδή οι εξομαλυμένες τιμές v_x υπολογίζονται ως σταθμισμένοι μέσοι όροι των μη εξομαλυμένων τιμών u_{x+r} είτε αυτές είναι παρατηρούμενες τιμές μιας συνεχούς σειράς είτε αρχικές εκτιμήσεις των ποσοστών θνησιμότητας ή της έντασης θνησιμότητας. Τα a_r ονομάζονται συντελεστές ή βάρη.

Σύμφωνα με τον τύπο (3.9) το πλήθος των όρων που χρησιμοποιούνται για τον υπολογισμό των μέσων όρων με βάρη τα a_r, $r = -m, ..., m$, είναι $2m+1$ και ο αριθμός αυτός ονομάζεται *εύρος* του τύπου. Επίσης αν και μπορούν να χρησιμοποιηθούν μη συμμετρικοί τύποι, στη βιβλιογραφία αναφέρονται κυρίως συμμετρικοί. Συμμετρικός είναι ένας τύπος αν ισχύει $a_r = a_{-r}$ για $r = 1, 2, ..., m$. Σημειώνεται ότι λόγω της κεντρικής φύσης (*central nature*) που παρουσιάζει ο τύπος (3.9), η μέθοδος αυτή δεν δίνει εξομαλυμένες τιμές για τις m πρώτες και m τελευταίες αρχικές εκτιμήσεις.

Γνωρίζουμε ότι $u_x = t_x + e_x$, $x = 1, 2, ..., n$ οπότε έχουμε

$$v_x = \sum_{r=-m}^{m} a_r (t_{x+r} + e_{x+r}) = \sum_{r=-m}^{m} a_r t_{x+r} + \sum_{r=-m}^{m} a_r e_{x+r} \ .$$

Εφόσον οι πραγματικές τιμές των t_x είναι άγνωστες, μπορούμε να υποθέσουμε μια συναρτησιακή μορφή για τη σειρά $\{t_x\}$. Επειδή η εξομάλυνση γίνεται στο διάστημα $[x-m, x+m]$, είναι λογικό να θεωρήσουμε ότι η σειρά $\{t_x\}$, αναπαριστάται από ένα τρίτου βαθμού πολυώνυμο.

Υπό την παραπάνω υπόθεση λοιπόν, ισχύει η αναπαραγωγική σχέση

$$\sum_{r=-m}^{m} a_r t_{x+r} = t_x$$

αν ισχύουν για τους συντελεστές a_r, οι σχέσεις

$$\sum_{r=-m}^{m} a_r = 1 \text{ και } \sum_{r=-m}^{m} r^2 a_r = 0 \tag{3.10}$$

οι οποίες ονομάζονται *εξισώσεις περιορισμών*. Εξαιτίας της συμμετρικότητας του τύπου (3.9), οι σχέσεις (3.10) μετατρέπονται σε

$$a_0 + 2\sum_{r=1}^{m} a_r = 1 \text{ και } \sum_{r=1}^{m} r^2 a_r = 0 .$$

Συνεπώς ισχύει $v_x = t_x + e'_x$, όπου $e'_x = \sum_{r=-m}^{m} a_r e_{x+r}$.

Υποθέτοντας ότι οι τυχαίες μεταβλητές που αντιστοιχούν στις τιμές u_{x+r}, δηλαδή οι U_{x+r}, $x = 1,2,...,n$, αποτελούν διωνυμικά ποσοστά με μέση τιμή t_{x+r} και ότι οι τυχαίες μεταβλητές που αντιστοιχούν στα σφάλματα, E_{x+r}, είναι ασυσχέτιστες με $E(E_{x+r}) = 0$ και έχουν κοινή διακύμανση s^2, αποδεικνύεται ότι

$$\text{var}(V_x) = s^2 \sum_{r=-m}^{m} a_r^2 ,$$

όπου V_x είναι οι τυχαίες μεταβλητές που αντιστοιχούν στις τιμές v_x, $x = 1,2,...,n$, αφού

$$V_x = \sum_{r=-m}^{m} a_r U_{x+r} \Rightarrow E(V_x) = \sum_{r=-m}^{m} a_r E(U_{x+r}) = \sum_{r=-m}^{m} a_r t_{x+r} = t_x .$$

Θέτοντας

$$R_0^2 = \sum_{r=-m}^{m} a_r^2 \tag{3.11}$$

μπορούμε να πούμε ότι οι συντελεστές a_r της σχέσης (3.9) είναι τέτοιοι ώστε να ελαχιστοποιούν τη σχέση (3.11) και να ικανοποιούν τις εξισώσεις περιορισμών.

Συνήθως στην πράξη η εξομάλυνση μέσω της ελαχιστοποίησης του R_0^2 δεν είναι ικανοποιητική ούτε ως προς την προσαρμογή ούτε ως προς την ομαλότητα. Για το λόγο αυτό ορίζουμε την ποσότητα

$$R_z^2 = \frac{\text{var}(\Delta^z V_x)}{\text{var}(\Delta^z U_x)} \tag{3.12}$$

όπου Δ^z είναι οι διαφορές τάξεως z και οι συντελεστές a_r της σχέσης (3.9) είναι τέτοιοι ώστε να ελαχιστοποιούν τη σχέση (3.12) και ταυτόχρονα να ικανοποιούν τις εξισώσεις περιορισμών. Η εμπειρία έχει δείξει ότι για τα κριτήρια της προσαρμογής και της ομαλότητας, η εξομάλυνση με βάση την ελαχιστοποίηση της ποσότητας R_z^2 δίνει ικανοποιητικά αποτελέσματα για $z =$ 2, 3 ή 4. Είναι φανερό ότι για $z = 0$ αναγόμαστε στην προηγούμενη περίπτωση.

Είναι επίσης εύκολο να δειχθεί, υπό τις υποθέσεις της μη συσχέτισης και των ίσων διασπορών των τυχαίων μεταβλητών E_{x+r}, ότι ισχύουν οι παρακάτω σχέσεις:

$$\text{var}(\Delta^z V_x) = s^2 \sum_{r=-m-z}^{m} (\Delta^z a_r)^2 \tag{3.13}$$

$$\text{var}(\Delta^z U_x) = s^2 \binom{2z}{z} \tag{3.14}$$

$$R_z^2 = \frac{1}{\binom{2z}{z}} \sum_{r=-m-z}^{m} (\Delta^z a_r)^2 , \tag{3.15}$$

αφού $\Delta^z V_x = \sum_{r=-m-z}^{m} \Delta^z a_r U_{x+z+r}$ και $\binom{2z}{z}$ είναι το άθροισμα των τετραγώνων των διωνυμικών συντελεστών σειράς z, δηλαδή $\sum_{k=0}^{z} \binom{z}{k}^2 = \binom{2z}{z}$.

Σημειώνουμε ότι για τον υπολογισμό των σχέσεων (3.13) και (3.14) θεωρούνται z τιμές των a_r ίσες με 0. Συγκεκριμένα ισχύει $a_r = 0$ για $|i - j| > n$ (Greville, 1983).

Μια εναλλακτική προσέγγιση για την υπό περιορισμούς ελαχιστοποίηση της ποσότητας R_z^2, η οποία δίνει έμφαση στην έννοια της ομαλότητας των εξομαλυμένων τιμών, ασχολείται όχι με τις μεταβλητές U_x και V_x, αλλά με τις τιμές τους.

Η ομαλότητα επιτυγχάνεται όταν η σχέση $\sum_{x=1}^{n-2m}(\Delta^z v_x)^2$ παίρνει μικρή τιμή. Για να γίνει η προηγούμενη ποσότητα μικρή θα πρέπει κάθε όρος του αθροίσματος $\Delta^z v_x$ να είναι μικρός. Επειδή

$$\Delta^z v_x = (-1)^z \sum_{x=-m-z}^{m} \Delta^z a_r u_{x+r+z}$$

και επειδή δεν μπορούμε να επέμβουμε στις τιμές u_{x+r+z}, θα πρέπει να επέμβουμε στις τιμές της σχέσης $\Delta^z a_r$. Για να μικρύνουμε αυτές τις ποσότητες, θα πρέπει να μικρύνουμε τη σχέση

$$\sum_{x=-m-z}^{m}(\Delta^z a_r)^2 .$$

Η παραπάνω σχέση μπορεί να θεωρηθεί σαν μέτρο της «*τραχύτητας*» (*roughness*) των τιμών v_x, καθώς μειώνοντας αυτή αυξάνεται η ομαλότητά τους. Το αντίστοιχο μέτρο για τις τιμές u_x ισούται με

$$\binom{2z}{z}$$

οπότε

$$R_z^2 = \frac{1}{\binom{2z}{z}} \sum_{r=-m-z}^{m}(\Delta^z a_r)^2 , \qquad (3.16)$$

το οποίο ισούται με τη σχέση (3.15).

Στη σχέση (3.15) το R_z^2 είναι ο λόγος της διακύμανσης του $\Delta^z V_x$ προς τη διακύμανση του $\Delta^z U_x$, ενώ στη σχέση (3.16) το R_z^2 είναι ο λόγος της τραχύτητας των v_x έναντι αυτής των u_x. Και στις δυο περιπτώσεις για να είναι ικανοποιητική η εξομάλυνση, δηλαδή τα v_x ομαλότερα των u_x, θα πρέπει να ισχύει $R_z^2 < 1$. Στην πράξη χρησιμοποιείται ως συντελεστής ομαλότητας η ποσότητα

$$R_3 = \sqrt{R_3^2} .$$

Όσο μικρότερη η τιμή του R_3 τόσο ομαλότερα είναι τα v_x σε σχέση με τα u_x.

Ας δούμε τώρα μια προσέγγιση για την ελαχιστοποίηση του R_z^2 (London, 1985).

Κατ' αρχήν εφόσον το R_z^2 είναι ένα άθροισμα τετραγώνων και συνεπώς θετική ποσότητα, το απόλυτο ελάχιστό της είναι το μηδέν. Αν θέσουμε $\frac{\partial R_z^2}{\partial a_r} = 0$ για όλα τα r και λύσουμε τις προκύπτουσες ταυτόχρονες εξισώσεις, θα πάρουμε $a_r = 0$ για όλα τα r. Αυτό συμβαίνει γιατί δεν έχουμε λάβει υπόψη τις εξισώσεις περιορισμών (3.10). Επειδή έχουμε δυο περιορισμούς, πρέπει να θεωρήσουμε δυο από τις μεταβλητές a_r να είναι εξαρτημένες ενώ όλες οι υπόλοιπες να είναι ανεξάρτητες. Δεν μας ενδιαφέρει ποιες δυο θα είναι εξαρτημένες, γι' αυτό το πιο βολικό είναι να θεωρήσουμε τις a_0 και a_1. Θεωρούμε δηλαδή ότι οι a_0 και a_1 είναι συναρτήσεις των υπόλοιπων a_r, $r = 2, 3, ..., m$. Χρησιμοποιώντας την ιδιότητα συμμετρίας $a_r = a_{-r}$ για $r = 1, 2, ..., m$ και τις εξισώσεις περιορισμών ως

$$a_0 + 2\sum_{r=1}^{m} a_r = 1 \text{ και } \sum_{r=1}^{m} r^2 a_r = 0$$

έχουμε ότι

$$a_0 + 2\sum_{r=1}^{m} a_r = 1 \Rightarrow \; a_0 + 2a_1 + 2\sum_{r=2}^{m} a_r = 1$$

οπότε

$$a_1 = -\sum_{r=2}^{m} r^2 a_r \text{ και } a_0 = 1 + 2\sum_{r=2}^{m} (r^2 - 1) a_r \,. \tag{3.17}$$

Συμβολίζοντας με R το $\binom{2z}{z} R_z^2$ έχουμε ότι

$$R = \sum_{r=-m-2}^{m} \left(\Delta^z a_r\right)^2 ,$$

το οποίο είναι συνάρτηση των ανεξάρτητων a_r, $r = 2, 3, ..., m$ και το Δ^z συμβολίζει τις διαφορές τάξης z.

Κάθε a_r για $r = -m, ..., m$ εμπλέκεται στο R μόνο στους όρους $\left(\Delta^z a_{r-z}\right)^2 + ... + \left(\Delta^z a_r\right)^2$. Αν όλοι οι συντελεστές a_r ήταν μαθηματικά ανεξάρτητοι, τότε θα ίσχυε

$$\frac{\partial R}{\partial a_r} = 2(-1)^z d^{2z} a_r$$

για $r = -m, ..., m$, όπου με d^z συμβολίζουμε τις κεντρικές d^z διαφορές τάξης z. Για παράδεγμα, για $z = 2$ και 4 έχουμε

$$d^2 a_r = a_{r+1} - 2a_r + a_{r-1} \text{ και } d^4 a_r = a_{r+2} - 4a_{r+1} + 6a_r - 4a_{r-1} + a_{r-2} \,.$$

Επειδή όμως ισχύει $a_r = a_{-r}$, έχουμε ότι

$$\frac{\partial R}{\partial a_r} = 4(-1)^z d^{2z} a_r,\ r = 1, 2, ..., m$$

και σαν ειδική περίπτωση έχουμε

$$\frac{\partial R}{\partial a_0} = 2(-1)^z d^{2z} a_0 .$$

Τελικά επειδή τα a_0 και a_1 είναι συναρτήσεις των υπόλοιπων a_r, παίρνουμε τη συνολική παράγωγο

$$\frac{\partial R}{\partial a_r} = 4(-1)^z d^{2z} a_r + \frac{\partial R}{\partial a_0} \cdot \frac{\partial a_0}{\partial a_r} + \frac{\partial R}{\partial a_1} \cdot \frac{\partial a_1}{\partial a_r},\ r = 2, 3, ..., m .$$

Από τη (3.17) βρίσκουμε ότι

$$\frac{\partial a_0}{\partial a_r} = 2(r^2 - 1) \text{ και } \frac{\partial a_1}{\partial a_r} = -r^2$$

οπότε τελικά παίρνουμε ότι

$$\frac{\partial R}{\partial a_r} = 4(-1)^z \left(d^{2z} a_r + (r^2 - 1) d^{2z} a_0 - r^2 d^{2z} a_1 \right),\ r = 2, 3, ..., m .$$

Εξισώνοντας την παραπάνω σχέση με 0, παίρνουμε

$$d^{2z} a_r = r^2 d^{2z} a_1 - (r^2 - 1) d^{2z} a_0 ,$$

το οποίο είναι μια τετραγωνική συνάρτηση του r έτσι ώστε το a_0 να είναι ένα πολυώνυμο του r με βαθμό $2z + 2$. Λόγω συμμετρίας των a_r το πολυώνυμο πρέπει να περιέχει μόνο ζυγούς όρους δυνάμεων, δηλαδή ισχύει

$$a_r = b_0 + b_2 r^2 + b_4 r^4 + ... + b_{2z+2} r^{2z+2}$$

όπου υπάρχουν $z + 2$ άγνωστες παράμετροι. Για να τις εκτιμήσουμε, χρησιμοποιούμε τις z συνθήκες

$$a_{m+1} = ... = a_{m+z} = 0$$

και τις συνθήκες

$$a_0 + 2\sum_{r=1}^{m} a_r = 1 \text{ και } \sum_{r=1}^{m} r^2 a_r = 0 .$$

Έτσι έχουμε μια σχέση που παράγει τους συντελεστές a_r που χρησιμοποιούνται στον τύπο (3.9).

Όπως έχουμε ήδη αναφέρει το πλεονέκτημα της μεθόδου αυτής, είναι η ευκολία στην κατανόηση και την εφαρμογή της. Το βασικό μειονέκτημά της, είναι ότι ένας τύπος $2m+1$ όρων δεν δίνει εξομαλυμένες τιμές για τις m πρώτες και m τελευταίες παρατηρήσεις εκτός και αν δοθούν επιπλέον παρατηρήσεις.

Οι Benjamin and Pollard (1980) αναφέρουν ότι απαιτείται μια μείωση στους βαθμούς ελευθερίας της c^2 κατανομής, της τάξεως $(n-2m)(2a_0 - f_E^2)$, όπου $f_E^2 = \left(a_0^2 + 2\sum_{r=1}^{m} a_r^2 \right)$ το οποίο ονομάζεται *δείκτης μείωσης σφάλματος* (*error – reducing index*) και n είναι ο αριθμός των ηλικιών.

3.4 Whittaker – Henderson εξομάλυνση

Μια μη παραμετρική μέθοδος εξομάλυνσης πινάκων θνησιμότητας, είναι και η μέθοδος που πρώτος επινόησε ο Whittaker το 1923 και επέκτεινε ο Henderson το 1924 και 1925. Προς τιμήν των δυο αυτών ερευνητών, η μέθοδος ονομάζεται *Whittaker – Henderson μέθοδος* (London, 1985).

Επειδή η εξομάλυνση γενικώς βασίζεται στις έννοιες της *προσαρμογής* και της *ομαλότητας*, σε αυτές στηρίζεται και η μέθοδος των Whittaker και Henderson, η οποία μπορεί να εφαρμοσθεί είτε στα ποσοστά είτε στην ένταση θνησιμότητας. Συγκεκριμένα, στηρίζεται στην ελαχιστοποίηση της συνάρτησης

$$M = F + hS = \sum_{x=1}^{n} w_x (u_x - v_x)^2 + h \sum_{x=1}^{n-z} (\Delta^z v_x)^2 , \qquad (3.18)$$

όπου v_x είναι οι εξομαλυμένες τιμές, u_x είναι οι αρχικές εκτιμήσεις, w_x είναι βάρη για $x = 1, 2, ..., n$, h είναι θετική παράμετρος, η οποία ελέγχει τη σχετική έμφαση που δίνεται στην προσαρμογή F και την ομαλότητα S, κατά την ελαχιστοποίηση της M και $\Delta^z v_x$ είναι οι διαφορές των εξομαλυμένων τιμών. Το z είναι η τάξη των διαφορών και συνήθως παίρνει τις τιμές 2 ή 3 (Greville, 1983 και Haberman, 1998).

Τα βάρη w_x είναι θετικοί πραγματικοί αριθμοί και μπορεί να ισούνται με τον αντίστροφο της $\text{var}(U_x)$, όπου U_x είναι η τυχαία μεταβλητή που αντιστοιχεί στην τιμή u_x. Αυτά συνήθως χρησιμοποιούνται για τον υπολογισμό ενός μέτρου για την προσαρμογή, μετά την εξομάλυνση των διαφόρων ασφαλιστικών ποσοτήτων.

Πριν την εξομάλυνση των ποσοστών θνησιμότητας, ως βάρη μπορούν να χρησιμοποιηθούν οι ποσότητες

$$w_x = \frac{l_x}{v_x (1 - v_x)},$$

πράγμα που δυσκολεύει τους υπολογισμούς. Χρησιμοποιώντας, στην παραπάνω σχέση, u_x αντί για v_x εξασφαλίζουμε την ανεξαρτησία των w_x από τα v_x. Συχνά ως βάρη χρησιμοποιούνται τα μεγέθη των δειγμάτων l_x, δηλαδή ο αριθμός των ατόμων σε κίνδυνο στην ηλικία x. Μια τέτοια επιλογή αυξάνει πολύ το μέγεθος της ποσότητας F, η οποία πρέπει να αντισταθμιστεί από την επιλογή μιας μεγάλης τιμής για την παράμετρο h. Εναλλακτικός τρόπος είναι η επιλογή

$$w_x = \frac{l_x}{\bar{l}},$$

όπου $\bar{l}$ είναι ο αριθμητικός μέσος των l_x για $x = 1, 2, ..., n$. Είναι σαφές ότι με την επιλογή αυτή δίνεται μεγαλύτερη έμφαση στην προσαρμογή στις ηλικίες με τα μεγαλύτερα μεγέθη δείγματος l_x. Σε περίπτωση που ισχύει $w_x = 1$ για όλες τις ηλικίες x, η μέθοδος ονομάζεται «*Τύπου Α*» ενώ σε κάθε άλλη περίπτωση, ονομάζεται «*Τύπου Β*».

Όσον αφορά την παράμετρο h, έχουμε τα παρακάτω: Όταν $h = 0$ ισχύει $M = F$ και συνεπώς, επειδή $F \geq 0$, το M ελαχιστοποιείται για $F = 0$. Αυτό σημαίνει ότι $v_x = u_x$ για $x = 1, 2, ..., n$, δηλαδή στην ουσία δεν έχει γίνει εξομάλυνση. Γενικά καθώς το h τείνει στο 0, οι εξομαλυμένες τιμές v_x τείνουν στις αρχικές εκτιμήσεις των τιμών u_x και περισσότερη έμφαση δίνεται στην προσαρμογή έναντι της ομαλότητας. Το αντίθετο ισχύει όταν το h παίρνει μεγάλες τιμές.

3.4.1 Ελαχιστοποίηση της συνάρτησης M

Έστω $\mathbf{u} = (u_1, ..., u_n)^T$, $\mathbf{v} = (v_1, ..., v_n)^T$, $\mathbf{W} = diag(w_1, ..., w_n)$ και $\mathbf{K_z}$ ένας $(n - z) \times n$ πίνακας με στοιχεία τους συντελεστές των διαφορών τάξης z. Έτσι το γινόμενο $\mathbf{K_z v}$ είναι το $(n - z) \times 1$ διάνυσμα με στοιχεία τις τιμές των διαφορών $\Delta^z v_x$. Στην περίπτωση αυτή, σε πινακική μορφή, η σχέση (3.18) γράφεται ως

$$M = (\mathbf{v} - \mathbf{u})^T \mathbf{W} (\mathbf{v} - \mathbf{u}) + h \mathbf{v}^T \mathbf{K_z}^T \mathbf{K_z} \mathbf{v},$$

όπου το T δηλώνει την αντιστροφή του διανύσματος ή του πίνακα.

Για να βρούμε τις τιμές του διανύσματος $\mathbf{v}$ που ελαχιστοποιεί τον πίνακα M, θα πρέπει να λύσουμε την εξίσωση

$$(\mathbf{W} + h \mathbf{K_z}^T \mathbf{K_z}) \mathbf{v} = \mathbf{W} \mathbf{u}. \tag{3.19}$$

Θέτοντας $\mathbf{C} = \mathbf{W} + h \mathbf{K_z}^T \mathbf{K_z}$, η σχέση (3.19) γίνεται

$$\mathbf{C} \mathbf{v} = \mathbf{W} \mathbf{u}. \tag{3.20}$$

Σημειώνουμε ότι η λύση της εξίσωσης (3.20) είναι μοναδική αφού ο πίνακας $\mathbf{C}$ είναι μη ιδιάζων (*non-singular*), δηλαδή η ορίζουσά του είναι διάφορη του μηδενός. Συγκεκριμένα ο πίνακας $h \mathbf{K_z}^T \mathbf{K_z}$ είναι ιδιάζων (*singular*) και μετατρέπεται σε μη ιδιάζοντα, με την πρόσθεση του διαγώνιου πίνακα $\mathbf{W}$ (London, 1985). Επίσης σημειώνουμε ότι η επίλυση της σχέσης (3.19) ή (3.20) είναι ισοδύναμη με την επίλυση των εξισώσεων

$$\frac{\partial M}{\partial v_r} = 0\,,\ r = 1, 2, ..., n\,.$$

Για την απόδειξη της σχέσης (3.20) παραπέμπουμε στον London (1985).

3.4.2 Παραλλαγές της βασικής μεθόδου των Whittaker – Henderson

Η πρώτη παραλλαγή αποτελεί μια γενίκευση του M της σχέσης (3.18). Συγκεκριμένα η συνάρτηση που πρέπει να ελαχιστοποιηθεί δίνεται από τον τύπο

$$M = \sum_{x=1}^{n} w_x (u_x - v_x)^2 + h_1 \sum_{x=1}^{n-1} (\Delta v_x)^2 + h_2 \sum_{x=1}^{n-2} (\Delta^2 v_x)^2 + ... + h_z \sum_{x=1}^{n-z} (\Delta^z v_x)^2\,.$$

Αν $h_z > 0$ και $h_j = 0$ για $j = 1, 2, ..., z-1$ παίρνουμε τη σχέση (3.18). Συνήθως χρησιμοποιείται η περίπτωση όπου $h_2 > 0$, $h_3 > 0$ και $h_j = 0$ για όλα τα υπόλοιπα j. Η παραλλαγή αυτή ονομάζεται *γενική «μικτών διαφορών» μορφή*.

Η δεύτερη παραλλαγή είναι η *εκθετική μορφή για το* S. Υποθέτουμε ότι η ένταση θνησιμότητας μοντελοποιείται σαν ένα πολυώνυμο βαθμού $z-2$ συν έναν εκθετικό όρο, όπως για παράδειγμα είναι ο πρώτος και ο δεύτερος νόμος του Makeham. Στην περίπτωση αυτή έχουμε $S = \sum_{x=1}^{n-z} (\Delta^z v_x - r\Delta^{z-1} v_x)^2$, όπου $r = C - 1$ και C είναι η παράμετρος που εμπλέκεται στους νόμους του Makeham.

Η *παραλλαγή του Lowry* συνδυάζει την προηγούμενη παραλλαγή με την αναφορά σε έναν τυπικό πίνακα θνησιμότητας. Ο τύπος του M δίνεται από τη σχέση

$$M = (1-h_1) \sum_{x=1}^{n} w_x (u_x - v_x)^2 + h_1 \sum_{x=1}^{n} w_x^s (v_x - s_x)^2 + h_2 \sum_{x=1}^{n-z} (\Delta^z v_x - r\Delta^{z-1} v_x)^2\,.$$

Οι τιμές s_x είναι οι πιθανότητες θνησιμότητας ή οι τιμές της έντασης θνησιμότητας του τυπικού πίνακα, w_x^s είναι τα βάρη που σχετίζονται με την προσαρμογή των v_x στα s_x και r όπως πριν. Η παράμετρος h_1 για την οποία ισχύει $0 \le h_1 \le 1$ ελέγχει την έμφαση που δίνεται στην προσαρμογή στον τυπικό πίνακα έναντι της προσαρμογής στις αρχικές εκτιμήσεις των τιμών u_x.

Η τελευταία παραλλαγή που αναφέρει ο London (1985) είναι αυτή που προτάθηκε από τον *Schuette*. Ο Schuette πρότεινε την ελαχιστοποίηση της συνάρτησης

$$M = \sum_{x=1}^{n} w_x \left| u_x - v_x \right| + h \sum_{x=1}^{n-z} \left| \Delta^z v_x \right|\,.$$

Ο λόγος που χρησιμοποιούνται απόλυτες τιμές είναι επειδή αν η κατανομή των σφαλμάτων E_x δεν είναι η κανονική, όπως υποθέτουν οι Whittaker – Henderson, θα παράγονται περισσότερες απομακρυσμένες τιμές (*outliers*) και συνεπώς το κριτήριο με τις τετραγωνικές δυνάμεις θα είναι πολύ ευαίσθητο σε αυτές.

Μια γενίκευση της μεθόδου των Whittaker – Henderson στις δυο διαστάσεις δίνεται από τον London (1985). Οι δυο μεταβλητές που θεωρούνται είναι η ηλικία επιλογής και ο χρόνος μέχρι την επιλογή. Ο ενδιαφερόμενος μπορεί να βρει μια περαιτέρω γενίκευση της μεθόδου σε περισσότερες διαστάσεις στον Broffit (1996).

Οι Guerrero et al. (2001) αποδεικνύουν ότι με βάση ορισμένες υποθέσεις, ο καλύτερος γραμμικός αμερόληπτος εκτιμητής (*BLUE*) για τα πραγματικά ποσοστά θνησιμότητας δίνεται από τον τύπο των Whittaker – Henderson.

Ο London (1985) αναφέρει επίσης πως το μπεϋζιανό υπόβαθρο της μεθόδου αυτής, οδήγησε στην ανάπτυξη της μπεϋζιανής μεθόδου εξομάλυνσης. Και τα δυο αυτά θέματα θα περιγράψουμε στη συνέχεια.

3.5 Εξομάλυνση μέσω μπεϋζιανής θεωρίας

Μια εναλλακτική μέθοδος εξομάλυνσης πινάκων θνησιμότητας με μέλλον, σύμφωνα με τον London (1985), είναι η *μπεϋζιανή μέθοδος*.

Η μέθοδος αυτή μπορεί να θεωρηθεί ότι αποτελείται από τέσσερα βασικά βήματα:

1. Κατασκευή της εκ των προτέρων (*priori*) κατανομής πιθανότητας των πραγματικών τιμών t_x, $x = 1, 2, ..., n$ που θέλουμε να εκτιμήσουμε.
2. Επιλογή του μοντέλου για την υπό συνθήκη κατανομή των παρατηρούμενων τιμών, είτε ποσοστών είτε έντασης θνησιμότητας, u_x, δοθεισών των πραγματικών τιμών t_x.
3. Χρησιμοποίηση του θεωρήματος του Bayes για την επίλυση ως προς την εκ των υστέρων (*posteriori*) κατανομή των πραγματικών τιμών t_x, δοθεισών των αρχικών εκτιμήσεων u_x.
4. Επιλογή των εξομαλυμένων τιμών v_x από την εκ των υστέρων κατανομή.

Η μπεϋζιανή προσέγγιση του προβλήματος της εξομάλυνσης είναι η μόνη, στην οποία οι πραγματικές τιμές t_x θεωρούνται ως τυχαίες μεταβλητές, έστω T_x. Στη συνέχεια θα θεωρήσουμε για ευκολία πινακικές μορφές. Το πρώτο βήμα είναι η κατασκευή της πολυμεταβλητής εκ των προτέρων συνάρτησης πυκνότητας $f_{\mathbf{T}}(\mathbf{t})$. Έπειτα πρέπει να βρούμε τη δεσμευμένη πυκνότητα του $\mathbf{u}$, δοθέντος του διανύσματος $\mathbf{t}$, δηλαδή την $f_{\mathbf{U}|\mathbf{T}}(\mathbf{u}|\mathbf{t})$. Χρησιμοποιώντας το θεώρημα του Bayes θα βρούμε την πολυμεταβλητή συνάρτηση πυκνότητας

$$f_{\mathbf{T}|\mathbf{U}}(\mathbf{t}|\mathbf{u}) = \frac{f_{\mathbf{U}|\mathbf{T}}(\mathbf{u}|\mathbf{t}) f_{\mathbf{T}}(\mathbf{t})}{f_{\mathbf{U}}(\mathbf{u})}.$$

Για ευκολία στις πράξεις, καλό είναι να επιλέγουμε την $f_{\mathbf{T}}(\mathbf{t})$ ως συζυγή (*conjugate*) κατανομή καθώς έτσι η $f_{\mathbf{T}|\mathbf{U}}(\mathbf{t}|\mathbf{u})$ θα είναι επίσης συζυγής και συνεπώς θα έχει τις ίδιες ιδιότητες με την εκ των προτέρων πυκνότητα $f_{\mathbf{T}}(\mathbf{t})$. Τέλος το διάνυσμα των εξομαλυμένων τιμών, $\mathbf{v}$, επιλέγεται από την πληροφορία που περιέχει η $f_{\mathbf{T}|\mathbf{U}}(\mathbf{t}|\mathbf{u})$. Επειδή στην εξομάλυνση των πινάκων θνησιμότητας θεωρούμε ως εξομαλυμένες τιμές τις πιο «όμοιες» τιμές του διανύσματος $\mathbf{T}$, επιλέγουμε ως $\mathbf{v}$, την κορυφή (*mode*) της συνάρτησης $f_{\mathbf{T}|\mathbf{U}}(\mathbf{t}|\mathbf{u})$.

Στη συνέχεια θα δούμε συνοπτικά δυο μεθόδους εξομάλυνσης με μπεϋζιανό υπόβαθρο.

3.5.1 Μέθοδος των Whittaker – Henderson

Η πρώτη μέθοδος στην οποία χρησιμοποιήθηκε η μπεϋζιανή θεωρία, ήταν η μέθοδος των Whittaker – Henderson. Συγκεκριμένα οι Whittaker και Henderson πρότειναν ως εκ των προτέρων πυκνότητα πιθανότητα για τα t_x την

$$f_T(t_x) = c_1 e^{-lS}, \tag{3.21}$$

όπου c_1 είναι σταθερά τέτοια ώστε $\int f_T(t_x)dt_x = 1$, l είναι σταθερά και $S = \sum_{x=1}^{n-z} (\Delta^z t_x)^2$ είναι το μέτρο ομαλότητας. Θεωρώντας ανεξαρτησία των u_x για $x = 1, 2, ..., n$ και την κανονική κατανομή με μέσο 0 και διακύμανση s_x^2 για την τυχαία μεταβλητή των σφαλμάτων $E_x = U_x - T_x$, πήραν

$$f_{U|T}(u_x \mid t_x) = c_2 \exp\left\{-\frac{1}{2}\sum_{x=1}^{n} w_x (t_x - u_x)^2\right\}, \tag{3.22}$$

όπου c_2 είναι σταθερά και $w_x = [\text{var}(E_x)]^{-1}$. Λόγω του θεωρήματος του Bayes και των σχέσεων (3.21) και (3.22) παίρνουμε

$$f_{T|U}(t_x \mid u_x) = \frac{c_3 \exp\left\{-lS - \frac{1}{2}\sum_{x=1}^{n} w_x (t_x - u_x)^2\right\}}{f_U(u_x)},$$

όπου $c_3 = c_1 c_2$ είναι σταθερά.

Οι Whittaker και Henderson όρισαν ως την πιο «όμοια» σειρά $\{t_x\}$ ή $\{v_x\}$, αυτή που μεγιστοποιεί την παραπάνω σχέση. Αυτό είναι ισοδύναμο με την ελαχιστοποίηση της ποσότητας

$$\frac{1}{2}\sum_{x=1}^{n} w_x (t_x - u_x)^2 + lS\,. \tag{3.23}$$

Επειδή η ελαχιστοποίηση της (3.23) είναι δύσκολη χρησιμοποιώντας σαν βάρη τα $w_x = [\text{var}(E_x)]^{-1}$, πρότειναν τη χρησιμοποίηση σταθερών βαρών για όλα τα x, η τιμή των οποίων ισούται με το μισό του αντιστρόφου (*reciprocal*) της κοινής διασποράς. Συνεπώς η (3.23) μπορεί να γραφεί ως

$$wF + lS\,, \tag{3.24}$$

όπου $F = \sum_{x=1}^{n} (u_x - t_x)^2$. Τέλος είναι φανερό ότι η σχέση (3.24) γράφεται ως $F + \frac{l}{w} S = F + hS$, που είναι ίδια με τη σχέση (3.18).

3.5.2 Μέθοδος των Kimeldorf – Jones

Η μέθοδος που περισσότερο εφαρμόζει τη μπεϋζιανή θεωρία στο πρόβλημα της εξομάλυνσης είναι η μέθοδος των Kimeldorf – Jones. Στην περίπτωση αυτή, η εκ των προτέρων πυκνότητα του $\mathbf{T}$ είναι η πολυδιάστατη κανονική με διάνυσμα μέσων $\mathbf{m}$, πίνακα διακυμάνσεων – συνδιακυμάνσεων $\mathbf{A}$ και τύπο

$$f_{\mathbf{T}}(\mathbf{t}) = k_1 \exp\left\{-\frac{1}{2}(\mathbf{t}-\mathbf{m})^T \mathbf{A}^{-1}(\mathbf{t}-\mathbf{m})\right\}. \qquad (3.25)$$

Ο $n \times n$ πίνακας $\mathbf{A}$ είναι θετικά ορισμένος και μη ιδιάζων, $k_1 = \left[(2p)^n |\mathbf{A}|\right]^{-\frac{1}{2}}$ και το T δηλώνει την αντιστροφή του διανύσματος. Αν και μια ικανοποιητική υπό συνθήκη κατανομή για τις τυχαίες μεταβλητές U_x δοθεισών των τιμών t_x είναι η διωνυμική (ή πιο καλά το διωνυμικό ποσοστό), οι Kimeldorf και Jones πήραν ως δεσμευμένη πυκνότητα του $\mathbf{U}$, δοθέντος του διανύσματος $\mathbf{t}$, την πολυδιάστατη κανονική με τύπο

$$f_{\mathbf{U}|\mathbf{T}}(\mathbf{u} \mid \mathbf{t}) = k_2 \exp\left\{-\frac{1}{2}(\mathbf{u}-\mathbf{t})^T \mathbf{B}^{-1}(\mathbf{u}-\mathbf{t})\right\}, \qquad (3.26)$$

όπου, θεωρώντας αμοιβαία ανεξαρτησία για τις μεταβλητές U_x, $\mathbf{B}$ είναι ένας $n \times n$ διαγώνιος και θετικά ορισμένος πίνακας με στοιχεία τις διακυμάνσεις των U_x και $k_2 = \left[(2p)^n |\mathbf{B}|\right]^{-\frac{1}{2}}$.

Με βάση το θεώρημα του Bayes και τις σχέσεις (3.25) και (3.26), παίρνουμε

$$f_{\mathbf{T}|\mathbf{U}}(\mathbf{t} \mid \mathbf{u}) = k_4 \exp\left\{-\frac{1}{2}\left[(\mathbf{t}-\mathbf{m})^T \mathbf{A}^{-1}(\mathbf{t}-\mathbf{m}) + (\mathbf{u}-\mathbf{t})^T \mathbf{B}^{-1}(\mathbf{u}-\mathbf{t})\right]\right\}, \qquad (3.27)$$

όπου $k_4 = \frac{k_1 k_2}{k_3}$ και $k_3 = f_{\mathbf{U}}(\mathbf{u})$ είναι σταθερά σε σχέση με το $\mathbf{t}$. Μπορεί να δειχθεί ότι η σχέση (3.27) γράφεται στην απλή μορφή

$$f_{\mathbf{T}|\mathbf{U}}(\mathbf{t} \mid \mathbf{u}) = k_5 \exp\left\{-\frac{1}{2}(\mathbf{t}-\mathbf{v})^T \mathbf{C}^{-1}(\mathbf{t}-\mathbf{v})\right\},$$

όπου η σταθερά k_5 δεν εξαρτάται από το $\mathbf{t}$, $\mathbf{v} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}(\mathbf{B}^{-1}\mathbf{u} + \mathbf{A}^{-1}\mathbf{m})$ και $\mathbf{C} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}$ που σημαίνει ότι η $f_{\mathbf{T}|\mathbf{U}}(\mathbf{t}\,|\,\mathbf{u})$ είναι επίσης πολυδιάστατη κανονική με διάνυσμα μέσων $\mathbf{v}$ και πίνακα διακυμάνσεων – συνδιακυμάνσεων $\mathbf{C}$. Επειδή η εκ των υστέρων κατανομή είναι πολυδιάστατη κανονική, σαν διάνυσμα των εξομαλυμένων τιμών $\mathbf{v}$ μπορούμε να πάρουμε είτε το μέσο είτε την κορυφή είτε τη διάμεσο της $f_{\mathbf{T}|\mathbf{U}}(\mathbf{t}\,|\,\mathbf{u})$ τα οποία ταυτίζονται. Άρα παίρνουμε

$$\mathbf{v} = (\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}(\mathbf{B}^{-1}\mathbf{u} + \mathbf{A}^{-1}\mathbf{m}).$$

Είναι φανερό ότι το διάνυσμα $\mathbf{v}$ είναι ένας σταθμισμένος μέσος των $\mathbf{m}$ και $\mathbf{u}$ με βάρη τον αντίστροφο (*reciprocal*) του αντίστοιχου πίνακα διακυμάνσεων – συνδιακυμάνσεων.

Όσον αφορά τις παραμέτρους $\mathbf{m}$, $\mathbf{A}$ και $\mathbf{B}$ αξίζει να κάνουμε τις παρακάτω παρατηρήσεις. Καθώς η κατανομή του $\mathbf{T}$ είναι πολυδιάστατη κανονική, το $\mathbf{m}$ μπορεί να είναι ο μέσος, η κορυφή ή η διάμεσος της $f_{\mathbf{T}}(\mathbf{t})$. Για τον πίνακα $\mathbf{A}$ αξίζει να πούμε ότι παρά τις κάποιες οδηγίες που υπάρχουν (London, 1985), απαιτείται και η προσωπική κρίση ή προτίμηση του ερευνητή. Για τα διαγώνια στοιχεία του $\mathbf{A}$ δεν υπάρχει αμφισβήτηση ότι είναι οι διασπορές των τυχαίων μεταβλητών του διανύσματος $\mathbf{T}$. Για τα μη διαγώνια στοιχεία του πίνακα υπάρχουν διάφορες επιλογές. Η πρώτη είναι

$$a_{ij} = c_{ij}\sqrt{a_{ii}a_{jj}}\,,$$

όπου c_{ij} είναι ο συντελεστής συσχέτισης μεταξύ των T_i και T_j. Βέβαια είναι δύσκολο να επιλέξουμε τα c_{ij}. Οι Kimeldorf and Jones (1967) πρότειναν την επιλογή των a_{ij} ως

$$a_{ij} = s^2 r^{|i-j|},\ s > 0 \text{ και } 0 \le r < 1,$$

όπου s^2 είναι η διασπορά της τυχαίας μεταβλητής T_i για όλα τα i και r ο συντελεστής συσχέτισης των T_i και T_j για $|i-j| = 1$. Σε περίπτωση που δεν μας ικανοποιεί η υπόθεση της σταθερότητας του συντελεστή συσχέτισης, επιλέγουμε

$$a_{ij} = s^2 r_i r_{i+1} \ldots r_{j-1},$$

όπου αν $c_{12} = r_1$, $c_{23} = r_2$, $c_{34} = r_3$ τότε $c_{13} = r_1 r_2$, $c_{24} = r_2 r_3$, $c_{14} = r_1 r_2 r_3$ κτλ.

Τέλος σε περίπτωση που δεν μας ικανοποιεί η υπόθεση των ίσων διασπορών για όλες τις μεταβλητές T_i και υποθέτουμε σταθερό συντελεστή συσχέτισης, έχουμε

$$a_{ij} = s_i s_j r^{|i-j|}.$$

Για την επιλογή των στοιχείων του πίνακα **B**, τα πράγματα είναι ευκολότερα, αφού ο **B** είναι διαγώνιος για να δηλώνει την ανεξαρτησία των τυχαίων μεταβλητών U_x και έτσι στην κύρια διαγώνιο βρίσκονται οι διασπορές των μεταβλητών U_x. Αν η U_x προσεγγίζει το διωνυμικό ποσοστό, τότε έχουμε $\text{var}(U_i) = \frac{t_i(1-t_i)}{l_i}$. Στην περίπτωση της μεθόδου των Kimeldorf – Jones, ο πίνακας **B** πρέπει να είναι ανεξάρτητος του **t** και γι' αυτό παίρνουμε $b_{ii} = \frac{m_i(1-m_i)}{l_i}$ καθώς το **m** είναι ο εκ των προτέρων «καλύτερος εκτιμητής» του **T**.

Όπως και στις προηγούμενες μεθόδους εξομάλυνσης, για κάθε παράμετρο που εκτιμούμε, αφαιρούμε και ένα βαθμό από τους βαθμούς ελευθερίας της c^2 κατανομής.

Κλείνοντας το θέμα της εξομάλυνσης των ασφαλιστικών ποσοτήτων με χρήση της μπεϋζιανής θεωρίας, να τονίσουμε ότι η μέθοδος είναι αντιφατική. Η αντίφαση έγκειται στο γεγονός ότι ενώ χρησιμοποιούνται στοιχεία μιας θεωρίας, υπεισέρχεται σε μεγάλο ποσοστό η υποκειμενικότητα του ερευνητή.

3.6 Εξομάλυνση μέσω εκτιμητών πυρήνα

Μια άλλη μη παραμετρική προσέγγιση στο πρόβλημα της εξομάλυνσης είναι οι *μέθοδοι των εκτιμητών πυρήνα* (Copas and Haberman, 1983).

Οι μέθοδοι αυτές χρησιμοποιούνται γενικά για την εκτίμηση της συνάρτησης πυκνότητας πιθανότητας $f(x)$ με βάση ένα δείγμα παρατηρήσεων $x_1,\ x_2,\ldots,\ x_n$. Η εκτίμηση της πυκνότητας στο σημείο x δίνεται από τη σχέση

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} y\left(\frac{x - x_i}{h}\right), \tag{3.28}$$

όπου $y(s)$ είναι συνεχής, θετική, συμμετρική και μονοκόρυφη συνάρτηση τέτοια ώστε $\int_{-\infty}^{+\infty} y(s)ds = 1$ και h είναι θετική συνάρτηση του n η οποία τείνει στο μηδέν όταν το n τείνει στο άπειρο. Η σχέση (3.28) μπορεί να θεωρηθεί ως σταθμισμένος μέσος όρος της δειγματικής συνάρτησης κατανομής ο οποίος για να πάρει όσο το δυνατόν περισσότερη πληροφορία για την πυκνότητα f στο σημείο x, χρησιμοποιεί τις παρατηρήσεις που βρίσκονται πιο κοντά στο σημείο x. Χρησιμοποιεί δηλαδή τις παρατηρήσεις που βρίσκονται σε μια περιοχή του x, το εύρος της οποίας ονομάζεται *παράμετρος παραθύρου* (*bandwidth parameter*). Η παράμετρος αυτή, συμβολίζεται με h και ελέγχει το βαθμό ομαλότητας του εκτιμητή $\hat{f}$.

Οι πιο ευρέως χρησιμοποιούμενες συνεχείς συναρτήσεις πυρήνα είναι οι εξής:
ο τριγωνικός (*triangular*) πυρήνας

$$y_T(s) = \begin{cases} 1 - |s|, & |s| < 1 \\ \\ 0, & |s| > 1, \end{cases}$$

ο κανονικός (*Gaussian*) πυρήνας

$$y_G(s) = \frac{1}{\sqrt{2p}} e^{-s^2/2},$$

ο τετραγωνικός (*rectangular*) πυρήνας

$$y_R(s) = \begin{cases} \frac{1}{2}, & |s| < 1 \\ 0, & |s| > 1 \end{cases}$$

και ο πυρήνας Epanechnikov

$$y_E(s) = \begin{cases} 0.75\left(1 - s^2\right), & |s| < 1 \\ 0, & |s| > 1 . \end{cases}$$

Για την εξομάλυνση των ποσοστών θνησιμότητας q_x, έχουν προταθεί δυο εκτιμητές πυρήνα. Ο πρώτος προτάθηκε από τους Copas and Haberman (1983) και δίνεται από τη σχέση

$$\hat{q}_x^{CH} = \frac{\sum_{i=1}^{n} d_i y\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} l_i y\left(\frac{x - x_i}{h}\right)}.$$

Παρατηρούμε ότι στον αριθμητή λαμβάνονται υπόψη μόνο οι περιπτώσεις για τις οποίες εμφανίζεται θάνατος.

Ο δεύτερος εκτιμητής προτάθηκε από τον Ramlau – Hausen, επίσης το 1983, και δίνεται από τον τύπο

$$\hat{q}_x^{NW} = \frac{\sum_{i=1}^{n} \overset{\circ}{q}_{x_i} y\left(\frac{x - x_i}{h}\right)}{\sum_{i=1}^{n} y\left(\frac{x - x_i}{h}\right)}.$$

Σύμφωνα με τον Haberman (1998) αυτός ο εκτιμητής μπορεί να θεωρηθεί ως το συνεχές ανάλογο της εξομάλυνσης μέσω κινητών σταθμισμένων μέσων όρων.

Όσον αφορά την παράμετρο h, αυτή ελέγχει το βαθμό ομαλότητας. Όσο μεγαλύτερη η τιμή της τόσο περισσότερα δεδομένα χρειάζονται για την εκτίμηση των ποσοστών θνησιμότητας και συνεπώς τόσο πιο ομαλή είναι η εκτίμηση. Η εμπειρία έχει δείξει ότι το πιο σημαντικό σημείο της μεθόδου μέσω εκτιμητών πυρήνα είναι η επιλογή της παραμέτρου h.

Αν και υπάρχουν κάποιες αυθαίρετες μέθοδοι επιλογής, δεν υπάρχει κάποιος γενικός τρόπος καθορισμού της και συνεπώς πρέπει να υπεισέλθει η προσωπική κρίση του ερευνητή. Μια προσέγγιση για την επιλογή τόσο της συνάρτησης πυρήνα όσο και της παραμέτρου παραθύρου, ο αναγνώστης μπορεί να βρει στις Peristera and Kostaki (2004).

Οι μέθοδοι μέσω εκτιμητών πυρήνα, χρειάζονται ιδιαίτερη προσοχή στην εφαρμογή τους στα άκρα των δεδομένων γιατί εκεί εμφανίζεται μεροληψία (Wang, 1998).

Κεφάλαιο 4

ΕΞΟΜΑΛΥΝΣΗ ΜΕ ΧΡΗΣΗ ΤΗΣ ΘΕΩΡΙΑΣ ΠΛΗΡΟΦΟΡΙΩΝ

Στο κεφάλαιο αυτό θα ασχοληθούμε με μια επίσης μη παραμετρική μέθοδο εξομάλυνσης πινάκων θνησιμότητας, η οποία κάνει χρήση της θεωρίας πληροφοριών (Brockett, 1991). Είναι μια όχι και τόσο διαδεδομένη μέθοδος, αλλά αξίζει προσοχής καθώς έχει ένα σοβαρό θεωρητικό υπόβαθρο. Πριν όμως παρουσιάσουμε τη μέθοδο αυτή, ας δούμε κάποιες βασικές έννοιες της θεωρίας πληροφοριών.

4.1 Στατιστική θεωρία πληροφοριών

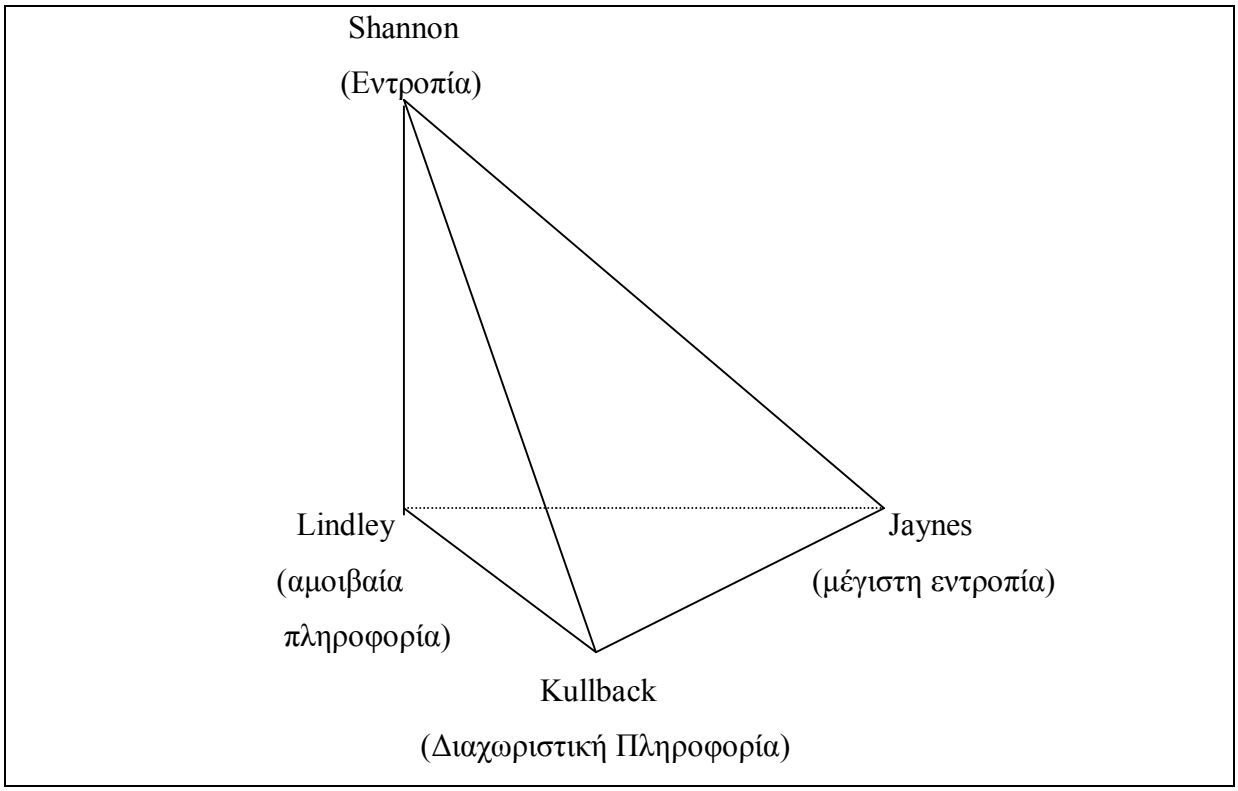
Σε κάθε στατιστικό πρόβλημα που εμπλέκεται δειγματοληψία, ο ερευνητής αναζητά την ποσότητα της πληροφορίας που δίνουν τα δεδομένα για την παράμετρο του πληθυσμού (Kullback, 1959).

Τεχνικά, σύμφωνα με τον Papaioannou, (1985) πληροφορία σημαίνει το ποσό της αβεβαιότητας για μια άγνωστη ποσότητα, το οποίο μειώνεται από το αποτέλεσμα ενός πειράματος. Ο όρος αβεβαιότητα δε σχετίζεται μόνο με την άγνωστη πυκνότητα f ή την παράμετρο q αλλά επίσης με το αποτέλεσμα ενός πειράματος τύχης.

Πριν την εκτέλεση του πειράματος έχουμε μηδενική πληροφορία για την f ή την q ενώ συλλέγοντας παρατηρήσεις από μια τυχαία μεταβλητή X η οποία ακολουθεί κατανομή f ή $f(x,q)$ μειώνεται η αβεβαιότητα για την f ή q. Δυο ομοιομόρφως κατανεμημένες παρατηρήσεις παρέχουν την ίδια ποσότητα πληροφορίας. Στην εκτιμητική, στόχος μας είναι η εύρεση επαρκών στατιστικών συναρτήσεων, δηλαδή στατιστικών που περιέχουν όλη την πληροφορία για την f ή q.

4.2 Μέτρα πληροφορίας

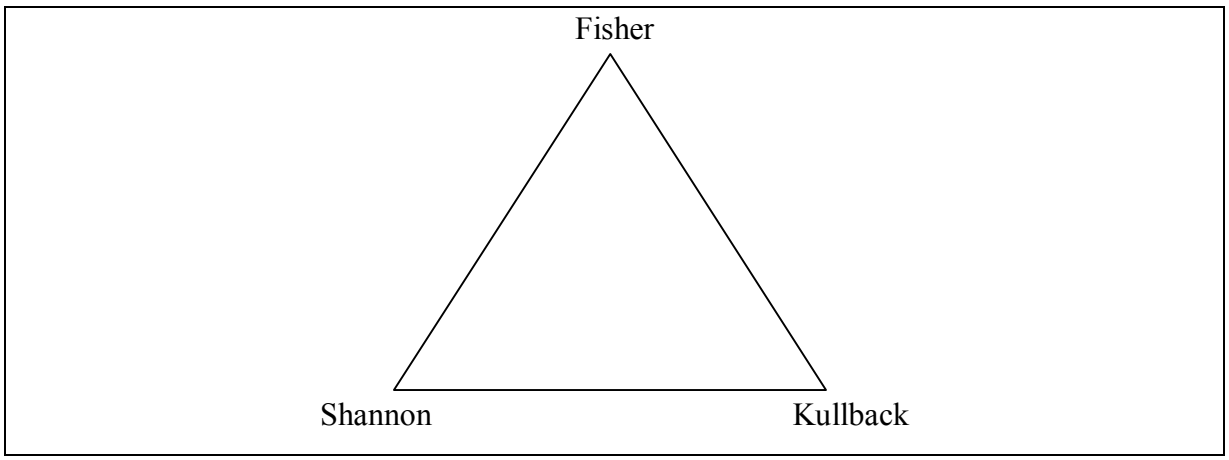
Υπάρχουν πολλά μέτρα πληροφορίας, εκ των οποίων τα πιο βασικά είναι του Fisher, των Kullback και Leibler και του Shannon. Ο Soofi (1994) συνοπτικά αναπαριστά τα μέτρα πληροφορίας με την πυραμίδα Shannon – Kullback – Lindley – Jaynes (Σχήμα 3.1). Στην κορυφή της πυραμίδας τοποθετείται το μέτρο του Shannon, γνωστό ως εντροπία (*entropy*). Στη βάση της πυραμίδας τοποθετούνται τρεις γενικεύσεις της εντροπίας. Το μέτρο του Kullback, η μέγιστη εντροπία του Jaynes και η αμοιβαία πληροφορία του Lindley. Η πλάγια πλευρά μεταξύ των κορυφών Shannon, Kullback και Jaynes είναι το επίπεδο της ελάχιστης διαχωριστικής πληροφορίας. Η πλάγια πλευρά μεταξύ των κορυφών Shannon, Lindley και Kullback είναι το επίπεδο της αμοιβαίας πληροφορίας, ενώ η τρίτη πλάγια πλευρά μεταξύ των κορυφών Shannon, Lindley και Jaynes είναι το επίπεδο της μπεϋζιανής θεωρίας πληροφοριών. Τα υπόλοιπα μέτρα μπορούν να τοποθετηθούν στο εσωτερικό της πυραμίδας.



Σχήμα 3.1: Πυραμίδα Shannon – Kullback – Lindley - Jaynes

Ο Papaioannou (2001) διαφωνεί με την παραπάνω αναπαράσταση, αφού ο Jaynes δεν ανακάλυψε ένα καινούριο μέτρο πληροφορίας αλλά εισήγαγε την αρχή της μέγιστης

εντροπίας. Συνεπώς, προτείνει ότι η στατιστική θεωρία πληροφοριών πρέπει να αναπαριστάται από ένα τρίγωνο (Σχήμα 3.2), στην κορυφή του οποίου θα βρίσκεται το μέτρο του Fisher ενώ στη βάση τα μέτρα των Shannon και Kullback.



Σχήμα 3.2: Τρίγωνο Fisher – Shannon - Kullback

Τα μέτρα πληροφορίας μπορούν να χωριστούν σε τρεις μεγάλες κατηγορίες (Papaioannou, 1985 και 2001). Σε παραμετρικά ή μέτρα τύπου Fisher, σε μη παραμετρικά ή τύπου απόκλισης και σε τύπου εντροπίας μέτρα πληροφορίας

Τα παραμετρικά μέτρα αναφέρονται σε παραμετρικές οικογένειες κατανομών $\{f(x,q),\ q \in \Theta\}$ και μετρούν το ποσό της πληροφορίας που παρέχουν τα δεδομένα για την άγνωστη παράμετρο q. Τα μέτρα αυτά αποτελούν συναρτήσεις του q. Ο κυριότερος εκπρόσωπος της κατηγορίας αυτής είναι το γνωστό *μέτρο πληροφορίας του Fisher*, το οποίο ορίζεται ως

$$I_X^F(q) = E\left(\frac{\partial \ln f(x,q)}{\partial q} \right)^2$$

αν η q είναι μονοδιάστατη παράμετρος ή

$$I_X^F(\boldsymbol{\theta}) = \left(E\left\{ \left(\frac{\partial \ln f(x,\boldsymbol{\theta})}{\partial q_i} \right) \left(\frac{\partial \ln f(x,\boldsymbol{\theta})}{\partial q_j} \right) \right\} \right)_{k \times k}$$

αν η q είναι k – διάστατο διάνυσμα.

Το μέτρο πληροφορίας του Fisher είναι ιδιαίτερης σημασίας στη στατιστική κυρίως επειδή εμφανίζεται στην ανισότητα Cramér – Rao.

Το αντιπροσωπευτικό μέτρο της μη παραμετρικής κατηγορίας (μέτρα απόκλισης) είναι το *μέτρο των Kullback – Leibler*, το οποίο δίνεται από τον τύπο

$$I_X^{KL}(f_1, f_2) = \int f_1(x) \ln \frac{f_1(x)}{f_2(x)} dx$$

και μετρά την απόσταση μεταξύ δυο κατανομών f_1 και f_2. Με άλλα λόγια δίνει την πληροφορία που κερδίζουμε αν αντικαταστήσουμε την f_2 με την f_1. Εναλλακτικά αναφέρεται ως *σχετική ή διασταυρούμενη εντροπία (relative* ή *cross – entropy)* και *πληροφορία διαχωρισμού*.

Το κυριότερο μέτρο της τρίτης κατηγορίας (τύπου εντροπίας) είναι η *εντροπία του Shannon* ή *διαφορική εντροπία* (*differential entropy*). Η εντροπία δίνεται από τον τύπο

$$H(x) = -\sum_x p(x) \log p(x)$$

ή τον

$$H(x) = -\int f(x) \ln f(x) dx$$

για τη διακριτή ή συνεχή περίπτωση αντίστοιχα.

Το μέτρο αυτό ποσοτικοποιεί την αναμενόμενη αβεβαιότητα που σχετίζεται με το αποτέλεσμα ενός πειράματος τύχης. Με άλλα λόγια παρέχει πληροφορία για την προβλεψιμότητα του αποτελέσματος μιας τυχαίας μεταβλητής Χ. Όσο μεγαλύτερη είναι η εντροπία τόσο λιγότερο συγκεντρωμένη είναι η κατανομή της Χ και συνεπώς μια παρατήρηση της Χ παρέχει λίγη πληροφορία. Γενικεύοντας την έννοια της εντροπίας σε δυο τυχαίες μεταβλητές, έχουμε την *κοινή εντροπία* (*joint entropy*)

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y) = -E[\log p(X,Y)],$$

όπου $p(x,y)$ είναι η από κοινού κατανομή των μεταβλητών Χ, Υ. Η *υπό συνθήκη εντροπία* (*conditional entropy*) ορίζεται ως

$$H(Y/X) = E_X\left(E_{Y|X}[\log(Y \mid X)]\right) = -\sum_x \sum_y p(x,y) \log p(y \mid x),$$

όπου $p(x \mid y)$ είναι η δεσμευμένη κατανομή της μεταβλητής Χ δοθέντος της μεταβλητής Υ. Από τον ορισμό της δεσμευμένης κατανομής έπεται ότι ισχύει $H(X,Y) = H(X) + H(Y \mid X)$ (Thomas and Cover, 1991).

Για περισσότερες πληροφορίες για τα παραπάνω μέτρα, καθώς και για άλλα μέτρα πληροφορίας, παραπέμπουμε στους Soofi (1994), Papaioannou (1985 και 2001), Papaioannou

and Ferentinos (2002), Papaioannou and Kempthorne (1971) και Ferentinos and Papaioannou (1981).

Τα μέτρα πληροφορίας έχουν κάποιες ιδιότητες τις οποίες πρέπει να ικανοποιούν τα νεοεμφανιζόμενα μέτρα. Ο Papaioannou (1985 και 2001) αναφέρει και παρουσιάζει με λεπτομέρεια τις ιδιότητες των μέτρων πληροφορίας. Ονομαστικά, αυτές είναι οι εξής: Μη αρνητικότητα (*Nonnegativity*), Προσθετικότητα και υποπροσθετικότητα (*Additivity – subadditivity*), Υπό συνθήκη ανισότητα (*Conditional Inequality*), Μέγιστη πληροφορία (*Maximal information*), Αναλλοίωτα (*Invariant*) υπό επαρκείς μετασχηματισμούς, Κυρτότητα (*Convexity*), Απώλεια πληροφορίας (*Loss of information*), Επάρκεια στα πειράματα (*Sufficiency in experiments*), Εμφάνιση στις ανισότητες Cramér – Rao, Αναλλοίωτα (*Invariant*) υπό παραμετρικούς μετασχηματισμούς, Ανισότητα της ενοχλητικής παραμέτρου (*Nuisance parameter inequality*), Ιδιότητα διατήρησης διάταξης (*Order preserving property*) και Ασυμπτωτική συμπεριφορά.

Από τις παραπάνω ιδιότητες δικαιολογείται η άποψη των στατιστικών επιστημόνων ότι περισσότερα δεδομένα σημαίνουν περισσότερη πληροφορία. Συγκεκριμένα, σ' αυτό συμβάλλουν οι ιδιότητες της προσθετικότητας, της μη αρνητικότητας και της μέγιστης δυνατής πληροφορίας.

Για περισσότερες πληροφορίες, ιδιότητες και παραδείγματα παραπέμπουμε στους Papaioannou (2001) και Kullback (1959).

4.3 Ελάχιστη Διαχωριστική Πληροφορία

Από την πλευρά της στατιστικής θεωρίας πληροφοριών, ο Brockett (1991) επιτυγχάνει την ομαλή εκτίμηση της πραγματικής κατανομής των πιθανοτήτων θνησιμότητας ενός ασφαλιστικού πίνακα θνησιμότητας μέσω της ελαχιστοποίησης του μέτρου των Kullback – Leibler,

$$I_X^{KL}(f,g) = \int f(x) \ln \frac{f(x)}{g(x)} dx .$$

Όπως είδαμε στην προηγούμενη παράγραφο, το μέτρο αυτό, μετρά την απόκλιση μεταξύ των δυο κατανομών πιθανότητας f και g, και ισχύει η σχέση $I_X^{KL}(f,g) \geq 0$ με ισότητα αν και μόνο αν $f = g$. Έτσι κανείς μπορεί να εκτιμήσει την g μέσω της f η οποία είναι όσο το δυνατόν πιο κοντά από πλευράς απόστασης πληροφορίας στην g. Αυτό επιτυγχάνεται με

την ελαχιστοποίηση του μέτρου των Kullback – Leibler. Είναι κατανοητό ότι η ελαχιστοποίηση του $I_X^{KL}(f,g)$ μπορεί να γίνει θέτοντας ορισμένους περιορισμούς για την f. Η μέθοδος εκτίμησης αυτή ονομάζεται *Ελάχιστη Διαχωριστική Πληροφορία* (*Minimum Discrimination Information*, *MDI*).

Γι' αυτό το σκοπό ο Kullback (1951) ελαχιστοποιεί τη Λαγκρανζιανή ποσότητα

$$\int\left(f(x)\ln\frac{f(x)}{g(x)} + kT(x)f(x) + lf(x)\right)dx,$$

με τον περιορισμό $q = \int T(x)f(x)dx$, όπου $T(x)$ είναι ένα στατιστικό, k, l είναι αυθαίρετοι πολλαπλασιαστές και καταλήγει ότι η ελάχιστη τιμή της παραπάνω ποσότητας είναι η

$$f^*(x) = g(x)e^{-kT(x)-l-1}.$$

Στη συνέχεια, αντικαθιστώντας το $-k$ με t, για εννοιολογική ευκολία, και θέτοντας

$$\mathrm{M}_2(t) = \int g(x)e^{t\,T(x)}dx,\ \mathrm{M}_2(t) < \infty$$

καταλήγει στο παρακάτω θεώρημα.

<u>Θεώρημα</u>

Αν $f(x)$ και $g(x)$ είναι πυκνότητες πιθανότητας, $T(x)$ είναι στατιστικό τέτοιο ώστε να υπάρχει περιορισμός $q = \int T(x)f(x)dx$ και το $\mathrm{M}_2(t) = \int g(x)e^{t\,T(x)}dx$ υπάρχει σε κάποιο διάστημα, τότε $I_X^{KL}(f,g) \geq q_t - \ln \mathrm{M}_2(t)$ όπου $q = \frac{\partial}{\partial t}\ln \mathrm{M}_2(t)$ με ισότητα αν και μόνο αν $f(x) = f^*(x) = \frac{e^{t\,\mathrm{T}(x)}g(x)}{M_2(t)}$.

Η πυκνότητα f^* ονομάζεται *εκτιμητής Ελάχιστης Διαχωριστικής Πληροφορίας* για την g υπό τον περιορισμό για την f, $q = \int T(x)f(x)dx$. Σημειώνουμε ότι η ελαχιστοποίηση του μέτρου των Kullback – Leibler, καταλήγει στην εκθετική οικογένεια κατανομών.

Είναι φανερό ότι για την εύρεση της εκτίμησης f^* είναι απαραίτητη η επίλυση της εξίσωσης $q = \frac{\partial}{\partial t}\ln \mathrm{M}_2(t)$ ως προς t, πράγμα όχι και τόσο εύκολο, ακόμα και στην περίπτωση που έχουμε περιορισμούς της μορφής γραμμικών ισοτήτων (Zhang and Brockett, 1987).

Οι περιορισμοί που θέτει ο ερευνητής για την πυκνότητα f, είναι της μορφής γραμμικών ισοτήτων, γραμμικών ανισοτήτων ακόμη και τετραγωνικών ανισοτήτων, όπως συμβαίνει στο πρόβλημα της εξομάλυνσης των πινάκων θνησιμότητας που θα μελετήσουμε στη συνέχεια. Σε διάφορες περιπτώσεις όπως στη μηχανική και τις κοινωνικές επιστήμες, το σύνολο των περιορισμών μπορεί να είναι ένας συνδυασμός των παραπάνω μορφών (Brockett, 1991).

Επειδή η επίλυση προβλημάτων με τέτοιας μορφής περιορισμούς είναι δύσκολη, ο Brockett (1991) προτείνει την προσέγγιση του θέματος από την πλευρά του δυϊκού κυρτού προγραμματισμού. Συνοπτικά αναφέρουμε ότι με βάση την προσέγγιση αυτή, αντί να ελαχιστοποιήσουμε μια ποσότητα, μεγιστοποιούμε τη δυϊκή έκφραση αυτής, πράγμα που είναι σαφώς ευκολότερο.

4.4 Ελαχιστοποίηση του διακριτού μέτρου των Kullback – Leibler με γραμμικούς και τετραγωνικούς περιορισμούς

Στη διακριτή περίπτωση, η αντικειμενική συνάρτηση την οποία θέλουμε να ελαχιστοποιήσουμε είναι η

$$I^{KL}(\boldsymbol{\delta},\mathbf{d}) = \sum_{i=1}^{n} d_i \ln \frac{d_i}{e d_i}$$

με γραμμικούς και τετραγωνικούς περιορισμούς, όπου $\boldsymbol{\delta} = (d_1, d_2, ..., d_n)^T$ και $\mathbf{d} = (d_1, d_2, ..., d_n)^T$ είναι δυο πεπερασμένα μέτρα από τα οποία τα d_i είναι γνωστά ενώ τα d_i είναι άγνωστα. Το T δηλώνει την αντιστροφή του διανύσματος. Ο παράγοντας e στον παρονομαστή του νεπέρειου λογαρίθμου της αντικειμενικής συνάρτησης δεν επηρεάζει το πρόβλημα της βελτιστοποίησης όταν το $\sum_{i=1}^{n} d_i$ είναι σταθερό, αλλά όπως ισχυρίζονται οι Zhang and Brockett (1987), διευκολύνει στη διατύπωση και λύση του προβλήματος μέσω του δυϊκού (*dual*) του όπως δίνεται πιο κάτω.

Το πρωτεύον (*primal*) πρόβλημα στην περίπτωση των γραμμικών περιορισμών, το οποίο είναι το

(P_L) $\min I^{KL}(\boldsymbol{\delta},\mathbf{d})$

υπό τους περιορισμούς $\mathbf{A}^T\boldsymbol{\delta} \le \mathbf{v}$, $\boldsymbol{\delta} \ge \mathbf{0}$.

όπου $\mathbf{A}$ είναι ένας $n \times k$ πίνακας, $\mathbf{v}$ είναι ένα $k \times 1$ διάνυσμα σταθερών ποσοτήτων, $\boldsymbol{\delta}$ και $\mathbf{d}$ όπως πριν ενώ το T δηλώνει την αντιστροφή του πίνακα ή του διανύσματος.

Οι Brockett, Charnes and Cooper (1980) αποδεικνύουν ότι το δυϊκό (*dual*) πρόβλημα του πρωτεύοντος προβλήματος (P_L) είναι το

(D_L) $\max - \mathbf{d}^T \mathbf{e}^{\mathbf{Az}} + \mathbf{v}^T \mathbf{z}$

υπό τον περιορισμό $\mathbf{z} \leq \mathbf{0}$.

Το δυϊκό πρόβλημα (D_L), μπορεί να γραφεί και στη μορφή $\max - \sum_{i=1}^{n} d_i \mathbf{e}^{{}_i\mathbf{A}\mathbf{z}} + \mathbf{v}^T \mathbf{z}$, όπου το ${}_i\mathbf{A}$ συμβολίζει την i γραμμή του πίνακα $\mathbf{A}$, το $\mathbf{z}$ είναι ένα $k \times 1$ διάνυσμα άγνωστων σταθερών ποσοτήτων και το $\mathbf{e}^{\mathbf{x}}$, με $\mathbf{x}$ ένα $n \times 1$ διάνυσμα, συμβολίζει το διάνυσμα $\left(e^{x_1}, e^{x_2}, ..., e^{x_n}\right)^T$.

Στην περίπτωση αυτή ισχύει η δυϊκή ανισότητα

$$\sum_{i=1}^{n} d_i \ln \frac{d_i}{e d_i} \geq -\mathbf{d}^T \mathbf{e}^{\mathbf{Az}} + \mathbf{v}^T \mathbf{z}$$

όταν ισχύουν ταυτόχρονα οι περιορισμοί $\mathbf{A}^T \boldsymbol{\delta} \leq \mathbf{v}$, $\boldsymbol{\delta} \geq \mathbf{0}$ και $\mathbf{z} \leq \mathbf{0}$.

Οι Zhang and Brockett, στην ίδια εργασία, γενίκευσαν το πρόβλημα (P_L), έτσι ώστε να συμπεριλάβουν και περιορισμούς της μορφής τετραγωνικών ανισοτήτων, όπως θα δούμε στην επόμενη ενότητα. Η περίπτωση αυτή αφορά εξομάλυνση πινάκων θνησιμότητας. Συγκεκριμένα το πρωτεύον πρόβλημα στην περίπτωση αυτή είναι το

(P) $\min I^{KL}(\boldsymbol{\delta}, \mathbf{d}) = \sum_{i=1}^{n} d_i \ln \frac{d_i}{e d_i}$

υπό τους περιορισμούς $\boldsymbol{\delta} \geq \mathbf{0}$ και $g_i(\boldsymbol{\delta}) = \frac{1}{2} \boldsymbol{\delta}^T \mathbf{D}_i \boldsymbol{\delta} + \mathbf{b}_i^T \boldsymbol{\delta} + c_i \leq 0$, $i = 1,2,...,r$,

όπου ο $\mathbf{D}_i$ είναι θετικά ημιορισμένος πίνακας για κάθε i και τα $\mathbf{b}_i$, c_i είναι σταθερές.

Παρατηρούμε ότι το πρόβλημα (P) περιέχει σαν ειδικές περιπτώσεις τα προβλήματα με περιορισμούς της μορφής γραμμικών ισοτήτων και γραμμικών ανισοτήτων.

Συγκεκριμένα χρησιμοποιώντας έννοιες και μεθόδους του μαθηματικού προγραμματισμού, οι Zhang and Brockett (1987) απέδειξαν ότι η λύση του προβλήματος (P), μπορεί να βρεθεί επιλύοντας το ακόλουθο δυϊκό πρόβλημα:

$$(D) \quad \max -\mathbf{d}^T \exp\left\{\left[-\sum_{i=1}^{r} y_i(\mathbf{A}_i^T \mathbf{w}_i + \mathbf{b}_i)\right]\right\} + \mathbf{c}^T \mathbf{y} - \frac{1}{2}\sum_{i=1}^{r} r_i$$

υπό τους περιορισμούς $\mathbf{y} \geq \mathbf{0}$ και $\mathbf{w}_i \in \mathbf{R}^{m_i}$,

του οποίου η αντικειμενική συνάρτηση είναι μη γραμμική και δεν έχει περιορισμούς.

Το $\mathbf{A}_i$ είναι ένας $m_i \times n$ πίνακας, ο οποίος προκύπτει από τον πίνακα $\mathbf{D}_i$ του πρωτεύοντος προβλήματος (P), μέσω της σχέσης $\mathbf{D}_i = \mathbf{A}_i^T \mathbf{A}_i$, όπου m_i είναι η τάξη του πίνακα $\mathbf{D}_i$, $\mathbf{w}_i$ είναι ένα $m_i \times 1$ διάνυσμα πραγματικών αριθμών, $\mathbf{y}$ είναι ένα $r \times 1$ διάνυσμα άγνωστων μεταβλητών και $r_i = y_i \|\mathbf{w}_i\|^2$.

Παρατηρούμε ότι το δυϊκό πρόβλημα (D) δεν έχει κανένα περιορισμό πέραν της μη αρνητικότητας ενός μικρού αριθμού μεταβλητών και συνεπώς επιλύεται ευκολότερα από το πρωτεύον πρόβλημα (P). Επίσης παρατηρούμε ότι το πρόβλημα (D) είναι μια επέκταση του προβλήματος (D_L) με την έννοια ότι όλοι οι περιορισμοί του (D_L) είναι γραμμικοί, δηλαδή αν $\mathbf{D}_i = \mathbf{0}$ τότε το (D) είναι όμοιο με το (D_L).

Συνοψίζοντας την παραπάνω μέθοδο, μπορούμε να πούμε ότι για να βρούμε το δυϊκό πρόβλημα ενός προβλήματος με κυρτής μορφής περιορισμούς, αρκεί να γνωρίζουμε τη δυϊκή μορφή ενός κυρτού προβλήματος με γραμμικούς περιορισμούς. Συγκεκριμένα, με ένα σύνολο βοηθητικών υπερεπιπέδων προσεγγίζουμε την εφικτή περιοχή, βρίσκουμε τη δυϊκή έκφραση για το προκύπτον πρόβλημα που έχει γραμμικούς περιορισμούς και στη συνέχεια παίρνουμε όρια και απλοποιούμε την προκύπτουσα έκφραση.

Για περισσότερες πληροφορίες για την εξαγωγή του προβλήματος (D), παραπέμπουμε στους Zhang and Brockett (1987).

4.5 Εφαρμογή σε πίνακες θνησιμότητας

Ας δούμε τώρα πώς η παραπάνω μέθοδος, μπορεί να χρησιμοποιηθεί στην αναλογιστική επιστήμη για την εξομάλυνση των πινάκων θνησιμότητας (Brockett and Zhang, 1986, Zhang and Brockett, 1987 και Brockett, 1991). Όπως έχουμε ήδη αναφέρει, η μέθοδος αυτή δίνει μια σειρά τιμών (ποσοστών θνησιμότητας) όσο το δυνατόν πιο κοντά στις παρατηρούμενες τιμές, οι οποίες επιπλέον ικανοποιούν κάποιους περιορισμούς, οι οποίοι αποτελούν την εκ των προτέρων άποψη σχετικά με το πραγματικό πρότυπο θνησιμότητας.

Σύμφωνα με τον Brockett (1991), οι εξομαλυμένες τιμές θα πρέπει να ικανοποιούν πέντε περιορισμούς. Συγκεκριμένα η σειρά των εξομαλυμένων τιμών v_x θα πρέπει (i) να είναι ομαλή (*ομαλότητα*), (ii) να αυξάνεται με την ηλικία (*μονοτονία*), και (iii) να αυξάνεται απότομα στις μεγάλες ηλικίες (*κυρτότητα*). Επίσης θα πρέπει (iv) ο αριθμός των θανάτων στα εξομαλυμένα δεδομένα v_x να ισούται με τον αντίστοιχο αριθμό των παρατηρούμενων τιμών u_x και (v) το σύνολο των εξομαλυμένων ηλικιών θανάτου να ισούται με τις παρατηρούμενες συνολικές ηλικίες θανάτου. Ως συνολική ηλικία θανάτου εννοούμε το άθροισμα του γινομένου του αριθμού θανάτων σε κάθε ηλικία επί την αντίστοιχη ηλικία. Αν συνδυάσουμε τους δυο τελευταίους περιορισμούς (αναλογιστικοί περιορισμοί), συμπεραίνουμε ότι η μέση ηλικία θανάτου θα πρέπει να είναι ίδια τόσο για τα εξομαλυμένα όσο και για τα παρατηρούμενα δεδομένα.

Οι παραπάνω περιορισμοί γράφονται μαθηματικά ως εξής:

(i) $\sum_x \left(\Delta^z v_x\right)^2 \le M$,

όπου $\sum_x \left(\Delta^z v_x\right)^2$, $x = 1, 2, ..., n - z$ είναι το μέτρο ομαλότητας με $z = 3$ ή 4, v_x είναι οι εξομαλυμένες τιμές και M είναι μια σταθερά ομαλότητας.

(ii) $\Delta v_x \ge 0$,

όπου $\Delta v_x = v_{x+1} - v_x$ για κάθε x .

(iii) $\Delta^2 v_x \ge 0$,

όπου $\Delta^2 v_x = v_x - 2v_{x+1} + v_{x+2}$ για κάθε x .

(iv) $\sum_x l_x v_x = \sum_x l_x u_x$,

όπου l_x είναι ο αριθμός των ατόμων σε κίνδυνο στην ηλικία x και u_x τα παρατηρούμενα δεδομένα.

(v) $\sum_x x l_x v_x = \sum_x x l_x u_x$.

Για να μοιάζουν οι παραπάνω περιορισμοί με τους αντίστοιχους του αρχικού προβλήματος (P), τους γράφουμε σε μορφή πινάκων ως εξής:

(i) $(\mathbf{Av})^T(\mathbf{Av}) = \mathbf{v}^T\mathbf{A}^T\mathbf{Av} \le M$,

όπου $\mathbf{A} = \begin{bmatrix} -1 & 3 & -3 & 1 & 0 & \mathbf{L} & 0 \\ 0 & -1 & 3 & -3 & 1 & \mathbf{L} & 0 \\ 0 & 0 & -1 & 3 & -3 & \mathbf{L} & 0 \\ \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} \\ 0 & 0 & 0 & 0 & 0 & \mathbf{L} & 1 \end{bmatrix}$ αν $z = 3$ (ομαλότητα).

(ii) $\mathbf{Bv} \ge \mathbf{0}$,

όπου $\mathbf{B} = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & \mathbf{L} & 0 \\ 0 & -1 & 1 & 0 & 0 & \mathbf{L} & 0 \\ 0 & 0 & -1 & 1 & 0 & \mathbf{L} & 0 \\ \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} \\ 0 & 0 & 0 & 0 & 0 & \mathbf{L} & 1 \end{bmatrix}$ (μονοτονία).

(iii) $\mathbf{Cv} \ge \mathbf{0}$,

όπου $\mathbf{C} = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 & \mathbf{L} & 0 \\ 0 & 1 & -2 & 1 & 0 & \mathbf{L} & 0 \\ 0 & 0 & 1 & -2 & 1 & \mathbf{L} & 0 \\ \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} & \mathbf{M} \\ 0 & 0 & 0 & 0 & 0 & \mathbf{L} & 1 \end{bmatrix}$ (κυρτότητα).

(iv) $\mathbf{l}^T\mathbf{v} = \mathbf{l}^T\mathbf{u}$,

όπου $\mathbf{l} = (l_x, l_{x+1}, ..., l_{x+n})^T$ και

(v) $\mathbf{m}^T\mathbf{v} = \mathbf{m}^T\mathbf{u}$,

όπου $\mathbf{m} = \left(xl_x, (x+1)l_{x+1}, ..., (x+n)l_{x+n}\right)^T$.

Να σημειώσουμε ότι οι περιορισμοί της μορφής $\mathbf{z} = \mathbf{w}$ γράφονται ως $\mathbf{z} - \mathbf{w} \ge \mathbf{0}$ και $-\mathbf{z} + \mathbf{w} \ge \mathbf{0}$, όπου $\mathbf{z}$, $\mathbf{w}$ είναι διανύσματα.

Σαν μέτρο προσαρμογής, χρησιμοποιείται το μέτρο των Kullback – Leibler μεταξύ των σειρών των εξομαλυμένων και των παρατηρούμενων τιμών, $I^{KL}(\mathbf{v},\mathbf{u}) = \sum_x v_x \ln \frac{v_x}{u_x}$, $x = 1, 2, ..., n$. Βέβαια όπως έχουμε αναφέρει τα $\mathbf{v}$ και $\mathbf{u}$ θα πρέπει να είναι κατανομές πιθανότητας. Παρόλο που αυτό δεν συμβαίνει στην περίπτωση αυτή, το μέτρο $I^{KL}(\mathbf{v},\mathbf{u})$ είναι μέτρο προσαρμογής γιατί τα ποσοστά θνησιμότητας είναι μη μηδενικά και εξαιτίας των περιορισμών που έχουμε ορίσει (Brockett, 1991). Ελαχιστοποιώντας αυτή την ποσότητα με βάση τους παραπάνω περιορισμούς, παίρνουμε μια σειρά εκτιμήσεων (εξομαλυμένων τιμών) $\{v_x\}$, η οποία ικανοποιεί τους περιορισμούς και επιπλέον είναι η λιγότερο διαχωρίσιμη από τη σειρά των πραγματικών τιμών $\{u_x\}$.

Η επίλυση του παραπάνω προβλήματος μπορεί να γίνει με οποιονδήποτε κώδικα μη γραμμικού προγραμματισμού.

Για την περίπτωση της πολυμεταβλητής εξομάλυνσης, παραπέμπουμε στον Brockett (1991), όπου περιγράφει την εξομάλυνση ενός διμεταβλητού πίνακα θνησιμότητας, ο οποίος έχει κατασκευαστεί με βάση την ηλικία επιλογής των ασφαλισμένων και το χρόνο μέχρι την επιλογή.

Όσον αφορά την τιμή της ποσότητας M μπορούμε να αναφέρουμε τα εξής (Brockett, 1991): Το M είναι μια ποσότητα που καθορίζει το πόσο ομαλή θα είναι η σειρά των εξομαλυμένων τιμών, οπότε αυξομειώνοντας το M θα αυξομειώνεται και ο βαθμός ομαλότητας της σειράς και περισσότερη ή λιγότερη αντίστοιχα έμφαση θα δίνεται στην ομαλότητα έναντι της προσαρμογής. Γενικός κανόνας για την επιλογή της τιμής του M δεν υπάρχει. Αυτό που μπορούμε να πούμε, είναι ότι εφόσον θέλουμε να πετύχουμε ομαλά δεδομένα, τότε το M θα πρέπει να πάρει μια μικρή τιμή, η οποία να προσεγγίζει το 0. Η τιμή του M μπορεί να επιλεχθεί είτε με βάση κάποια άλλη εξομάλυνση που έχει γίνει για τα ίδια δεδομένα με διαφορετική βέβαια μέθοδο είτε κάνοντας διάφορες εξομαλύνσεις με διαφορετικές τιμές της ποσότητας M οπότε επιλέγουμε αυτή που δίνει τα καλύτερα αποτελέσματα.

Η εφαρμογή της παραπάνω μεθόδου σε δεδομένα, δίνεται στο επόμενο κεφάλαιο.

Κεφάλαιο 5

ΕΦΑΡΜΟΓΗ ΟΡΙΣΜΕΝΩΝ ΜΕΘΟΔΩΝ ΕΞΟΜΑΛΥΝΣΗΣ ΣΕ ΔΕΔΟΜΕΝΑ

Στο κεφάλαιο αυτό θα δούμε τα αποτελέσματα της εξομάλυνσης μιας σειράς αρχικών εκτιμήσεων, που προκύπτουν από κάποιες από τις μεθόδους που περιγράψαμε στα προηγούμενα κεφάλαια.

Τα δεδομένα που θα χρησιμοποιήσουμε είναι ποσοστά θνησιμότητας, τα έχουμε πάρει από τον London (1985), σελ. 20 και εμφανίζονται στον Πίνακα 5.1. Συγκεκριμένα στον Πίνακα 5.1 εμφανίζονται η ηλικία, ο αριθμός των ατόμων που βρίσκονται σε κίνδυνο σε κάθε ηλικία, ο αριθμός των θανάτων σε κάθε ηλικία και οι αντίστοιχες αρχικές εκτιμήσεις των ποσοστών θνησιμότητας. Όπως έχουμε ήδη αναφέρει, θεωρούμε ότι ο αριθμός των θανάτων στην ηλικία x, d_x, κατανέμεται ως διωνυμική κατανομή, που βασίζεται σε ένα τυχαίο δείγμα μεγέθους l_x, όπου l_x είναι ο αριθμός των επιζώντων ατόμων στην ηλικία x, $x = 1, 2, ..., n$. Συνεπώς η αρχική εκτίμηση των πραγματικών ποσοστών θνησιμότητας t_x είναι η $u_x = \frac{d_x}{l_x}$.

Σύμφωνα με τον London (1985), η εξομάλυνση θα γίνει έτσι ώστε να λάβουμε υπόψη την εκ των προτέρων άποψη για τη σειρά των πραγματικών ποσοστών θνησιμότητας. Συγκεκριμένα υποθέτουμε ότι το πραγματικό πρότυπο θνησιμότητας είναι (α) ομαλό (β) μονότονο, δηλαδή τα ποσοστά αυξάνονται με την ηλικία και (γ) κυρτό, δηλαδή τα ποσοστά αυξάνονται απότομα στις μεγάλες ηλικίες.

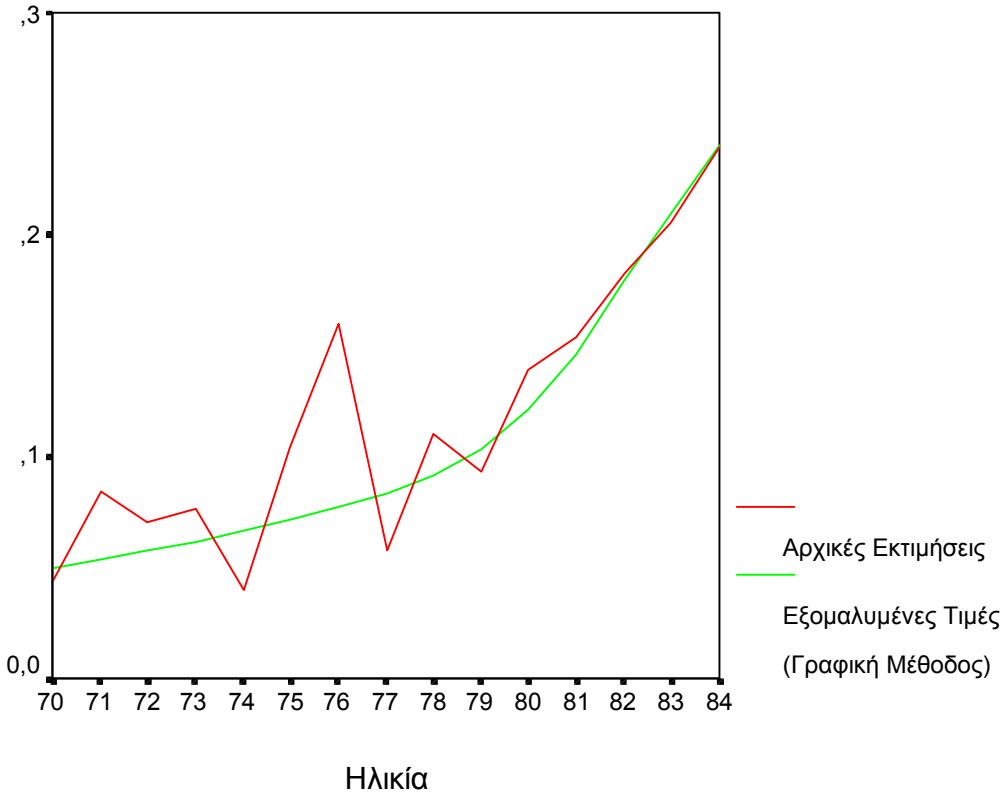
Στη συνέχεια θα δούμε τα εξομαλυμένα αποτελέσματα που παίρνουμε από την εφαρμογή της μεθόδου των Whittaker – Henderson, της μεθόδου των splines και της μεθόδου του Brockett με χρήση της θεωρίας πληροφοριών. Για κάθε μέθοδο θα υπολογίσουμε το μέτρο ομαλότητας, ως $S = \sum_{x=1}^{n-3} \left(\Delta^3 v_x\right)^2$, όπου v_x είναι τα εξομαλυμένα ποσοστά και το μέτρο

προσαρμογής $F = \sum_{x=1}^{n} w_x (u_x - v_x)^2$, που είναι ίδιο με το μέτρο F_2 που είδαμε στην ενότητα 1.4, χρησιμοποιώντας ως βάρη τα $w_x = \frac{l_x}{v_x(1-v_x)}$.

Στον Πίνακα 5.1, εμφανίζονται και τα αποτελέσματα της εξομάλυνσης με τη *γραφική μέθοδο* όπως τα δίνει ο London (1985). Επίσης στο Σχήμα 5.1, εμφανίζεται η γραφική παράσταση των αρχικών εκτιμήσεων και των εξομαλυμένων τιμών με τη γραφική μέθοδο. Υπολογίζοντας τα μέτρα ομαλότητας και προσαρμογής για τα δεδομένα της γραφικής μεθόδου, παίρνουμε αντίστοιχα $S = 0.000202$ και $F = 24.6109$.

Ηλικία x	**Αριθμός Ατόμων σε Κίνδυνο** l_x	**Αριθμός Θανάτων** d_x	**Αρχικές Εκτιμήσεις** u_x	**Εξομαλυμένα Ποσοστά** v_x
70	135	6	0.044	0.050
71	143	12	0.084	0.054
72	140	10	0.071	0.058
73	144	11	0.076	0.062
74	149	6	0.040	0.067
75	154	16	0.104	0.072
76	150	24	0.160	0.077
77	139	8	0.058	0.083
78	145	16	0.110	0.091
79	140	13	0.093	0.103
80	137	19	0.139	0.121
81	136	21	0.154	0.146
82	126	23	0.183	0.180
83	126	26	0.206	0.210
84	109	26	0.239	0.240

Πίνακας 5.1: Αρχικές εκτιμήσεις u_x και εξομαλυμένα ποσοστά θνησιμότητας v_x με τη γραφική μέθοδο



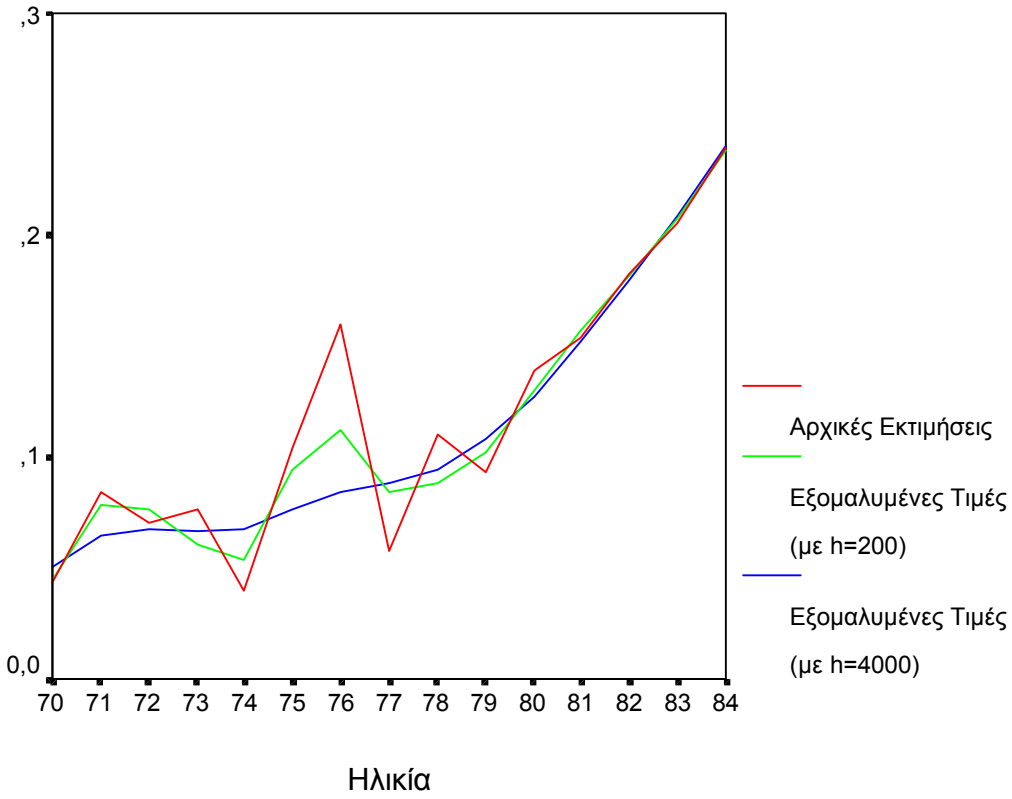
Σχήμα 5.1: Γραφική παράσταση αρχικών εκτιμήσεων και εξομαλυμένων τιμών με τη γραφική μέθοδο

5.1 Μέθοδος Whittaker – Henderson

Όπως είδαμε στην ενότητα 3.4, η μέθοδος των Whittaker – Henderson, χρησιμοποιεί τον τύπο $M = F + hS = \sum_{x=1}^{n} w_x (u_x - v_x)^2 + h\sum_{x=1}^{n-z} (\Delta^z v_x)^2$, όπου u_x είναι οι αρχικές εκτιμήσεις, v_x είναι οι εξομαλυμένες τιμές, h είναι ο παράγοντας που ελέγχει την ομαλότητα και z είναι η τάξη των διαφορών. Ως βάρη θα χρησιμοποιήσουμε τα $w_x = \frac{l_x}{u_x(1-u_x)}$. Για να τονίσουμε το ρόλο που παίζει η παράμετρος h θα κάνουμε δυο εξομαλύνσεις, μια με $h = 200$ και μια δεύτερη με $h = 4000$. Τα αποτελέσματα και των δυο εξομαλύνσεων φαίνονται στον Πίνακα 5.2. Παρατηρούμε ότι η μεγαλύτερη τιμή του h δίνει ομαλότερα αποτελέσματα. Βέβαια η τιμή 4000 για την παράμετρο h επιλέχθηκε έτσι ώστε να πάρουμε τον ίδιο περίπου βαθμό ομαλότητας με αυτόν που πέτυχε ο London (1985) με τη γραφική μέθοδο. Τα αποτελέσματα φαίνονται γραφικά στο Σχήμα 5.2.

Ηλικία	Εξομαλυμένα ποσοστά με h = 200	Εξομαλυμένα ποσοστά με h = 4000
70	0.045	0.051
71	0.078	0.065
72	0.076	0.068
73	0.061	0.067
74	0.054	0.068
75	0.094	0.076
76	0.112	0.084
77	0.084	0.088
78	0.088	0.094
79	0.102	0.108
80	0.130	0.127
81	0.157	0.152
82	0.182	0.180
83	0.208	0.209
84	0.238	0.240
	$S = 0.0146145$	$S = 0.000253768$
	$F = 7.24123$	$F = 18.4375$

Πίνακας 5.2: Εξομαλυμένα ποσοστά θνησιμότητας με h = 200 και h = 4000



Σχήμα 5.2: Γραφική παράσταση αρχικών εκτιμήσεων και εξομαλυμένων τιμών με τη μέθοδο Whittaker – Henderson με h = 200 και h = 4000

5.2 Μέθοδος των splines

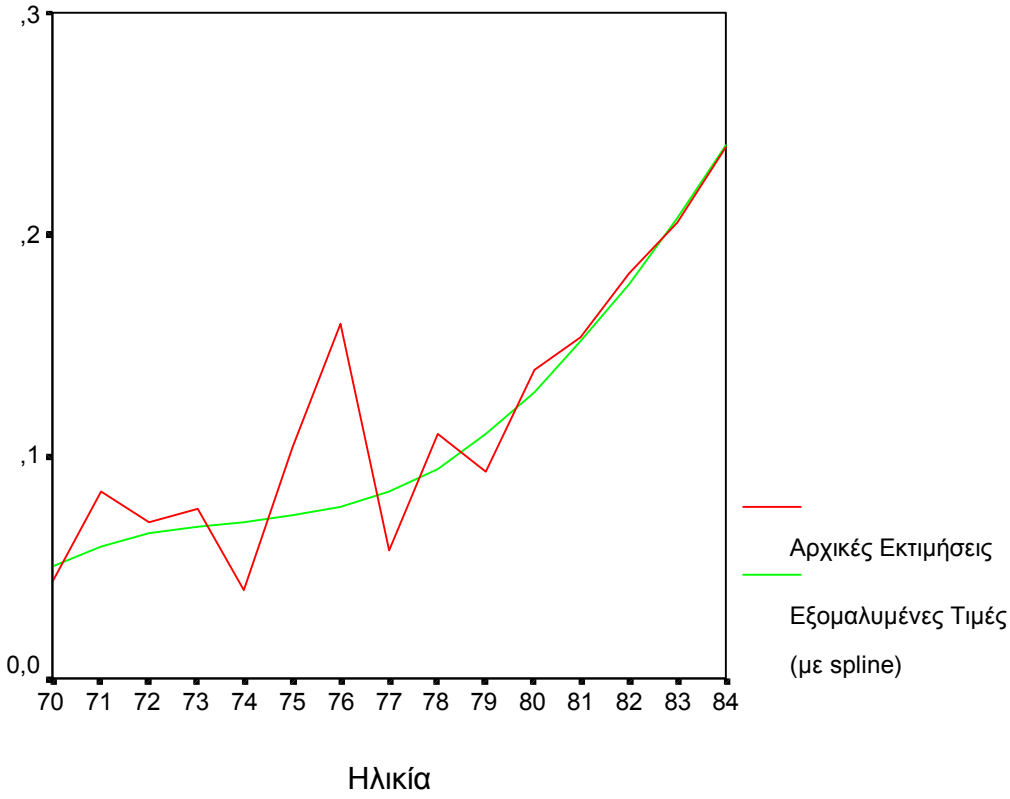
Για να εξομαλύνουμε τις αρχικές εκτιμήσεις των ποσοστών θνησιμότητας, θα χρησιμοποιήσουμε μια δίτοξη τρίτου βαθμού spline με ένα δεσμό στην ηλικία $x = 77.5$. Δηλαδή οι εξομαλυμένες τιμές v_x δίνονται από τη σχέση

$$v_x = \begin{cases} c_0 + c_1 x + c_2 x^2 + c_3 x^3, & 70 \le x \le 77.5 \\ c_0 + c_1 x + c_2 x^2 + c_3 x^3 + c_4 (x - 77.5)^3, & 77.84 \le x \le b, \end{cases}$$

όπου τα c_0, c_1, c_2, c_3 και c_4 είναι άγνωστες παράμετροι. Σημειώνουμε ότι ο δεσμός επιλέχθηκε αυθαίρετα. Τα εξομαλυμένα ποσοστά θνησιμότητας v_x δίνονται στον Πίνακα 5.3 ενώ αναπαρίστανται γραφικά στο Σχήμα 5.3. Επειδή η μέθοδος των splines ανήκει στην παραμετρική οικογένεια των μεθόδων εξομάλυνσης, δεν υπολογίζουμε το μέτρο ομαλότητας S αφού η ομαλότητα είναι εξασφαλισμένη.

Ηλικία	**Εξομαλυμένες τιμές με δεσμό στην ηλικία x = 77.5**
70	0.051
71	0.060
72	0.066
73	0.069
74	0.071
75	0.074
76	0.077
77	0.084
78	0.094
79	0.110
80	0.129
81	0.152
82	0.178
83	0.208
84	0.240
	$F = 22.5507$

Πίνακας 5.3: Εξομαλυμένες τιμές με τη μέθοδο των splines και δεσμό στην ηλικία x = 77.5



Σχήμα 5.3: Γραφική παράσταση αρχικών εκτιμήσεων και εξομαλυμένων τιμών με τη μέθοδο των splines, με δεσμό στο x = 77.5

5.3 Μέθοδος θεωρίας πληροφοριών

Ο Brockett (1991) εξομαλύνει τα ίδια δεδομένα, κάνοντας δυο επιπλέον υποθέσεις για τα πραγματικά ποσοστά θνησιμότητας. Συγκεκριμένα υποθέτει ότι (α) ο εξομαλυμένος αριθμός θανάτων ισούται με τον παρατηρούμενο αριθμό θανάτων και ότι (β) η συνολική εξομαλυμένη ηλικία θανάτου ισούται με τη συνολική παρατηρούμενη ηλικία θανάτου, όπου όπως έχουμε ήδη αναφέρει στην ενότητα που περιγράψαμε τη μέθοδο αυτή, η συνολική ηλικία θανάτου είναι το άθροισμα του γινομένου της ηλικίας επί τον αριθμό θανάτων σε κάθε ηλικία.

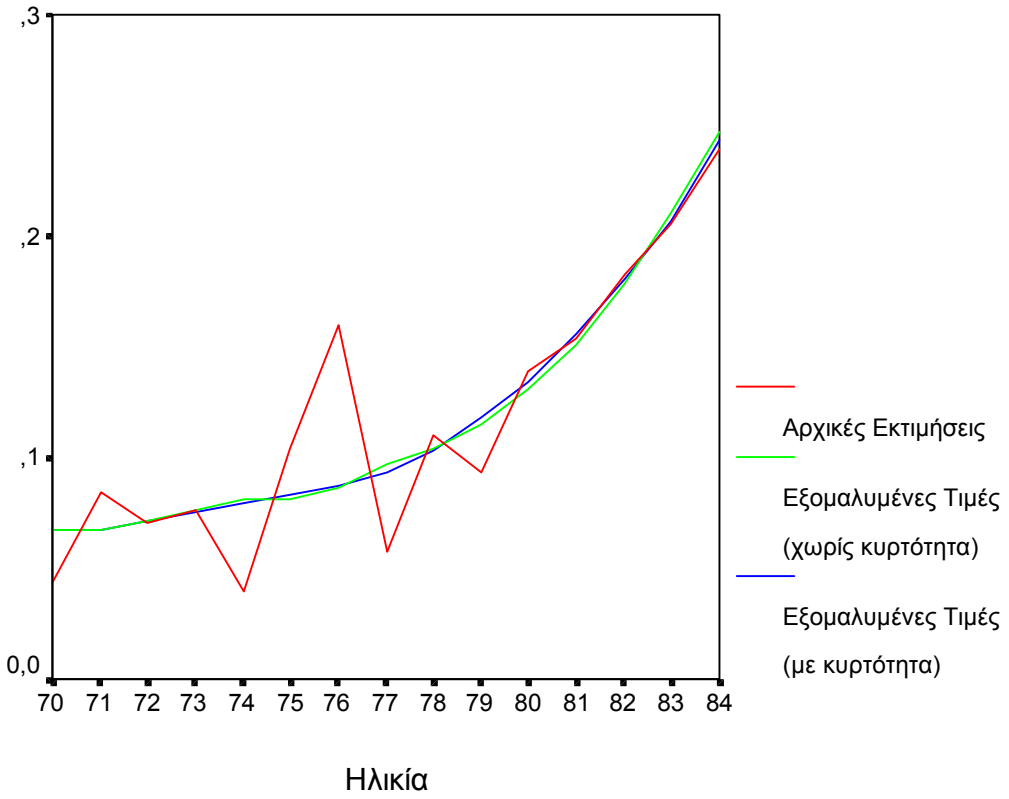
Ελαχιστοποιώντας το μέτρο απόκλισης $I^{KL}(\mathbf{v},\mathbf{u}) = \sum_{x=70}^{84} v_x \ln \frac{v_x}{u_x}$, όπου $\mathbf{v}$ είναι το διάνυσμα των εξομαλυμένων τιμών και $\mathbf{u}$ είναι το διάνυσμα των αρχικών εκτιμήσεων, με τους παραπάνω περιορισμούς, ο Brockett (1991) βρίσκει τα εξομαλυμένα αποτελέσματα. Στον Πίνακα 5.4 φαίνονται τα εξομαλυμένα ποσοστά που προκύπτουν τόσο με τον περιορισμό της

ομαλότητας όσο και χωρίς αυτόν. Η γραφική τους παράσταση φαίνεται στο Σχήμα 5.4. Να σημειώσουμε ότι η εξομάλυνση έγινε με $M = 0.0002$, όπου M είναι η παράμετρος που εμφανίζεται στον περιορισμό της ομαλότητας. Η τιμή αυτή επιλέχθηκε έτσι ώστε να είναι κοντά στην τιμή του μέτρου ομαλότητας που προέκυψε από τη γραφική μέθοδο. Παρατηρούμε ότι καλύτερη από τις δυο εξομαλύνσεις είναι η δεύτερη, καθώς επιτυγχάνει τόσο καλύτερη ομαλότητα όσο και καλύτερη προσαρμογή.

Ηλικία	**Εξομαλυμένες τιμές χωρίς τον περιορισμό κυρτότητας**	**Εξομαλυμένες τιμές με όλους τους περιορισμούς**
70	0.068	0.068
71	0.068	0.068
72	0.071	0.072
73	0.077	0.076
74	0.082	0.080
75	0.081	0.084
76	0.086	0.088
77	0.097	0.093
78	0.105	0.103
79	0.115	0.118
80	0.131	0.134
81	0.151	0.155
82	0.179	0.181
83	0.211	0.207
84	0.248	0.244
	$S = 0.000344$	$S = 0.000211$
	$F = 20.1271$	$F = 18.5212$

Πίνακας 5.4: Εξομαλυμένα ποσοστά θνησιμότητας χωρίς και με τον περιορισμό κυρτότητας

Να σημειώσουμε ότι τα αποτελέσματα παρατίθενται όπως στο παράδειγμα του Brockett (1991). Σε μελλοντική διερεύνηση του θέματος, θα εξετάσουμε την ακρίβεια των υπολογισμών με χρήση κώδικα μη γραμμικού (κυρτού) προγραμματισμού.



Σχήμα 5.4: Γραφική παράσταση αρχικών εκτιμήσεων και εξομαλυμένων τιμών με τη μέθοδο της θεωρίας πληροφοριών, με και χωρίς τον περιορισμό κυρτότητας

Από τις παραπάνω εξομαλύνσεις, μπορούμε να βγάλουμε το συμπέρασμα ότι όλες οι μέθοδοι δίνουν περίπου τις ίδιες εξομαλυμένες τιμές καθώς και περίπου την ίδια τιμή για το μέτρο ομαλότητας.

5.4 Υπολογισμός ασφαλίστρου

Αναφέραμε στην εισαγωγή ότι ο λόγος που εξομαλύνουμε τα ποσοστά ή την ένταση θνησιμότητας, είναι για να εξασφαλίσουμε ότι θα μεταβάλλονται ομαλά και οι διάφορες ασφαλιστικές ποσότητες που υπολογίζονται με βάση τις τιμές αυτές (Miller, 1949). Ας δούμε αν αυτό συμβαίνει όντως στον υπολογισμό του ενιαίου καθαρού ασφαλίστρου στην *ασφάλιση μελλοντικού κεφαλαίου*, το οποίο αποτελεί και τη βάση υπολογισμού των ασφαλίστρων όλων των άλλων ασφαλίσεων ζωής.

Υποθέτουμε ότι έχουμε ένα άτομο ηλικίας x, το οποίο ασφαλίζεται στο πρόγραμμα μελλοντικού κεφαλαίου για διάρκεια ενός έτους και για κεφάλαιο μιας νομισματικής μονάδας. Δηλαδή το συγκεκριμένο άτομο θα εισπράξει μια νομισματική μονάδα μετά από ένα έτος (στην ηλικία $x+1$) εφόσον βέβαια βρίσκεται στη ζωή.

Η παρούσα αξία της παραπάνω ασφάλισης ισούται με το ασφαλισμένο κεφάλαιο της μιας νομισματικής μονάδας, το οποίο πληρώνεται με βεβαιότητα μετά από ένα έτος, πολλαπλασιασμένο με την πιθανότητα ότι αυτό θα πληρωθεί. Η πιθανότητα αυτή είναι στην ουσία η πιθανότητα το άτομο ηλικίας x να επιβιώσει μέχρι την ηλικία $x+1$, δηλαδή η p_x.

Γενικά στην περίπτωση που το ποσό πληρώνεται μετά από n έτη, το *ενιαίο καθαρό ασφάλιστρο* συμβολίζεται με ${}_nE_x$ και υπολογίζεται ως

$$ {}_nE_x = 1\cdot(1+i)^{-n}\,{}_np_x = 1\cdot(1+i)^{-n}(1-{}_nq_x)\,. $$

Όταν το $n=1$, έχουμε

$$ E_x = 1\cdot(1+i)^{-1}\,p_x = 1\cdot(1+i)^{-1}(1-q_x)\,. $$

Το i ονομάζεται *τεχνικό επιτόκιο* και συνήθως ισούται με 2.5% (ΦΕΚ, 847). Για περισσότερες πληροφορίες παραπέμπουμε στον Μπλέσιο (1998).

Στον Πίνακα 5.5 εμφανίζονται τα ασφάλιστρα που υπολογίστηκαν με βάση τα αδρά και τα εξομαλυμένα, με τη μέθοδο των Whittaker – Henderson με $h=4000$, ποσοστά. Αν υπολογίσουμε το μέτρο ομαλότητας για τις τιμές των ασφαλίστρων, θα πάρουμε $S=0.2271$ και $S'=0.00026$ για τα ασφάλιστρα που υπολογίστηκαν με τα αδρά ποσοστά και με τα εξομαλυμένα ποσοστά αντίστοιχα. Παρατηρούμε λοιπόν ότι η ομαλότητα των ποσοστών θνησιμότητας, διατηρείται κατά τον υπολογισμό των ασφαλίστρων. Μάλιστα αν συγκρίνουμε τις παραπάνω τιμές των μέτρων ομαλότητας με τις αντίστοιχες τιμές για τα ποσοστά θνησιμότητας, παρατηρούμε ότι αυτές δεν διαφέρουν πολύ. Σημειώνουμε ότι οι αντίστοιχες τιμές για τα ποσοστά ήταν $S=0.2386$ και $S'=0.00025$.

Ηλικία	Ασφάλιστρο (αδρά ποσοστά)	Ασφάλιστρο (εξομαλυμένα ποσοστά)
70	0.93268	0.92585
71	0.89366	0.91220
72	0.90634	0.90927
73	0.90146	0.91024
74	0.93659	0.90927
75	0.87415	0.90146
76	0.81951	0.89366
77	0.91902	0.88976
78	0.86829	0.88390
79	0.88488	0.87024
80	0.84000	0.85171
81	0.82537	0.82732
82	0.79707	0.80000
83	0.77463	0.77171
84	0.74244	0.74146

Πίνακας 5.5: Ενιαία καθαρά ασφάλιστρα, για την ασφάλιση μελλοντικού κεφαλαίου, υπολογισμένα με βάση τα αδρά και εξομαλυμένα ποσοστά θνησιμότητας αντίστοιχα

Κεφάλαιο 6

ΣΥΜΠΕΡΑΣΜΑΤΑ

6.1 Κριτική των μεθόδων εξομάλυνσης

Ύστερα από την παρουσίαση όλων των μεθόδων εξομάλυνσης πινάκων θνησιμότητας, γεννιέται το εύλογο ερώτημα ποιά από τις μεθόδους είναι η καλύτερη και συνεπώς ποιά πρέπει να χρησιμοποιείται. Σύμφωνα με τον London (1985), «καμία μέθοδος εξομάλυνσης δεν μπορεί να θεωρηθεί ως "καλύτερη" ή "πιο σωστή" από γενικής άποψης».

Επίσης αναφέρει ότι υπάρχουν κάποιοι παράγοντες, οι οποίοι μπορούν να καθοδηγήσουν ως προς το ποιά μέθοδος είναι κατάλληλη για χρήση. Πιο συγκεκριμένα, αν μεγαλύτερη έμφαση δίνεται στην ομαλότητα, μέθοδοι όπως η μπεϋζιανή, η μέθοδος των κινητών σταθμισμένων μέσων με ελαχιστοποίηση του R_0 και η γραφική μέθοδος, θα πρέπει να αποφεύγονται. Επιπλέον, οι παραμετρικές μέθοδοι δίνουν συγκεκριμένο βαθμό ομαλότητας, ο οποίος μπορεί να ελεγχθεί και να αυξομειωθεί από τον ερευνητή αν χρησιμοποιήσει μια μη παραμετρική μέθοδο.

Βασικός παράγοντας είναι επιπλέον, η μορφή και το εύρος των δεδομένων. Πολλές φορές τα αναλογιστικά δεδομένα δίνονται ομαδοποιημένα και συνήθως η μέθοδος της παρεμβολής ή κάποια άλλη παραμετρική μέθοδος χρησιμοποιείται για να υπολογίσουμε τις εξομαλυμένες τιμές στις ενδιάμεσες ηλικίες. Με άλλα λόγια, αν θέλουμε να πάρουμε εξομαλυμένες τιμές και για ενδιάμεσες ηλικίες από αυτές που δίνονται, τότε πρέπει απαραίτητα να χρησιμοποιήσουμε μια από τις παραμετρικές μεθόδους. Επίσης όταν τα δεδομένα είναι λίγα, δεν είναι κατάλληλες μέθοδοι, όπως αυτές των κινητών σταθμισμένων μέσων ή της παρεμβολής με κεντρικές διαφορές, οι οποίες δεν δίνουν εξομαλυμένες τιμές για κάποιον αριθμό από τις πρώτες και τελευταίες παρατηρήσεις. Σε περιπτώσεις λίγων δεδομένων, συνήθως χρησιμοποιούνται η γραφική μέθοδος και η μέθοδος με αναφορά σε τυπικό πίνακα θνησιμότητας. Σε αντίθετη περίπτωση καλό είναι να χρησιμοποιείται μια από τις παραμετρικές μεθόδους εξομάλυνσης.

Ως ένας άλλος παράγοντας, μπορεί να θεωρηθεί η πολυπλοκότητα των υπολογιστικών πράξεων κάθε μεθόδου. Μια τέτοια μέθοδος είναι η μπεϋζιανή ενώ η γραφική μέθοδος

απαιτεί μεγάλη εμπειρία από τον ερευνητή. Μια εύκολη στην εφαρμογή της μέθοδος είναι αυτή με αναφορά σε τυπικό πίνακα θνησιμότητας καθώς οι παράμετροι που πρέπει να εκτιμηθούν είναι λίγες. Βέβαια, με τη χρήση του ηλεκτρονικού υπολογιστή είναι εύκολη η εφαρμογή σχεδόν όλων των μεθόδων.

Ένα πρόβλημα που ίσως να αντιμετωπίσει ο ερευνητής, είναι ο προσδιορισμός των παραμέτρων που εμπλέκονται στις μεθόδους. Στις παραμετρικές μεθόδους, δεν υπάρχει πρόβλημα με τις παραμέτρους των μοντέλων καθώς αυτές προσδιορίζονται από τα ίδια τα δεδομένα. Βέβαια τα μοντέλα που χρησιμοποιούνται στη μέθοδο των splines είναι συνήθως μικρότερα, όσον αφορά τις παραμέτρους, από αυτά της μεθόδου μέσω των μοντέλων θνησιμότητας. Επίσης, οι παραμετρικές μέθοδοι εγγυώνται την ομαλότητα των αποτελεσμάτων. Έτσι το πρόβλημα με αυτές τις μεθόδους είναι τί μορφής θα είναι η συνάρτηση που θα χρησιμοποιηθεί, καθώς υπάρχει πληθώρα μοντέλων και πολλά από αυτά δεν είναι ικανοποιητικά για όλο το εύρος των ηλικιών, και ίσως η επιλογή της μεθόδου υπολογισμού των παραμέτρων (London, 1985).

Αντίθετα, οι μη παραμετρικές μέθοδοι έχουν το πλεονέκτημα ότι οδηγούνται από τα δεδομένα (*data driven*) και έτσι δεν υπάρχει το πρόβλημα να μην ισχύουν κάποιες υποθέσεις που προϋποθέτουν τα παραμετρικά μοντέλα (Wang, 1998). Επίσης οι μέθοδοι αυτές, μπορεί να μην εξασφαλίζουν την ομαλότητα σε βαθμό ανάλογο των παραμετρικών μεθόδων, δίνουν όμως στον ερευνητή τη δυνατότητα να επιτύχει τον επιθυμητό βαθμό ομαλότητας αλλάζοντας τις τιμές των παραμέτρων.

Όσον αφορά τη μπεϋζιανή μέθοδο εξομάλυνσης, αυτή είναι σύμφωνα με τον London (1985) η πιο αντιφατική εξαιτίας του προβλήματος ορισμού και επιλογής των παραμέτρων που εμπλέκονται σε αυτή. Εξαιτίας της πολυπλοκότητας ως προς τους μαθηματικούς υπολογισμούς, είναι μη πρακτικό ο ερευνητής να κάνει πολλές εξομαλύνσεις των ίδιων δεδομένων έτσι ώστε να επιλέξει την καλύτερη εξομάλυνση. Έτσι απαιτείται σε μεγάλο βαθμό η προσωπική κρίση του ερευνητή στην επιλογή των παραμέτρων πράγμα που οδηγεί σε μεροληψία. Να τονίσουμε στο σημείο αυτό ότι η υποκειμενικότητα του ερευνητή είναι χαρακτηριστικό που εμπλέκεται σε όλες τις μεθόδους εξομάλυνσης καθώς αυτή εκφράζει την εκ των προτέρων άποψη για τα δεδομένα που εξομαλύνονται. Άλλωστε αν δεν υπήρχε εκ των προτέρων άποψη, τότε οι αρχικές εκτιμήσεις θα ήταν οι καλύτερες εκτιμήσεις για τα πραγματικά δεδομένα και δεν θα χρειαζόταν η εξομάλυνση.

Αξίζει επίσης να αναφερθεί, ότι η μέθοδος εξομάλυνσης μέσω της θεωρίας πληροφοριών, παρουσιάζει ένα σοβαρό πρόβλημα όσον αφορά τη θεωρία της. Συγκεκριμένα, το πρόβλημα

έγκειται στο ότι τα μέτρα **u** και **v** που χρησιμοποιούνται στο μέτρο των Kullback – Leibler δεν είναι μέτρα πιθανότητας, όπως θα έπρεπε, καθώς δεν αθροίζουν στη μονάδα αλλά την υπερβαίνουν. Σύμφωνα με τους Mathai and Rathie (1975) ορισμένα μέτρα πληροφορίας μπορούν να χρησιμοποιηθούν ακόμα και αν δεν εμπλέκουν πιθανότητες, στην περίπτωση όμως που το άθροισμά τους δεν υπερβαίνει τη μονάδα. Ο Brockett (1991), όπως είδαμε, δίνει μια ασαφή εξήγηση για το λόγο που μπορεί να χρησιμοποιηθεί το μέτρο των Kullback – Leibler. Πέρα όμως από αυτό το πρόβλημα, η μέθοδος δίνει αποτελέσματα κοντά σε αυτά των άλλων μεθόδων.

Για την αντιμετώπιση του παραπάνω προβλήματος, προτείνουμε δυο λύσεις:

i) Η πρώτη λύση είναι η τυποποίηση των ασφαλιστικών ποσοτήτων που θέλουμε να εξομαλύνουμε. Δηλαδή να διαιρέσουμε κάθε τιμή u_x με το άθροισμά τους, $\sum_{x=1}^{n} u_x$, έτσι ώστε $\sum_{x=1}^{n} u'_x = 1$, όπου $u'_x = u_x \frac{1}{\sum_{x=1}^{n} u_x}$. Τότε το διάνυσμα $\mathbf{u}^T = (u_1,...,u_n)$, όπου το T δηλώνει την αντιστροφή του διανύσματος, αποτελεί μέτρο πιθανότητας και μπορεί να χρησιμοποιηθεί στο μέτρο των Kullback – Leibler.

ii) Η δεύτερη λύση είναι να χρησιμοποιήσουμε τη συνάρτηση πιθανότητας της τυχαίας μεταβλητής T που δηλώνει το χρόνο ζωής ενός ατόμου. Στην περίπτωση αυτή, γνωρίζουμε ότι ισχύει $f(x) = h(x)S(x)$. Έστω τώρα ότι η τυχαία μεταβλητή T είναι διακριτή με σύνολο τιμών το $R_T = \{x_0, x_1, x_2, ...\}$, τα οποία x_j, $j = 0, 1, 2, ...$ είναι διατεταγμένα κατ' αύξουσα σειρά. Αν συμβολίσουμε τη συνάρτηση πιθανότητας $f(x)$ με f_x, τη συνάρτηση κινδύνου $h(x)$ με q_x και τη συνάρτηση επιβίωσης $S(x)$ με S_x, τότε ισχύει $f_x = S_x q_x$. Έτσι το μέτρο των Kullback – Leibler μπορεί να χρησιμοποιηθεί χωρίς πρόβλημα καθώς το διάνυσμα $\mathbf{f}^T = (f_1,...,f_n)$, όπου το T δηλώνει την αντιστροφή του διανύσματος, είναι μέτρο πιθανότητας.

Γνωρίζοντας τις πιθανότητες θανάτου q_x για κάθε ηλικία x, μπορούμε να υπολογίσουμε τη συνάρτηση επιβίωσης ως $S_x = \prod_{x=0}^{n} (1 - q_{x-1})$ με $S_0 = 1$ και συνεπώς η συνάρτηση πιθανότητας ισούται με $f_x = S_x q_x$. Μετά την εξομάλυνση, και αφού πάρουμε τις εξομαλυμένες τιμές της συνάρτησης πιθανότητας, έστω g_x, ισχύει $S'_x = S'_{x-1} - g_{x-1}$, όπου S'_x

είναι οι εξομαλυμένες τιμές της συνάρτησης επιβίωσης, με $S_0' = 1$, και οι εξομαλυμένες τιμές για τα ποσοστά θνησιμότητας υπολογίζονται ως $q_x' = \frac{g_x}{S_x'}$, $x = 1, 2, ..., n$.

Ο London (1985), χωρίς να θέλει επίσημα να προτείνει κάποια από τις μεθόδους εξομάλυνσης, τονίζει την ευστροφία (*versatility*) της μεθόδου Whittaker – Henderson. Επιπλέον, όπως έχουμε ήδη αναφέρει, οι Guerrero et al. (2000) απέδειξαν ότι ο τύπος των Whittaker and Henderson, αποτελεί τον καλύτερο γραμμικό αμερόληπτο εκτιμητή (BLUE) για τα πραγματικά ποσοστά θνησιμότητας.

6.2 Ερωτήματα για την εξομάλυνση

Στο κεφάλαιο 1 και στην παράγραφο που ορίσαμε την εξομάλυνση, τέθηκε το ερώτημα αν οι μέθοδοι εξομάλυνσης μπορούν να χρησιμοποιηθούν σε οποιαδήποτε σειρά παρατηρούμενων τιμών. Η απάντηση που δόθηκε εκεί, ήταν ότι η εξομάλυνση μπορεί να γίνει μόνο στις σειρές για τις οποίες πιστεύεται ότι τα στοιχεία τους είναι συσχετισμένα μεταξύ τους (London, 1985).

Ένα δεύτερο ερώτημα που μπορεί να τεθεί, είναι αν έχοντας στη διάθεσή μας έναν πίνακα θνησιμότητας κατά πόσο είναι εύκολο να διαπιστώσουμε αν αυτός είναι εξομαλυμένος. Συγκεκριμένη απάντηση στη βιβλιογραφία δεν υπάρχει. Στη συνέχεια θα προσπαθήσουμε να δώσουμε δυο πιθανούς τρόπους οι οποίοι στηρίζονται απλά στη λογική.

i) Ο πρώτος τρόπος είναι – στην περίπτωση που έχουμε στη διάθεσή μας τον αριθμό των θανάτων d_x και τον αριθμό των ατόμων σε κίνδυνο l_x στην ηλικία x, $x = 1, 2, ..., n$ – να υπολογίσουμε τον αδρό δείκτη θνησιμότητας $u_x = \frac{d_x}{l_x}$, $x = 1, 2, ..., n$. Αν οι τιμές αυτές είναι διαφορετικές από τις αντίστοιχες τιμές του πίνακα θνησιμότητας τότε ο πίνακας είναι εξομαλυμένος. Να σημειώσουμε όμως στο σημείο αυτό ότι η εύρεση των πραγματικών (παρατηρούμενων) αριθμών θανάτων και ατόμων σε κίνδυνο είναι δύσκολη έως και αδύνατη ενώ οι δημοσιευμένοι πίνακες θνησιμότητας δίνουν τους αντίστοιχους εξομαλυμένους αριθμούς.

ii) Ο δεύτερος τρόπος είναι να παραστήσουμε γραφικά τα δεδομένα. Αν παράγεται μια ομαλή καμπύλη τότε ο πίνακας θνησιμότητας είναι εξομαλυμένος. Επίσης μπορούμε να υπολογίσουμε το μέτρο ομαλότητας $S = \sum_{x=1}^{n-3} \left(\Delta^3 v_x\right)^2$, όπου v_x είναι οι τιμές του πίνακα. Αν το παραπάνω μέτρο παίρνει πολύ μικρή τιμή τότε υποθέτουμε ότι ο πίνακας έχει υποστεί εξομάλυνση. Βέβαια η έννοια «μικρή τιμή» είναι σχετική καθώς δεν υπάρχει κάποιο συγκεκριμένο όριο γι' αυτή.

Ένα άλλο ερώτημα είναι αν οι μέθοδοι εξομάλυνσης που περιγράψαμε για τους πίνακες θνησιμότητας, μπορούν να χρησιμοποιηθούν σε δεδομένα όπως είναι τα βιοστατιστικά και αυτά της ανάλυσης επιβίωσης, στα οποία επιπλέον υπεισέρχεται η έννοια της λογοκρισίας. Σύμφωνα με την απάντηση που δώσαμε στο πρώτο ερώτημα, σαφώς και μπορούν να χρησιμοποιηθούν οι μέθοδοι αυτές. Ο Wang (1998) αναφέρει δυο τρόπους εξομάλυνσης της συνάρτησης κινδύνου στην περίπτωση της τυχαίας λογοκρισίας. Ο πρώτος είναι μέσω εκτιμητών πυρήνα και ο δεύτερος με χρήση συναρτήσεων splines. Και στις δυο περιπτώσεις, εξομαλύνονται οι αρχικές εκτιμήσεις της αθροιστικής συνάρτησης κινδύνου, χρησιμοποιώντας τις εκτιμήσεις Nelson – Aalen. Για περισσότερες πληροφορίες, παραπέμπουμε στους Anderson, Borgan, Gill and Keiding (1993).

Βέβαια στην περίπτωση των παραπάνω δεδομένων μπαίνει το δίλημμα κατά πόσο είναι θεμιτό να αλλάξουμε τις εκτιμήσεις που έχουμε υπολογίσει. Στην περίπτωση των πινάκων θνησιμότητας, ο λόγος για τον οποίο γίνεται η εξομάλυνση είναι για να είναι ομαλές οι διάφορες ασφαλιστικές ποσότητες που υπολογίζονται βάσει των πινάκων. Κάτι τέτοιο όμως δεν ισχύει για τα βιοστατιστικά δεδομένα ή τα δεδομένα ανάλυσης επιβίωσης. Επίσης να σημειώσουμε ότι οι εκτιμητές Kaplan – Meier και Nelson – Aalen δίνουν ομαλά αποτελέσματα για τη συνάρτηση κινδύνου.

Η δεύτερη προσέγγιση που είδαμε στην προηγούμενη ενότητα για την επίλυση του προβλήματος που υπάρχει στη μέθοδο της θεωρίας πληροφοριών, γεννά το ερώτημα αν γνωρίζοντας τις τιμές των πιθανοτήτων θανάτου q_x ή της έντασης θνησιμότητας m_x, μπορούμε να υπολογίσουμε τη συνάρτηση επιβίωσης S_x και το αντίστροφο. Όσον αφορά το πρώτο, η απάντηση έχει ήδη δοθεί στην προηγούμενη ενότητα ενώ για το δεύτερο ισχύουν τα παρακάτω:

Όταν η τυχαία μεταβλητή T, που συμβολίζει το χρόνο ζωής ενός ατόμου, είναι διακριτή, η συνάρτηση επιβίωσης S_x υπολογίζεται ως $S_x = \sum_{j:x_j>x} f_{x_j}$, $x \geq 0$ με $S_0 = 1$, όπου f_x είναι η συνάρτηση πιθανότητας της T. Άρα από την προηγούμενη σχέση ισχύει $f_x = S_x - S_{x-1}$. Επιπλέον από τη σχέση $f_x = S_x q_x$ παίρνουμε $q_x = \frac{f_x}{S_x}$, οπότε γνωρίζοντας τη συνάρτηση επιβίωσης μπορούμε να υπολογίσουμε την πιθανότητα θανάτου για κάθε ηλικία.

Ένα συνεπακόλουθο ερώτημα είναι αν η ομαλότητα διατηρείται για τις τιμές μιας ποσότητας, η οποία υπολογίζεται με βάση τα εξομαλυμένα δεδομένα. Για παράδειγμα, θέλουμε να δούμε αν εξομαλύνοντας τα ποσοστά θνησιμότητας και στη συνέχεια υπολογίσουμε τη συνάρτηση επιβίωσης, οι τιμές της θα είναι ομαλές.

Ξεκινάμε, υπολογίζοντας τη συνάρτηση επιβίωσης για τα δεδομένα του κεφαλαίου 5. Στον Πίνακα 6.1 φαίνονται οι τιμές των συναρτήσεων επιβίωσης S_x και S'_x που βασίζονται στα αρχικά και στα εξομαλυμένα με τη μέθοδο των Whittaker – Henderson με $h = 4000$ ποσοστά θνησιμότητας αντίστοιχα, καθώς και ο βαθμός ομαλότητας τους.

Ηλικία	S_x	S'_x
70	1.000	1.000
71	0.956	0.949
72	0.876	0.887
73	0.814	0.827
74	0.752	0.772
75	0.722	0.719
76	0.647	0.664
77	0.543	0.609
78	0.512	0.555
79	0.455	0.503
80	0.413	0.449
81	0.356	0.392
82	0.301	0.332
83	0.246	0.272
84	0.195	0.215
	$S = 0.033$	$S = 0.00023$

Πίνακας 6.1: Συναρτήσεις επιβίωσης για τα αρχικά και εξομαλυμένα, με τη μέθοδο των Whittaker – Henderson με $h = 4000$, ποσοστά θνησιμότητας αντίστοιχα

Από τα παραπάνω αποτελέσματα, παρατηρούμε ότι πραγματικά οι τιμές της συνάρτησης επιβίωσης που υπολογίστηκαν με βάση τα εξομαλυμένα ποσοστά θνησιμότητας, είναι ομαλές. Πιο συγκεκριμένα, η ομαλότητα είναι καλύτερη από αυτή των εξομαλυμένων ποσοστών θνησιμότητας αφού η μέθοδος των Whittaker – Henderson με $h = 4000$ δίνει τιμή για το μέτρο ομαλότητας ίση με $S = 0.000253768$. Επίσης και η ομαλότητα των τιμών της συνάρτησης επιβίωσης για τα αρχικά δεδομένα είναι κατά πολύ βελτιωμένη αφού το μέτρο ομαλότητας των αρχικών δεδομένων ήταν $S = 0.238581$.

Από το παραπάνω παράδειγμα λοιπόν, βγάζουμε το συμπέρασμα ότι για τη συνάρτηση επιβίωσης που υπολογίστηκε με βάση τα αρχικά (μη εξομαλυμένα) δεδομένα, η ομαλότητα βελτιώνεται ενώ για τη συνάρτηση επιβίωσης που στηρίχθηκε στα εξομαλυμένα δεδομένα, η ομαλότητα παρέμεινε περίπου σταθερή.

6.3 Επίλογος

Σκοπός αυτής της διπλωματικής εργασίας ήταν γενικότερα να εξετάσουμε τις βασικές πτυχές της εξομάλυνσης και να επικεντρωθούμε ειδικότερα στο πώς αυτή επιτυγχάνεται σε αναλογιστικά δεδομένα, όπως είναι τα ποσοστά θνησιμότητας και η ένταση θνησιμότητας.

Ένας από τους στόχους της αναλογιστικής επιστήμης είναι η περιγραφή της θνησιμότητας του πληθυσμού. Για να επιτευχθεί αυτό, υπολογίζονται από πρωτογενή δεδομένα οι αδροί δείκτες θνησιμότητας που παρουσιάζονται σε πίνακες, οι οποίοι είναι γνωστοί ως πίνακες θνησιμότητας. Επειδή όμως οι αρχικές εκτιμήσεις συνήθως παρουσιάζουν μεγάλες διακυμάνσεις, συνηθίζεται να εξομαλύνονται με στατιστικούς και μη τρόπους. Έτσι, *εξομάλυνση* είναι η αναθεώρηση των αρχικών εκτιμήσεων των ποσοστών θνησιμότητας $\overset{0}{q}_x$ ή της έντασης θνησιμότητας $\overset{0}{m}_x$, έτσι ώστε να απαλειφθούν οι μεγάλες διακυμάνσεις που αυτές παρουσιάζουν και να προκύψουν με τον τρόπο αυτό καλύτερες εκτιμήσεις.

Αφού αναφέραμε γενικά στοιχεία και έννοιες που αφορούν τους πίνακες θνησιμότητας καθώς και την κατασκευή τους, ορίσαμε την εξομάλυνση και δώσαμε τα βασικά χαρακτηριστικά της, τα οποία είναι η ομαλότητα (*smoothness*) και η καλή προσαρμογή (*goodness of fit*). Επίσης, παρουσιάσαμε τεστ και κριτήρια που χρησιμοποιούνται για τον

έλεγχο του κατά πόσο είναι ικανοποιητική μια εξομάλυνση και τις κατηγορίες των μεθόδων εξομάλυνσης, οι οποίες διακρίνονται σε παραμετρικές και μη παραμετρικές μεθόδους.

Έγινε προσπάθεια να παρουσιασθούν όλες οι μέθοδοι εξομάλυνσης που υπάρχουν στη βιβλιογραφία. Στόχος μας βέβαια, δεν ήταν η λεπτομερής παρουσίαση των μεθόδων, γι' αυτό περιοριστήκαμε μόνο στη βασική ιδέα και σε κάποια μεθοδολογικά ζητήματά τους. Σε ξεχωριστό κεφάλαιο, παρουσιάστηκε η μέθοδος της θεωρίας πληροφοριών, καθώς αυτή έδωσε το έναυσμα να ασχοληθούμε με το θέμα της εξομάλυνσης.

Για πρακτικούς λόγους, εφαρμόσθηκαν σε δεδομένα ορισμένες μόνο από τις μεθόδους, για να δείξουμε ότι όλες δίνουν περίπου τα ίδια αποτελέσματα. Ο υπολογισμός του ενιαίου καθαρού ασφαλίστρου στην περίπτωση της ασφάλισης μελλοντικού κεφαλαίου, έγινε για να αποδείξουμε πρακτικά τα λόγια του Miller (1949), ότι ο αναλογιστής επιθυμεί τα ποσοστά ή η ένταση θνησιμότητας να μεταβάλλονται ομαλά, γιατί έτσι εξασφαλίζει ότι θα μεταβάλλονται ομαλά και οι διάφορες ασφαλιστικές ποσότητες που υπολογίζονται με βάση τις τιμές αυτές.

Στο τέλος, προσπαθήσαμε να κάνουμε μια κριτική και σύγκριση των μεθόδων εξομάλυνσης, καταλήγοντας στο συμπέρασμα ότι δεν μπορούμε με βεβαιότητα να προτείνουμε κάποια από αυτές ως καλύτερη. Όλες έχουν πλεονεκτήματα και μειονεκτήματα και έγκειται στον ερευνητή ποιά μέθοδο θα χρησιμοποιήσει. Επίσης αναφέραμε κάποια ανοικτά προβλήματα για την εξομάλυνση και κάναμε κάποιες σκέψεις για την επίλυσή τους.

Κλείνοντας τη διπλωματική εργασία, να σημειώσουμε ότι κατά τη διάρκεια της μελέτης και της συλλογής των στοιχείων, αντιμετωπίσαμε κάποια προβλήματα, όπως για παράδειγμα, ότι κάποιες έννοιες δεν έχουν μονοσήμαντο ορισμό. Τέτοιες έννοιες είναι οι πίνακες επιβίωσης (*life tables*) και η ομαλότητα. Επίσης κάποιοι συγγραφείς διατύπωναν προτάσεις και ισχυρισμούς, που κατά τη γνώμη μας δεν είναι τόσο ξεκάθαροι, όπως για παράδειγμα οι Benjamin and Pollard (1980) στη γραφική μέθοδο εξομάλυνσης.

ΠΑΡΑΡΤΗΜΑ

Α. Πίνακας Θνησιμότητας 1990 Ανδρών

Β. Πίνακας Θνησιμότητας 1990 Γυναικών

Οι πίνακες αυτοί έχουν δημιουργηθεί από την Εθνική Στατιστική Υπηρεσία, αναφέρονται σε στοιχεία 1990 και έχουν εξομαλυνθεί από την Ένωση Αναλογιστών Ελλάδος.

Α. Πίνακας Θνησιμότητας 1990 Ανδρών

x	l_x	d_x	q_x	$\overset{0}{e}_x$
0	1000000	10070	0.010070	74.63
1	989930	490	0.000495	74.38
2	989440	389	0.000393	73.42
3	989051	316	0.000319	72.45
4	988735	279	0.000282	71.47
5	988456	253	0.000256	70.49
6	988203	230	0.000233	69.51
7	987973	206	0.000209	68.52
8	987767	186	0.000188	67.54
9	987581	173	0.000175	66.55
10	987408	174	0.000176	65.56
11	987234	191	0.000193	64.57
12	987043	222	0.000225	63.58
13	986821	269	0.000273	62.60
14	986552	342	0.000347	61.62
15	986210	448	0.000454	60.64
16	985762	581	0.000589	59.66
17	985181	730	0.000741	58.70
18	984451	877	0.000891	57.74
19	983574	1008	0.001025	56.79
20	982566	1109	0.001129	55.85
21	981457	1174	0.001196	54.91
22	980283	1202	0.001226	53.98
23	979081	1196	0.001222	53.04
24	977885	1166	0.001192	52.11
25	976719	1121	0.001148	51.17
26	975598	1074	0.001101	50.23
27	974524	1037	0.001064	49.28
28	973487	1013	0.001041	48.34
29	972474	1007	0.001036	47.39
30	971467	1016	0.001046	46.43
31	970451	1039	0.001071	45.48
32	969412	1076	0.001110	44.53
33	968336	1123	0.001160	43.58
34	967213	1181	0.001221	42.63
35	966032	1250	0.001294	41.68
36	964782	1330	0.001379	40.73
37	963452	1425	0.001479	39.79
38	962027	1535	0.001596	38.85
39	960492	1662	0.001730	37.91
40	958830	1799	0.001876	36.97
41	957031	1939	0.002026	36.04
42	955092	2077	0.002175	35.11
43	953015	2214	0.002323	34.19
44	950801	2354	0.002476	33.27
45	948447	2510	0.002646	32.35
46	945937	2692	0.002846	31.43
47	943245	2912	0.003087	30.52
48	940333	3176	0.003378	29.62
49	937157	3489	0.003723	28.71
50	933668	3847	0.004120	27.82
51	929821	4250	0.004571	26.93
52	925571	4697	0.005075	26.05
53	920874	5189	0.005635	25.18
54	915685	5730	0.006258	24.32

x	l_x	d_x	q_x	$\overset{0}{e}_x$
55	909955	6324	0.006950	23.47
56	903631	6975	0.007719	22.63
57	896656	7688	0.008574	21.81
58	888968	8467	0.009525	20.99
59	880501	9312	0.010576	20.19
60	871189	10223	0.011735	19.40
61	860966	11193	0.013001	18.62
62	849773	12213	0.014372	17.86
63	837560	13262	0.015834	17.12
64	824298	14338	0.017394	16.38
65	809960	15499	0.019136	15.66
66	794461	16782	0.021124	14.96
67	777679	18164	0.023357	14.27
68	759515	19608	0.025816	13.60
69	739907	21095	0.028510	12.95
70	718812	22613	0.031459	12.31
71	696199	24149	0.034687	11.70
72	672050	25684	0.038217	11.10
73	646366	27198	0.042078	10.52
74	619168	28668	0.046301	9.96
75	590500	30065	0.050914	9.42
76	560435	31359	0.055955	8.90
77	529076	32515	0.061456	8.40
78	496561	33499	0.067462	7.91
79	463062	34271	0.074010	7.45
80	428791	34794	0.081144	7.01
81	393997	35033	0.088917	6.58
82	358964	34953	0.097372	6.17
83	324011	34529	0.106567	5.78
84	289482	33740	0.116553	5.42
85	255742	32578	0.127386	5.06
86	223164	31049	0.139131	4.73
87	192115	29171	0.151841	4.41
88	162944	26980	0.165578	4.11
89	135964	24529	0.180408	3.83
90	111435	21884	0.196384	3.56
91	89551	19125	0.213565	3.31
92	70426	16339	0.232002	3.08
93	54087	13616	0.251743	2.86
94	40471	11042	0.272837	2.65
95	29429	8691	0.295321	2.45
96	20738	6619	0.319173	2.27
97	14119	4863	0.344429	2.10
98	9256	3435	0.371111	1.95
99	5821	2323	0.399072	1.80
100	3498	1498	0.428245	1.67
101	2000	917	0.458500	1.54
102	1083	530	0.489381	1.42
103	553	288	0.520796	1.31
104	265	148	0.558491	1.18
105	117	71	0.606838	1.05
106	46	31	0.673913	0.89
107	15	12	0.800000	0.70
108	3	3	1.000000	0.50

Β. Πίνακας Θνησιμότητας 1990 Γυναικών

x	l_x	d_x	q_x	$\overset{0}{e}_x$
0	1000000	9540	0.009540	79.47
1	990460	364	0.000368	79.24
2	990096	239	0.000241	78.26
3	989857	209	0.000211	77.28
4	989648	187	0.000189	76.30
5	989461	175	0.000177	75.31
6	989286	169	0.000171	74.33
7	989117	164	0.000166	73.34
8	988953	158	0.000160	72.35
9	988795	149	0.000151	71.36
10	988646	138	0.000140	70.37
11	988508	128	0.000129	69.38
12	988380	121	0.000122	68.39
13	988259	124	0.000125	67.40
14	988135	139	0.000141	66.41
15	987996	167	0.000169	65.42
16	987829	203	0.000206	64.43
17	987626	244	0.000247	63.44
18	987382	285	0.000289	62.46
19	987097	322	0.000326	61.48
20	986775	352	0.000357	60.50
21	986423	375	0.000380	59.52
22	986048	391	0.000397	58.54
23	985657	404	0.000410	57.56
24	985253	416	0.000422	56.59
25	984837	426	0.000433	55.61
26	984411	439	0.000446	54.63
27	983972	451	0.000458	53.66
28	983521	462	0.000470	52.68
29	983059	472	0.000480	51.71
30	982587	480	0.000489	50.73
31	982107	487	0.000496	49.76
32	981620	496	0.000505	48.78
33	981124	507	0.000517	47.80
34	980617	523	0.000533	46.83
35	980094	545	0.000556	45.85
36	979549	575	0.000587	44.88
37	978974	615	0.000628	43.91
38	978359	664	0.000679	42.93
39	977695	725	0.000742	41.96
40	976970	800	0.000819	40.99
41	976170	887	0.000909	40.03
42	975283	986	0.001011	39.06
43	974297	1095	0.001124	38.10
44	973202	1212	0.001245	37.14
45	971990	1331	0.001369	36.19
46	970659	1450	0.001494	35.24
47	969209	1569	0.001619	34.29
48	967640	1689	0.001745	33.34
49	965951	1815	0.001879	32.40
50	964136	1952	0.002025	31.46
51	962184	2105	0.002188	30.52
52	960079	2282	0.002377	29.59
53	957797	2488	0.002598	28.66
54	955309	2728	0.002856	27.73
55	952581	3006	0.003156	26.81
56	949575	3326	0.003503	25.89
57	946249	3694	0.003904	24.98
58	942555	4112	0.004363	24.08
59	938443	4587	0.004888	23.18
60	933856	5100	0.005461	22.29
61	928756	5630	0.006062	21.41
62	923126	6164	0.006677	20.54
63	916962	6705	0.007312	19.68
64	910257	7313	0.008034	18.82
65	902944	8108	0.008980	17.97
66	894836	9144	0.010219	17.12
67	885692	10419	0.011764	16.30
68	875273	11891	0.013585	15.48
69	863382	13560	0.015706	14.69
70	849822	15424	0.018150	13.92
71	834398	17476	0.020944	13.16
72	816922	19700	0.024115	12.44
73	797222	22077	0.027692	11.73
74	775145	24580	0.031710	11.05
75	750565	27169	0.036198	10.40
76	723396	29799	0.041193	9.77
77	693597	32415	0.046735	9.17
78	661182	34950	0.052860	8.59
79	626232	37330	0.059610	8.04
80	588902	39475	0.067032	7.52
81	549427	41298	0.075166	7.02
82	508129	42714	0.084061	6.55
83	465415	43640	0.093766	6.11
84	421775	44004	0.104331	5.69
85	377771	43746	0.115800	5.30
86	334025	42832	0.128230	4.92
87	291193	41253	0.141669	4.57
88	249940	39031	0.156161	4.25
89	210909	36226	0.171761	3.94
90	174683	32928	0.188501	3.65
91	141755	29263	0.206434	3.39
92	112492	25377	0.225589	3.14
93	87115	21430	0.245997	2.90
94	65685	17582	0.267671	2.69
95	48103	13980	0.290626	2.49
96	34123	10744	0.314861	2.30
97	23379	7957	0.340348	2.13
98	15422	5661	0.367073	1.97
99	9761	3855	0.394939	1.82
100	5906	2503	0.423806	1.69
101	3403	1544	0.453717	1.56
102	1859	900	0.484131	1.44
103	959	494	0.515120	1.33
104	465	254	0.546237	1.20
105	211	126	0.597156	1.05
106	85	59	0.694118	0.86
107	26	21	0.807692	0.69
108	5	5	1.000000	0.50

ΒΙΒΛΙΟΓΡΑΦΙΑ

Α. ΕΛΛΗΝΙΚΗ

Εφημερίδα της Κυβερνήσεως της Ελληνικής Δημοκρατίας (4 Ιουλίου 2001), Τεύχος Δεύτερο, Αρ. Φύλλου 847

Μπλέσιος, Ν. (1998). *Μαθηματικά Ασφαλίσεων Ζωής*, Εκδοτικές Επιχειρήσεις «Το Οικονομικό», Αθήνα.

Παπαϊωάννου, Τ. (2000). *Εισαγωγή στις Πιθανότητες*, Εκδόσεις Αθ. Σταμούλης, Αθήνα.

Παπαϊωάννου, Τ. και Φερεντίνος, Κ. (2000). *Μαθηματική Στατιστική*, Εκδόσεις Αθ. Σταμούλης, Αθήνα.

Β. ΞΕΝΗ

Agresti, A. (2002). *Categorical Data Analysis*, Second Edition, Wiley – Interscience, New Jersey.

Anderson, P. K, Borgan ∅., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer Verlag, New York.

Benjamin, P. and Pollard, J.H. (1980). *The Analysis of Mortality and Other Actuarial Statistics*, Heinemann, London.

Brockett, P.L. (1991). Information Theoretic Approach to Actuarial Science: A Unification and Extension of Relevant Theory and Applications, *TSA* 43, 73 – 114.

Brockett, P.L, Charnes, A. and Cooper, W. (1980). MDI Estimation via Unconstrained Convex Programming, *Communications in Statistics*, IX, Ser. B, No. 3, 223 – 234.

Brockett, P.L. and Cox, S. (1984). Statistical Adjustment of Mortality Tables to Reflect Known Information, *TSA* 36, 63 – 71.

Brockett, P.L., Cox, S., Golang, B., Phillips, F. and Song, Y. (1995). Actuarial Usage of Grouped Data: An Approach to Incorporating Secondary Data, *TSA* 47, 89 – 113.

Brockett, P.L. and Zhang, J. (1986). Information Theoretical Mortality Graduation, *Scandinavian Actuarial Journal*, 131 – 140.

Broffit, J.D. (1996). On Smoothness Terms in Multidimensional Whittaker Graduation, *Insurance: Mathematics and Economics*, 18, 13 – 27.

Copas, J. and Haberman, S. (1983). Non – Parametric Graduation Using Kernel Methods, *Journal of the Institute of Actuaries*, 110, 135 – 156.

Cox, D.R and Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall/CRC.

Ferentinos, K. and Papaioannou, T. (1981). New Parametric Measures of Information, *Information and Control* 51, 193 – 208.

Gavrilov, L. and Gavrilova, N.S. (1991). *The Biology of Life Span: A Quantitative Approach*, Academician V.P. Skulachev.

Gerber, H. (1997). *Life Insurance Mathematics*, Springer, New York.

Greville, T.N.E. (1983). *Graduation*, Encyclopedia of Statistical Sciences, 3, (Eds., S. Kotz and N.L. Johnson), 463 – 469, John Wiley & Sons, New York.

Guerrero, V., Juarez, R. and Poncela, P. (2001). Data Graduation Based on Statistical Time Series Methods, *Statistics & Probability Letters* 52, 169 – 175.

Haberman, S. (1998). *Actuarial Methods*, Encyclopedia of Biostatistics, 1, (Eds., P. Armitage and Th. Colton), 37 – 49, John Wiley & Sons, New York.

Haberman, S. and Pitacco, E. (1999). *Actuarial Models for Disability Insurance*, Chapman & Hall/CRC.

Hatzopoulos, P. (1997). Statistical and Mathematical Modeling for Mortality Trends and the Comparison of Mortality Experiences through Generalised Linear Models and GLIM, *PhD Thesis*, The City University, London.

Heligman, L. and Pollard, J.H. (1980), The Age Pattern of Mortality, *Journal of the Institute of Actuaries*, 107, 49 – 80.

Johnson, R.C.E. and Johnson, N.L. (1980). *Survival Models and Data Analysis*, John Wiley & Sons, New York.

Kellison, S.G. (1975). *Fundamentals of Numerical Analysis*, Homewood: Richard D. Irwin, Inc.

Kimeldorf, G.S. and Jones, D.A. (1967). Bayesian Graduation, *TSA*, XIX, 66.

Kostaki, A. (1992). A nine – Parameter Version of the Heligman – Pollard Formula, *Mathematical Population Studies*, 3(4), 277 – 288.

Kullback, S. (1959). *Information Theory and Statistics*, John Wiley & Sons, New York.

London, D. (1985). *Graduation: The Revision of Estimates*, ACTEX Publications, Winsted, Connecticut.

Lytrokapi, A. (1998). PARAMETRIC MODELS OF MORTALITY: Their Use in Demography and Actuarial Science, *Master Thesis*, Department of Statistics, Athens University of Economics and Business.

Mac Lane, S. (1986). *Mathematics Form and Function*, Springer – Verlag, New York.

Mathai, A.M and Rathie, P.N. (1975). *Basic Concepts in Information Theory and Statistics. Axiomatic Foundations and Applications*, Wiley Eastern Limited, New Delhi.

Miller, M.D. (1949). *Elements of Graduation*, Actuarial Society of America, New York.

Pagano, M. and Gauvreau, K. (2000). *Principles of Biostatistics*, Pacific Grove, CA: Duxbury.

Papaioannou, T. (1985). *Measures of Information*, Encyclopedia of Statistical Sciences, 5, (Eds., S. Kotz and N.L. Johnson), 391 – 497, John Wiley & Sons, New York.

Papaioannou, T. (2001). *On Distances and Measures of Information: A Case of Diversity, Probability and Statistical Models with Applications*, (Eds., C.A Charalambides, M.V. Koutras and N. Balakrishnan), 503 – 515, Chapman & Hall/CRC.

Papaioannou, T. and Ferentinos, K. (2002). Is Fisher's Information Number a Measure of Statistical Information?, *Technical Report*, Department of Statistics & Insurance Science, University of Piraeus.

Papaioannou, T. and Kempthorne, O. (1971). On Statistical Information Theory and Related Measures of Information, *Technical Report No ARL. 71-0059*, Aerospace Research Laboratories, Wright – Patterson A.F.B., Ohio.

Papaioannou, T. and Tsairidis, Ch. (2001). A Note on Defining Information in Random Censoring, *Technical Report*, Department of Mathematics, University of Ioannina.

Peristera, P. and Kostaki, A. (2004). Graduation of Mortality Data Through Kernel Estimates, *Presentation in 17th Statistical Conference*, 14 – 18 April 2004, Lefkada, Greece.

Renshaw, A.E., Haberman, S. and Hatzopoulos, P. (1996α). On the Duality of Assumptions Underpinning the Constructions of Life Tables, *ASTIN Bulletin 27*, 1, 1 – 18.

Renshaw, A.E., Haberman, S. and Hatzopoulos, P. (1996β). The Modelling of Mortality Trends in United Kingdom Male Assured Lives, *B.A.J. 2*, II, 449 – 477.

Renshaw, A.E. and Hatzopoulos, P. (1996). On the Graduation of Amounts, *B.A.J.* 2, I, 185 – 205.

Soofi, E.S. (1994). Capturing the Intangible Concept of Information, *Journal of the American Statistical Association 89*, 1243 – 1254.

Thomas, J. and Cover, T. (1991). *Elements of Information Theory*, John Wiley & Sons, New York.

Tsairidis, Ch., Zografos, K., Ferentinos, K. and Papaioannou, T. (2001). Information in Quantal Response Data and Random Censoring, *Ann. Inst. Statist. Math*, 53(3), 528 – 542.

Wang, J.L. (1998). *Smoothing Hazard Rates*, Encyclopedia of Biostatistics, 5, (Eds., P. Armitage and Th. Colton), 4140 – 4150, John Wiley & Sons, New York.

Weber, E.J. (1976). *Mathematical Analysis. Business and Economic Applications*, 4th Edition, Haprer and Row Publishers, New York.

Zhang, J. and Brockett, P.L. (1987). Quadratically Constrained Information Theoretic Analysis, *SIAM Journal of Applied Mathematics 47*, no 4, 871 – 885.