ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΙΑΣ

ΤΜΗΜΑ ΙΑΤΡΙΚΗΣ

ΠΜΣ «Μεθοδολογία Βιοϊατρικής Έρευνας Βιοστατιστική και Κλινική Βιοπληροφορική»

2017-2018

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΤΣΙΤΩΤΑ ΒΑΣΙΛΕΙΟΥ

ΣΥΓΚΡΙΣΗ ΣΤΑΤΙΣΤΙΚΩΝ ΜΕΘΟΔΩΝ ΓΙΑ ΤΗΝ ΑΝΙΧΝΕΥΣΗ ΣΥΝΔΕΣΗΣ 2Χ2

COMPARISON OF STATISTICAL METHODS IN DETECTING 2X2 ASSOCIATION

Τριμελής Επιτροπή Εξέτασης:

Μπατσίδης Απόστολος,

Επίκουρος Καθηγητής, Τμήμα Μαθηματικών, Πανεπιστήμιο Ιωαννίνων (*Επιβλέπων Καθηγητης*).

Στεφανίδης Ιωάννης,

Καθηγητής, Τμήμα Ιατρικής,Πανεπιστήμιο Θεσσαλίας.

Δοξάνη Χρυσούλα,

Επιστημονικός Συνεργάτης Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας.

Σεπτέμβριος 2018

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ΠΕΡΙΕΧΟΜΕΝΑ

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ΠΡΟΛΟΓΟΣ-ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να εκφράσω τις ευχαριστίες μου στα μέλη της Τριμελούς Επιτροπής κρίσης της παρούσας διατριβής, κ.κ, Στεφανίδη Ιωάννη, Καθηγητή, Κοσμήτορα της Σχολής Επιστημών Υγείας του Πανεπιστημίου Θεσσαλίας, Δοξάνη Χρυσούλα, Επιστημονικός Συνεργάτης Τμήματος Ιατρικής Πανεπιστημίου Θεσσαλίας και Μπατσίδη Απόστολο (Επιβλέπων), Επίκουρο Καθηγητή του Τμήματος Μαθηματικών του Πανεπιστημίου Ιωαννίνων για την αξιολόγηση της διπλωματικής διατριβής. Τέλος θα ήθελα να αφιερώσω την εργασία μου αυτή στους γονείς μου, οι οποίοι είναι οι ηθικοί μου αρωγοί σε κάθε μου προσπάθεια.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ΠΕΡΙΛΗΨΗ

Τη σημερινή εποχή αναγνωρίζεται η χρήση των μαθηματικών και της στατιστικής σε πολλούς τομείς επιστημών και ιδιαίτερα στον τομέα της Ιατρικής. Η παρούσα διπλωματική εργασία πραγματεύεται τις στατιστικές μεθόδους που χρησιμοποιούνται για την εύρεση της πιθανής σχέσης μεταξύ δυο ποιοτικών μεταβλητών και βρίσκουν εφαρμογή στην επιστήμη της Ιατρικής. Συγκεκριμένα το αντικείμενο αυτής της εργασίας είναι οι στατιστικές μέθοδοι: X^2 του Pearson, το ακριβές test του Fisher, η διόρθωση συνέχειας του Yates και η μέθοδος της Λογιστικής Παλινδρόμησης.

Η εργασία αυτή πραγματοποιήθηκε στα πλαίσια του μεταπτυχιακού προγράμματος: «Μεθοδολογία Βιοϊατρικής Έρευνας, Βιοστατιστική και Κλινική Βιοπληροφορική» του Τμήματος Ιατρικής του Πανεπιστημίου Θεσσαλίας και έχει ως σκοπό να κάνει φανερή τη χρησιμότητα αυτών των στατιστικών μεθόδων, ιδιαίτερα στον τομέα της Ιατρικής. Στο πρώτο κεφάλαιο γίνεται αναφορά στη στατιστική μέθοδο X^2. Συγκεκριμένα περιγράφονται οι στατιστικές μέθοδοι: X^2 του Pearson, το ακριβές test του Fisher και η διόρθωση συνέχειας του Yates. Αρχικά δίνεται η μορφή της στατιστικής συνάρτησης αυτών των μεθόδων και στη συνέχεια αναφερόμαστε στις υποθέσεις που πρέπει να ελεγχθούν, καθώς και στο πότε εφαρμόζονται αυτές οι μέθοδοι. Παράλληλα δίνονται παραδείγματα εφαρμογής αυτών των μεθόδων στην Ιατρική. Στο δεύτερο κεφάλαιο αναφερόμαστε στη μέθοδο της Λογιστικής Παλινδρόμησης και συγκεκριμένα παρουσιάζεται ο μαθηματικός τύπος που χρησιμοποιείται στη μέθοδο αυτή καθώς οι υποθέσεις και οι εφαρμογές αυτής.

Λέξεις κλειδιά: Στατιστικές μέθοδοι, ποιοτικές μεταβλητές, X^2 του Pearson, Test Fisher, Διόρθωση του Yates, Λογιστική Παλινδρόμηση.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ABSTRACT

Nowadays the use of mathematics and statistics is recognized in many sectors of science especially in medicine. This diplomatic thesis deals with the statistical methods that are being used in medicine in order to find a possible association between two quality variables. To be more specific, the subject of this thesis is the following statistical methods: Pearson's X^2, Fisher's exact test, Yates Continuity Correction and Logistic Regression.

This assignment was done under the postgraduate program: "Biostatistic research methodology, biostatistics and clinical bioinformatics" of the medical department of the University of Thessaly and aims to show how useful these statistical methods are, especially in medicine. At the first chapter is reported the statistical method X^2. Specifically the following statistical methods are described: X^2 of Pearson, Fisher's exact test and the correction of continuity of Yates. First of all the form of the statistical function of these methods are given and then we refer to the assumption that should be satisfied for these methods to be valid. Alongside, examples of implementation of these methods in medicine are given. At the second chapter, we refer to the subject of Logistic Regression. In this frame the mathematical relation that is used in this method is presented as well as the assumptions and its implementation.

Keywords: Statistical methods, qualitative variables, Pearson's X^2, Fisher's exact test, Correction of Yates, logistic regression

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ΕΙΣΑΓΩΓΗ

Τη σημερινή εποχή αναγνωρίζεται η χρήση της στατιστικής και διάφορων στατιστικών μεθόδων σε πολλούς τομείς επιστημών. Βασικές έννοιες της Στατιστικής έχουν εισχωρήσει και ενσωματωθεί στη: Φυσική, Χημεία, Βιολογία, Ιατρική, Μετεωρολογία, Γενετική, Αστρονομία, Γεωργία, Βιομηχανία, στη μελέτη του φυσικού περιβάλλοντος, στη μελέτη των ανθρωπίνων ιδεών και προθέσεων, στη θεωρία των αποφάσεων, στον έλεγχο ποιότητας των προϊόντων κ.λ.π. (Ζαχαροπούλου, Χ. 2001).

Σύμφωνα με τον πατέρα της σύγχρονης Στατιστικής Sir R. A. Fisher (1890-1962): «Στατιστική είναι ένα σύνολο αρχών και μεθοδολογιών για:

- το σχεδιασμό της διαδικασίας συλλογής δεδομένων (σχεδιασμός πειραμάτων-δειγματοληψία)
- τη συνοπτική και αποτελεσματική παρουσίασή τους (περιγραφική στατιστική)
- την ανάλυση και εξαγωγή αντίστοιχων συμπερασμάτων, για όλο το σύνολο ή την ικανότητα μιας διαδικασίας, κάτω από συνθήκες αβεβαιότητας (επαγωγική στατιστική ή στατιστική συμπερασματολογία) (Τσίπος, Στ. & Κωνσταντινίδης, Θ., 2010).

Η Στατιστική χρησιμοποιείται ευρέως στην Ιατρική έρευνα, μιας και η σύγχρονη ιατρική προσπαθεί να αξιολογήσει καινούριες απόψεις και θεραπείες μέσω της παρατήρησης και του πειράματος. Η πλειοψηφία των εργασιών που δημοσιεύονται σε ιατρικά περιοδικά περιέχουν σημαντική ποσότητα στατιστικού υλικού. Περισσότερο από 90% των δημοσιευμένων εργασιών σε έγκυρα βιοϊατρικά περιοδικά επικαλούνται βιοστατιστικές έννοιες ή εφαρμόζουν στατιστικές μεθόδους. Το σύνολο των ιατρικών γνώσεων οικοδομείται από τις επιστημονικές αυτές εργασίες που αποτελούν έτσι, τη βάση για την τεκμηριωμένη ιατρική (evidence based medicine), ανεξάρτητα από την επικέντρωση τους στον χώρο μιας συγκεκριμένης ιατρικής ειδικότητας (π.χ. παθολογίας, χειρουργικής) ή σε άλλο γνωστικό πεδίο των επιστημών υγείας (Τριχοπουλος Δ., Τζώνου Α. & Κατσουγιάννη Κ. 2001).

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

Η στατιστική στην Ιατρική έρευνα περιλαμβάνει:

- Σχεδιασμό πειραμάτων και μελετών.
- Συλλογή δεδομένων
- Επεξεργασία δεδομένων
- Ανάλυση δεδομένων
- Παρουσίαση δεδομένων και συμπερασμάτων
- Ερμηνεία των αποτελεσμάτων και παραγωγή νέας γνώσης

(Τριχοπουλος Δ., Τζώνου Α. & Κατσουγιάννη Κ. 2001).

Ένα βασικό θέμα κατά τη διαδικασία της μελέτης ενός φαινομένου (ή πειράματος) είναι αυτό του προσδιορισμού των μεταβλητών που θα αποτελέσουν το αντικείμενο έρευνας.

Για κάθε μεταβλητή που επιλέγεται για να μελετηθεί, πρέπει να υπάρχει δυνατότητα να της αποδοθούν τιμές. Η απόδοση τιμών σε μια μεταβλητή δε σημαίνει κατά ανάγκη ποσοτικοποίηση, δηλαδή, μέτρηση ή απαρίθμηση. Υπάρχουν μεταβλητές όπως το βάρος, η θερμοκρασία ή ο αριθμός των μελών μιας οικογένειας που χαρακτηρίζονται ποσοτικές και οι οποίες, πράγματι, είναι μετρήσιμες. Υπάρχουν όμως άλλες, όπως το άγχος, η στάση ως προς κάποιο θέμα, η οικογενειακή κατάσταση ή οι πολιτικές πεποιθήσεις που χαρακτηρίζονται ποιοτικές και στις οποίες μπορούν να αποδοθούν τιμές, όμως, οι τιμές αυτές δεν εκφράζουν κάτι το μετρήσιμο αλλά κατηγοριοποίηση ή διάταξη (Τσίπος Στ. & Κωνσταντινίδης Θ. 2010).

Όταν θέλουμε να μελετήσουμε δύο ποιοτικές μεταβλητές χρησιμοποιούνται συνήθως οι στατιστικές μέθοδοι: X^2 του Pearson, το ακριβές test του Fisher, η Διόρθωση συνέχειας του Yates και η μέθοδος της Λογιστικής Παλινδρόμησης.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ΚΕΦΑΛΑΙΟ 1^{Ο}

ΣΤΑΤΙΣΤΙΚΗ ΜΕΘΟΔΟΣ $Χ^2$

1.1 Διάκριση Μεταβλητών

Μια στατιστική ανάλυση δεν περιορίζεται στη μελέτη μιας μεταβλητής, αλλά πάντοτε απαιτείται η μελέτη της σχέσης μεταξύ δυο ή και περισσότερων μεταβλητών. Η τεχνική που ακολουθείται για την στατιστική ανάλυση εξαρτάται αποκλειστικά από τη διάκριση των μεταβλητών σε ποιοτικές και ποσοτικές (Μπατσίδης Α. 2014).

Ποιοτικές μεταβλητές είναι εκείνες που δεν επιδέχονται αριθμητικές μετρήσεις, αλλά περιγράφονται οι κατηγορίες στις οποίες ταξινομούνται οι παρατηρήσεις, όπως π.χ το φύλο, το χρώμα της ίριδας, η επιβίωση ή όχι μετά από ορισμένο χρονικό διάστημα των πασχόντων από μια νόσο κ.ο.κ. Η απλούστερη μορφή ποιοτικών παρατηρήσεων είναι εκείνη με δυο μόνο κατηγορίες, όπως είναι το φύλο (άνδρας- γυναίκα). Τα δεδομένα αυτά ονομάζονται δυαδικά (binary). Άλλες ποιοτικές μεταβλητές έχουν περισσότερες από δυο κατηγορίες, όπως η ομάδα αίματος (Α,Β,ΑΒ,Ο)

Ποσοτικές μεταβλητές είναι εκείνες οι οποίες επιδέχονται αριθμητικές μετρήσεις, όπως τα φυσικά μεγέθη (ανάστημα, βάρος), τα βιολογικά μεγέθη (χοληστερίνη, σάκχαρο). Οι ποσοτικές μεταβλητές χωρίζονται σε συνεχείς και ασυνεχείς. Οι συνεχείς ποσοτικές μεταβλητές μπορούν να πάρουν θεωρητικά όλες τις τιμές των πραγματικών αριθμών, τουλάχιστον σε ένα διάστημα. Στην πράξη βέβαια, συνήθως μετρούνται με μια καθορισμένη εκ των προτέρων ακρίβεια. Ασυνεχείς λέγονται εκείνες που είναι δυνατόν να λάβουν μόνο ορισμένες αριθμητικές τιμές. (Τριχόπουλος Δ.,Τζώνου Α., Κατσουγιάννη Κ. 2001).

1.2 Διαξονική Ταξινόμηση Των Ποιοτικών Παρατηρήσεων

Όταν οι παρατηρήσεις έχουν ποιοτικό χαρακτήρα ο στατιστικός έλεγχος γίνεται με σύγκριση των συχνοτήτων των παρατηρήσεων στις διάφορες ταξινομητικές κατηγορίες. Οι ερευνητικές μονάδες, ως μέλη ενός συνόλου, μπορεί να ταξινομηθούν με πολλούς τρόπους. Για παράδειγμα, τα άτομα ενός δείγματος μπορεί να ταξινομηθούν ανάλογα με το φύλο, το επάγγελμα, την ομάδα αίματος, το αν πάσχουν από κάποια νόσο κ.ο.κ. Τα μέλη ενός συνόλου μπορεί να ταξινομηθούν είτε διαδοχικά είτε ταυτόχρονα. Η ταυτόχρονη ταξινόμηση με δυο τρόπους καλείται

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ταξινόμηση δυο διευθύνσεων ή διαξονική (two-way classification). Η διαξονική διάταξη αποτελεί προϋπόθεση εφαρμογής της στατιστικής δοκιμασίας X^2 για τον έλεγχο της συσχέτισης, συνάφειας ποιοτικών χαρακτηριστικών. Η αξιοπιστία της συσχέτισης δεν επηρεάζεται από την αναλογική σχέση των αριθμών των παρατηρήσεων στις διάφορες κατηγορίες. Για δεδομένο ολικό αριθμό παρατηρήσεων, η ισχύς της δοκιμασίας X^2 (δηλαδή η ικανότητα τεκμηρίωσης μιας πραγματικής σχέσης) είναι μικρότερη, όταν η ανισότητα του αριθμού των παρατηρήσεων μεταξύ των διαφόρων κατηγοριών είναι μεγαλύτερη. (Τριχόπουλος Δ.,Τζώνου Α., Κατσουγιάννη Κ. 2001).

Μια από τις στατιστικές μεθόδους που χρησιμοποιούνται για την εύρεση της πιθανής σχέσης μεταξύ δυο ποιοτικών μεταβλητών είναι το X^2

1.3 Πίνακας Συνάφειας- X^2 Του Pearson

Για να εξετάσουμε την αλληλεπίδραση δύο (ποιοτικών) μεταβλητών χρησιμοποιούμε έναν πίνακα συνάφειας. Πίνακας συνάφειας ή πίνακας συχνοτήτων είναι ένας πίνακας, συνήθως διδιάστατος (στο επίπεδο), με r γραμμές (rows) και c στήλες (columns) και rxc κυψελίδες (κελιά) στις οποίες καταγράφονται οι συχνότητες των περιπτώσεων εμφάνισης σε ένα τυχαίο δείγμα n στοιχείων (π.χ. ατόμων) των τιμών δυο χαρακτηριστικών Α και Β τα οποία παίρνουν τιμές 1,2,3,….,r και 1,2,3,….,c , αντίστοιχα.(Παπαϊωάννου Τ., Λουκάς Σ. 2002) Σε κάθε διαξονικό πίνακα υπάρχουν τόσες στήλες, όσες και οι κατηγορίες της οριζόντιας ταξινόμησης (εξαιρείται η στήλη των συνόλων) και τόσες σειρές, όσες και οι κατηγορίες της κάθετης ταξινόμησης (εξαιρείται η σειρά των συνόλων). Όταν έχουμε δυαδικές μεταβλητές απόκρισης, κατασκευάζουμε ένα 2x2 πίνακα συνάφειας, όπως φαίνεται στον παρακάτω πίνακα

Α (Έκθεση στον παράγοντα κινδύνου)	**Β** (Έκβαση)		ΣΥΝΟΛΟ
	ΝΑΙ	ΟΧΙ	
ΝΑΙ	α	β	α+β
ΟΧΙ	γ	δ	γ+δ
ΣΥΝΟΛΟ	α+γ	β+δ	α+β+γ+δ

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

Οι δυο μεταβλητές μας Α και Β, ακολουθούν διωνυμική κατανομή, με πιθανότητες επιτυχίας ρ_1 και ρ_2 αντίστοιχα.

Παράδειγμα 2x2 πίνακα (Ταξινόμηση 500 ατόμων ανάλογα με το αν είναι γεωργοί ή όχι και την προσβολή τους ή μη από καρκίνο.

ΚΑΡΚΙΝΟΣ	**ΓΕΩΡΓΟΙ**		ΣΥΝΟΛΟ
	ΝΑΙ	ΟΧΙ	
ΝΑΙ	140	55	195
ΟΧΙ	240	65	305
ΣΥΝΟΛΟ	380	120	500

Σε καθένα από τα rxc κελιά αντιστοιχεί μια παρατηρηθείσα συχνότητα που παριστάνεται με το σύμβολο Ο (από το αγγλικό Observed frequency). Σε όσα ακολουθούν συμβολίζεται με O_{ij} η παρατηρούμενη συχνότητα του (i,j) κελιού δηλαδή ο αριθμός των περιπτώσεων που ανήκουν στην i και j κατηγορία της πρώτης και δεύτερης ποιοτικής μεταβλητής, αντίστοιχα. Έτσι, στην κατηγορία των γεωργών-καρκινοπαθών η παρατηρηθείσα συχνότητα είναι O_{11}=140, στην κατηγορία των γεωργών-μη καρκινοπαθών είναι O_{21}=240, στην κατηγορία μη γεωργών-καρκινοπαθών είναι O_{12} =55 και στην κατηγορία μη γεωργών-μη καρκινοπαθών είναι O_{22}=65.

Επιπρόσθετα, με E_{ij} συμβολίζεται η αναμενόμενη συχνότητα του (i,j) κελιού, δηλαδή ο αριθμός των περιπτώσεων κάθε κελιού αν οι προς μελέτη μεταβλητές ήταν στατιστικά ανεξάρτητες. Η αναμενόμενη συχνότητα E_{ij}, που παριστάνεται με το σύμβολο Ε (από το αγγλικό Expected frequency), δίνεται από τη σχέση:

$$E_{ij}=\frac{\sum_{i=1}^{r} O_{ij} \sum_{j=1}^{c} O_{ij}}{\sum_{i=1}^{r} \sum_{j=1}^{c} O_{ij}} = \frac{\sum_{i=1}^{r} O_{ij} \sum_{j=1}^{c} O_{ij}}{n},$$

όπου n το μέγεθος δείγματος. (Μπατσίδης Α.2014).

Εναλλακτικά, ο υπολογισμός των αναμενόμενων συχνοτήτων σε οποιοδήποτε κελί πραγματοποιείται με τη βοήθεια του τύπου:

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

$$E=\frac{(Αντίστοιχο\ οριζόντιο\ σύνολο)*(Αντίστοιχο\ κάθετο\ σύνολο)}{(Γενικό\ σύνολο)}$$

Ο έλεγχος της στατιστικής σημαντικότητας της σχέσης που ενδέχεται να υπάρχει μεταξύ των δυο χαρακτηριστικών του διαξονικού πίνακα, δηλαδή στο συγκεκριμένο παράδειγμα, της σχέσης που ενδέχεται να υπάρχει μεταξύ επαγγέλματος (γεωργός ή όχι) και ανάπτυξης ή όχι καρκίνου γίνεται συνήθως με τη δοκιμασία X^2. (Τριχόπουλος Δ.,Τζώνου Α., Κατσουγιάννη Κ. 2001). Το X^2 test του Pearson αναπτύχθηκε κατά το έτος 1900 και είναι επίσης γνωστό ως X^2 test για την ανεξαρτησία. Το X^2 test του Pearson αποτελεί μια στατιστική δοκιμή η οποία εφαρμόζεται σε σύνολα κατηγοριοποιημένων δεδομένων για να αξιολογήθει πόσο πιθανό είναι οποιαδήποτε παρατηρούμενη διαφορά μεταξύ των συνόλων να προέκυψε κατά τύχη. Αν τα δεδομένα είναι ταιριαστά, εξαρτημένα και ο πίνακας 2x2 τότε χρησιμοποιείται η διαδικασία McNemar. Αφού κατασκευάσουμε τον πίνακα συνάφειας πραγματοποιούμε ένα X^2 test για να ελέγξουμε την ύπαρξη ή όχι ανεξαρτησίας μεταξύ δυο ποιοτικών μεταβλητών. Συγκεκριμένα ελέγχουμε την υπόθεση

- H_0: Οι δυο μεταβλητές μας είναι ανεξάρτητες

έναντι της εναλλακτικής

- $H_α$: Οι δυο μεταβλητές μας είναι εξαρτημένες.

Το X^2 στατιστικό test δίνεται από τη σχέση:

$$X^2=\sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

όπου O_{ij} η παρατηρούμενη συχνότητα του (i,j) κελιού και E_{ij} η αναμενόμενη συχνότητα του (i,j) κελιού. Γίνεται αντιληπτό ότι μεγάλες αποκλίσεις των αναμενόμενων τιμών από τις παρατηρούμενες τιμές υποδηλώνει πιθανή ύπαρξη σχέσης εξάρτησης, άρα η υπόθεση της ανεξαρτησίας απορρίπτεται για μεγάλες τιμές της παραπάνω στατιστικής συνάρτησης.
Οι βαθμοί ελευθερίας για το X^2 στατιστικό test δίνονται από τη σχέση: (αριθμός γραμμών πίνακα-1)*(αριθμός στηλών πίνακα-1)=(r-1)*(c-1).(Norusis, M.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

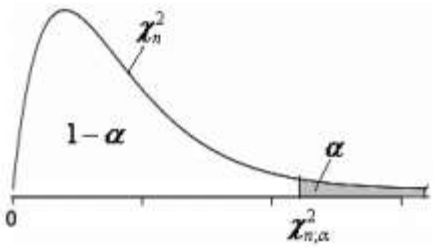
2011) και επομένως η υπόθεση της ανεξαρτησίας απορρίπτεται σε επίπεδο σημαντικότητας α, όταν

$$X^2 \geq X^2_{(r-1)(c-1),a}$$

ή όταν

$$\text{p-τιμή} = P\left(X^2_{(r-1)(c-1)} > X^2\right) < \alpha.$$

Παρατήρηση Το παρακάτω σχήμα παρουσιάζει τη μορφή της X^2 κατανομής



Το σχήμα καθορίζεται από τους βαθμούς ελευθερίας (β.ε) που στην περίπτωση των 2x2 πινάκων έχουμε β.ε =(2-1)*(2-1)=1 και το 5% σημείο της X^2 κατανομής με 1 βαθμό ελευθερίας είναι $X^2_{1,a} = 3{,}84$.

Στην περίπτωση όπου η υπόθεση της ανεξαρτησίας απορρίπτεται πρέπει να διαπιστώσουμε ποια κελία «προκαλούν» το πρόβλημα της εξάρτησης των δυο μεταβλητών. Προκειμένου να το πετύχουμε αυτό αρκεί να παρατηρήσουμε τις αναμενόμενες τιμές ή καλύτερα τις τιμές των Adjusted Standardized Residuals που δίνονται από τη σχέση

$$d_{ij} = \frac{\frac{\left(O_{ij}-E_{ij}\right)}{\sqrt{E_{ij}}}}{\sqrt{\left(1-\frac{n_i}{n}\right)\left(1-\frac{n_j}{n}\right)}},$$

και τα οποία ακολουθούν κατά προσέγγιση κανονική κατανομή όταν οι μεταβλητές του πίνακα συνάφειας είναι ανεξάρτητες μεταξύ τους. Έτσι οι τιμές των d_{ij} μπορούν να θεωρηθούν ως z-τιμές. Οι τιμές που είναι μεγαλύτερες κατά απόλυτη τιμή από το

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

$z_{0.025}$=1.96 υποδεικνύουν κελιά που διαφέρουν από το μοντέλο της ανεξαρτησίας σε επίπεδο σημαντικότητας α=5%. Επιπλέον, για να διερευνήσουμε την ένταση της σχέσης των δυο μεταβλητών υπάρχουν διαθέσιμα στατιστικά μέτρα.

Ορισμένα από τα στατιστικά αυτά μέτρα είναι:

- Ο συντελεστής Phi που αναφέρεται και ως συντελεστής του Pearson με τύπο $\Phi=\sqrt{\frac{X^2}{n}}$

Η μέγιστη τιμή του εξαρτάται από το μέγεθος του πίνακα και όταν το Φ παίρνει την τιμή 0 τότε οι μεταβλητές είναι ανεξάρτητες.

- Ο συντελεστής συνάφειας με τύπο $C=\sqrt{\frac{X^2}{(X^2+n)}}$

Όταν οι τιμές του C είναι κοντά στο 0 τότε οι μεταβλητές είναι ανεξάρτητες, ενώ η μέγιστη τιμή του είναι μικρότερη του 1

- Ο συντελεστής V του Cramer με τύπο $V=\sqrt{\frac{X^2}{n \min{(r-1,c-1)}}}$

Ο συντελεστής V στην περίπτωση των 2x2 πινάκων ταυτίζεται με τον συντελεστή Phi και παίρνει τιμές από 0 έως 1. Όταν το V λαμβάνει την τιμή 0 τότε οι μεταβλητές μας είναι ανεξάρτητες ενώ όταν παίρνει την τιμή 1 τότε έχουμε απόλυτη συνάφεια. (Μπατσίδης Α. 2014)

Το X^2 test εφαρμόζεται όταν πληρούνται οι προϋποθέσεις:

1. Το μέγεθος του δείγματος είναι τετραπλάσιο του πλήθους των κελιών.
2. Οι αναμενόμενες συχνότητες δεν είναι μικρότερες του 1 και το 25% αυτών δεν είναι μικρότερες του 5

Όταν δεν πληρούνται αυτές οι δυο προϋποθέσεις, τότε στην περίπτωση των 2x2 κελιών χρησιμοποιείται το ακριβές στατιστικό του Fisher, ενώ σε κάθε άλλη περίπτωση πρέπει να γίνει συγχώνευση γειτονικών κελιών, με τέτοιο τρόπο ώστε να εξαλείφεται το πρόβλημα αλλά συγχρόνως να υπάρχει φυσική ερμηνεία των νέων κελιών. Η συγχώνευση των κελιών επιτυγχάνεται με επανακωδικοποίηση μίας εκ των

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

δυο ποιοτικών μεταβλητών. Επιπλέον στην περίπτωση των 2x2 πινάκων χρησιμοποιείται αντί του κλασσικού X^2 test η διόρθωση συνέχειας του Yates.

1.4 Το Ακριβές test Του Fisher (Fisher's exact test)

Το ακριβές test του Fisher αρχικά βασίστηκε στον έλεγχο ανεξαρτησίας δυο μεταβλητών για πίνακες συνάφειας 2x2. Εφαρμόζεται όμως και για πίνακες συνάφειας μεγαλύτερης διάστασης. Στη συνέχεια θα περιγράψουμε τον έλεγχο αυτό για πίνακες συνάφειας 2x2.

Η γενική μορφή του 2x2 πίνακα συνάφειας είναι η ακόλουθη:

X	Y		Σύνολο
	1	2	
1	O_{11}	O_{12}	$O_{1\bullet}$
2	O_{21}	O_{22}	$O_{2\bullet}$
Σύνολο	$O_{\bullet 1}$	$O_{\bullet 2}$	$O=O_{\bullet\bullet}$

Μας ενδιαφέρει να ελέγξουμε κατά πόσο θεωρείται τυχαία μια πραγματοποίησή του δηλαδή κατά πόσον είναι τυχαίες οι τιμές των O_{11}, O_{12}, O_{21}, O_{22} εφόσον είναι γνωστά τα αθροίσματα των γραμμών $O_{1\bullet}$, $O_{2\bullet}$ και των στηλών $O_{\bullet 1}$, $O_{\bullet 2}$

Παρατηρούμε ότι αρκεί να ελέγξουμε κατά πόσο ήταν τυχαίο το O_{11} που εμφανίστηκε καθώς επειδή γνωρίζουμε τα αθροίσματα των γραμμών και των στηλών τα O_{12}, O_{21}, O_{22}, μπορούν να εξαχθούν από το O_{11}. Έστω λοιπόν ότι από το δείγμα που πήραμε βρέθηκε ότι $O_{11}=o_{11}$. Η πιθανότητα να συμβεί αυτό τυχαία δεδομένου ότι

$O_{1\bullet}=o_{1\bullet}$,$O_{2\bullet}=o_{2\bullet}$,$O_{\bullet 1}=o_{\bullet 1}$,$O_{\bullet 2}=o_{\bullet 2}$ δίνεται από την υπεργεωμετρική κατανομή:

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

$$p=\frac{\binom{o_{1.}}{o_{11}}\binom{o_{2.}}{o_{.1}-o_{11}}}{\binom{o_{1.}+o_{2.}}{o_{.1}}}$$

Το p-value του ελέγχου της υπόθεσης H_0 : το αποτέλεσμα στα τέσσερα κελιά είναι τυχαίο (δεδομένων των αθροισμάτων των γραμμών και των στηλών) είναι ίσο με την πιθανότητα (υπό την H_0) να εμφανιστεί το δείγμα ακόμα πιο «ακραίο» από αυτό δηλαδή:

$$\text{p-value} =\sum_{i=0}^{o_{11}} \frac{\binom{o_{1.}}{i}\binom{o_{2.}}{o_{.1}-i}}{\binom{o_{1.}+o_{2.}}{o_{.1}}} ,$$

ή

$$\text{p-value} = \sum_{i=o_{11}}^{o_{1.}} \frac{\binom{o_{1.}}{i}\binom{o_{2.}}{o_{.1}-i}}{\binom{o_{1.}+o_{2.}}{o_{.1}}}$$

ανάλογα με το αν $o_{11}<E(O_{11})=\frac{o_{.1}\ o_{1.}}{O}$ ή $o_{11}>E(O_{11})=\frac{o_{.1}\ o_{1.}}{O}$ αντίστοιχα.

Το παραπάνω p-value αντιστοιχεί σε μονόπλευρο έλεγχο. Για δίπλευρο έλεγχο συνήθως λαμβάνεται το

$$\text{p-value}=\sum_{i=0}^{o_{11}} \frac{\binom{o_{1.}}{i}\binom{o_{2.}}{o_{.1}-i}}{\binom{o_{1.}+o_{2.}}{o_{.1}}} + \sum_{i=\left[2\frac{o_{.1}\ o_{1.}}{O}-o_{11}+0,5\right]}^{o_{1.}} \frac{\binom{o_{1.}}{i}\binom{o_{2.}}{o_{.1}-i}}{\binom{o_{1.}+o_{2.}}{o_{.1}}} ,$$

αν $o_{11}< \frac{o_{.1}\ o_{1.}}{O}$

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

1.5 Διόρθωση Συνέχειας Του Yates

Προκειμένου να αντισταθμιστεί εν μέρει η ανακρίβεια που εισάγεται από τη χρήση μιας συνεχούς συνάρτησης κατανομής (της X^2) για την προσέγγιση της διακριτής συνάρτησης κατανομής της στατιστικής συνάρτησης $T=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$, ο Yates (1934) πρότεινε τη λεγόμενη διόρθωση συνέχειας (correction for continuity). Η κατά Yates τροποποιημένη μορφή της στατιστικής συνάρτησης T έχει τη μορφή $T'=\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{(|O_{ij}-E_{ij}|-0,5)^2}{E_{ij}}\sim\chi_1^2$ Η διόρθωση αυτή δεν πραγματοποιείται για να έχουμε καλύτερη προσέγγιση της X^2 κατανομής αλλά για να έχουμε p-value πιο κοντά στο p-value που προκύπτει από το Fisher's exact test. Επομένως η διόρθωση αυτή γίνεται όταν θέλουμε και δεν μπορούμε να υπολογίσουμε το p-value του Fisher's exact test. Ωστόσο σήμερα με τη χρήση των ηλεκτρονικών υπολογιστών ο ακριβής υπολογισμός του p-value του Fisher's exact test είναι εφικτός ακόμη και για μεγάλα δείγματα και επομένως η διόρθωση αυτή έχει μικρότερη αξία από αυτήν που είχε στο παρελθόν. (Μπούτσικας, Μ. 2004).

1.6 ΠΑΡΑΔΕΙΓΜΑ

Θέλουμε να εξετάσουμε αν η εμφάνιση της στεφανιαίας νόσου επηρεάζεται από την ηλικιακή ομάδα των ατόμων. Επιλέγουμε 200 άτομα από τα οποία τα 92 έχουν ηλικία μικρότερη ή ίση των 40 ετών, ενώ τα υπόλοιπα 108 άτομα έχουν ηλικία μεγαλύτερη των 40 ετών. Καταγράφηκαν τα υγιή (0) και ασθενή (1) άτομα και των δυο ηλικιακών ομάδων. Τα αποτελέσματα καταγράφονται στον παρακάτω 2x2 πίνακα

Ηλικία	Νόσος		
	0:Υγιής	1:Ασθενής	Σύνολο
≤40	86	6	92
>40	88	20	108
Σύνολο	174	26	200

Συγκεκριμένα έχουμε 200 περιπτώσεις-cases και δυο μεταβλητές (στεφανιαία νόσος με τιμές Υγιής, Ασθενής και ηλικία με τιμές ≤40, >40). Δηλαδή πρόκειται για έλεγχο

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ανεξαρτησίας. Η ανάλυση των δεδομένων θα γίνει με τη βοήθεια του στατιστικού προγράμματος SPSS από το output του οποίου λαμβάνουμε τον παρακάτω πίνακα.

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6,322[a]	1	,012		
Continuity Correction[b]	5,306	1	,021		
Likelihood Ratio	6,695	1	,010		
Fisher's Exact Test				,019	,009
Linear-by-Linear Association	6,290	1	,012		
N of Valid Cases	200				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 11,96.

b. Computed only for a 2x2 table

Ο πίνακας αυτός μας δίνει την τιμή της στατιστικής συνάρτησης (Pearson Chi-Square) και την τιμή της στατιστικής συνάρτησης Τ΄ με «διόρθωση» (Continuity Correction). Οι βαθμοί ελευθερίας είναι (J-1)(K-1)=1*1=1 Τα αντίστοιχα p-value είναι 0.012 και 0.021. Ακόμα δίνεται και η τιμή του p-value για τον μονόπλευρο (0.009) και δίπλευρο (0.019) έλεγχο που προκύπτει από το Fisher's Exact Test. Επομένως με βάση τα παραπάνω (οι p-value<0.05) απορρίπτουμε ότι υπάρχει ανεξαρτησία της ηλικιακής ομάδας των ατόμων με την εμφάνιση στεφανιαίας νόσου με επίπεδο σημαντικότητας α=5%. Γίνεται αντιληπτό ότι η εμφάνιση της στεφανιαίας νόσου σχετίζεται με την ηλικία με επίπεδο σημαντικότητας α=5%.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ΚΕΦΑΛΑΙΟ 2^{Ο}

Λογιστική Παλινδρόμηση

Τα μοντέλα Παλινδρόμησης αποτελούν βασικό εργαλείο κάθε στατιστικής ανάλυσης που αφορά τη διερεύνηση πιθανής εξάρτησης μιας μεταβλητής απόκρισης (response variable) από μια ή περισσότερες επεξηγηματικές μεταβλητές (explanatory variables). Συχνά η μεταβλητή απόκρισης είναι διακριτή και λαμβάνει δύο ή περισσότερες τιμές (π.χ η πρόβλεψη αν ένα νεογνό θα είναι ελλιποβαρές ή όχι). Σε τέτοιες περιπτώσεις, δηλαδή όταν ενδιαφερόμαστε να εξετάσουμε ως μεταβλητή απόκρισης μία δίτιμη ποιοτική μεταβλητή και να διερευνήσουμε κατά πόσο επηρεάζεται από ποσοτικά ή ποιοτικά χαρακτηριστικά, εφαρμόζουμε το μοντέλο της λογιστικής παλινδρόμησης. Σε αντίθεση με τα στατιστικά μοντέλα γραμμικής παλινδρόμησης τα οποία έχουν ως εξαρτημένη μια ποσοτική μεταβλητή, η λογιστική παλινδρόμηση απαιτεί ως εξαρτημένη να χρησιμοποιηθεί μία ποιοτική μεταβλητή που λαμβάνει δύο ή περισσότερες τιμές. Η διαφορά μεταξύ γραμμικής και λογιστικής παλινδρόμησης αντικατοπτρίζεται τόσο στην επιλογή του παραμετρικού μοντέλου, όσο και στις υποθέσεις που πρέπει να ισχύουν για την έγκυρη εφαρμογή τους (Τριανταφύλλου, Ι. 2016). Υπάρχει μια τεράστια γκάμα εφαρμογών στις οποίες η εξαρτημένη μεταβλητή είναι δυαδική. Μια τέτοια περίπτωση είναι η μελέτη της στεφανιαίου νόσου καρδιοπάθειας ως μια συνάρτηση της ηλικίας, του φύλου, του ιστορικού καπνίσματος, του επιπέδου χοληστερίνης, του ποσοστού του ιδανικού βάρους του σώματος, και της πίεσης του αίματος. Η αποκρινόμενη μεταβλητή Y ορίστηκε να έχει τα δυο πιθανά αποτελέσματα: άτομο που εκδήλωσε καρδιοπάθεια κατά την διάρκεια της μελέτης και άτομο που δεν εκδήλωσε καρδιοπάθεια κατά την διάρκεια της μελέτης. Αυτά τα αποτελέσματα μπορούν να κωδικοποιηθούν με 1 και 0 αντίστοιχα.

Οι Dixon and Massey(1983) παρουσίασαν μια μελέτη για την εμφάνιση στεφανιαίας νόσου σε 200 άνδρες. Για τα άτομα αυτά κατέγραψαν αν είχαν υποστεί κάποιο καρδιακό επεισόδιο κατά τη διάρκεια των τελευταίων 10 ετών (CNT=1) ή όχι(CNT=0). Επίσης ήταν γνωστά μία σειρά από άλλα χαρακτηριστικά όπως η ηλικία των ατόμων, μετρήσεις της συστολικής και διαστολικής πίεσης του αίματος, τα επίπεδα χοληστερίνης, το εύρος και το βάρος των ατόμων. Μια κλασσική επιδημιολογική προσέγγιση θα προσπαθούσε να αναλύσει την πιθανότητα εμφάνισης στεφανιαίας νόσου σε σχέση με έναν παράγοντα κινδύνου π.χ η ηλικία. Στη μελέτη

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

αυτή η ηλικία μετριέται σε συνεχή κλίμακα αλλά ας υποθέσουμε ότι μπορούμε να χωρίσουμε τα άτομα σε άνδρες ηλικίας 40 ετών και άνω των 40 ετών. Τότε μπορούμε να κατασκευάσουμε έναν 2x2 πίνακα και να υπολογίσουμε μέτρα κινδύνου.

Πίνακας με συχνότητες εμφάνισης στεφανιαίας νόσου ανά ηλικιακή ομάδα

Ηλικία	Νόσος		
	0:Υγιής	1:Ασθενής	Σύνολο
≤40	86	6	92
>40	88	20	108
Σύνολο	174	26	200

Ο σχετικός κίνδυνος είναι ίσος με $RR=\frac{\frac{a}{a+b}}{\frac{c}{c+d}}$ =1,15 Αυτό σημαίνει ότι ο κίνδυνος εμφάνισης στεφανιαίας νόσου σε άνδρες άνω των 40 ετών είναι 1,15 φορές μεγαλύτερος από τον αντίστοιχο κίνδυνο που έχουν οι άντρες μικρότεροι των 40 ετών. Ο λόγος συμπληρωματικών πιθανοτήτων (Odds Ratio) υπολογίζεται ως $OR=\frac{a*d}{b*c}$ = 3,25 Δηλαδή η σχετική πιθανότητα εμφάνισης της νόσου σε άντρες άνω των 40 ετών είναι 3,25 φορές τον κίνδυνο εμφάνισης που έχουν οι νεώτεροι. Αυτή η μορφή ανάλυσης παρέχει ορισμένες πληροφορίες για τον κίνδυνο εμφάνισης της νόσου ανά ηλικία ωστόσο είναι ελλιπής καθώς δεν λαμβάνονται υπόψη άλλοι σημαντικοί παράγοντες που μπορεί να σχετίζονται με τη στεφανιαία νόσο, όπως μετρήσεις για τη συστολική και διαστολική πίεση των ανθρώπων αυτών, τα επίπεδα χοληστερίνης, τις διατροφικές συνήθειες και άλλες μεταβλητές που μπορεί να επηρεάζουν την υγεία των ατόμων που συμμετέχουν στη μελέτη. Επιπλέον, ο τρόπος που δημιουργήθηκε ο πίνακας αυτός βασίζεται στην αυθαίρετη διχοτόμηση μιας συνεχούς μεταβλητής. Με αυτόν τον τρόπο προκύπτουν αρκετά προβλήματα. Αρχικά δεν υπάρχει καμία επιστημονική επιχειρηματολογία που να υπαγορεύει ότι ο διαχωρισμός των ατόμων πρέπει να γίνει στα 40 έτη. Αν αντί αυτής της ηλικίας επιλεγεί η ηλικία των 50 ετών τότε οι εκτιμήσεις αλλάζουν. Ακόμα, η απώλεια πληροφορίας από την κατηγοριοποίηση αυτή είναι πολύ σημαντική. Ενώ υπάρχουν στοιχεία για την ηλικία των 200 ατόμων, η επιλογή να χωριστούν αυτά τα άτομα σε δυο μεγάλες ομάδες οδηγεί σε μη ασφαλή συμπεράσματα. Έτσι, ένας άντρας 41 ετών

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

θεωρείται ότι έχει 225% μεγαλύτερο κίνδυνο να πάθει στεφανιαία νόσο σε σχέση με έναν άντρα που είναι μόλις ένα χρόνο νεώτερος του, ενώ έχει τον ίδιο ακριβώς κίνδυνο με έναν άντρα που μπορεί να είναι 30 χρόνια μεγαλύτερος. Μια εναλλακτική προσέγγιση θα ήταν να χωρισθεί η συνεχής μεταβλητή της ηλικίας σε περισσότερες από δύο ομάδες, για παράδειγμα σε ομάδες εύρους 10 ετών. Με τον τρόπο αυτό η απώλεια πληροφορίας είναι μικρότερη, αλλά εξακολουθεί να υπάρχει. Για να ξεπεραστούν τέτοιου είδους προβλήματα θα ήταν χρήσιμη μια μορφή ανάλυσης που να είναι στην ίδια λογική με την ανάλυση παλινδρόμησης. Σε ένα γραμμικό μοντέλο η ανεξάρτητη μεταβλητή, στο παράδειγμά μας η ηλικία, μπορεί να χρησιμοποιηθεί ως συνεχής, ενώ ταυτόχρονα μπορούν να συμπεριληφθούν στο στατιστικό μοντέλο και άλλες μεταβλητές που πιθανόν να επηρεάζουν την εξαρτημένη μεταβλητή. Σε ένα απλό μοντέλο γραμμικής παλινδρόμησης ο μέσος μιας συνεχούς απόκρισης y μπορεί να περιγραφεί ως μια σχέση με μια ανεξάρτητη μεταβλητή (συνεχής ή κατηγορική) με την ακόλουθη σχέση:

$$E(y/X)=\beta_0+\beta_1X$$

Μια τέτοια μορφή αναπαράστασης θα ήταν πολύ χρήσιμη για να περιγραφεί η σχέση της ηλικίας με τον κίνδυνο εμφάνισης στεφανιαίας νόσου. Ωστόσο, η βασική υπόθεση που πρέπει να ισχύει για να έχει νόημα το γραμμικό μοντέλο $E(y/X)=\beta_0+\beta_1X$ είναι ότι η εξαρτημένη μεταβλητή y ακολουθεί κανονική κατανομή. Στο παράδειγμά μας, εξαρτημένη μεταβλητή είναι η εμφάνιση ή όχι στεφανιαίας νόσου. Μια δίτιμη μεταβλητή όμως δεν μπορεί να ακολουθεί κανονική κατανομή. Επίσης, μας ενδιαφέρει ο κίνδυνος εμφάνισης της νόσου, δηλαδή η δεσμευμένη πιθανότητα $P(Y=1|X)$ να εμφανιστεί η νόσος με βάση μια οποιαδήποτε έκθεση. Έτσι θα χρειαζόταν ένα μοντέλο της μορφής:

$$P(Y=1|X)= \beta_0+\beta_1X$$

Στην περίπτωση αυτή θα πρέπει να αντιμετωπιστεί το στατιστικό πρόβλημα ότι η εξαρτημένη μεταβλητή δεν ακολουθεί κανονική κατανομή καθώς επίσης και το αριθμητικό πρόβλημα ότι το δεξί μέλος της εξίσωσης θα πρέπει να περιοριστεί να δίνει τιμές στο διάστημα (0,1). Τα προβλήματα αυτά αντιμετωπίζονται με τη χρήση της λογιστικής παλινδρόμησης.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

2.1 Το μαθηματικό μοντέλο

Το μοντέλο παλινδρόμησης που περιγράφει τον κίνδυνο εμφάνισης μιας ασθένειας θα πρέπει να δίνει τιμές μέσα στο διάστημα (0,1). Θα πρέπει λοιπόν να χρησιμοποιηθεί ο κατάλληλος μαθηματικός μετασχηματισμός του δεξιού μέλους της εξίσωσης $E(y/X)=\beta_0+\beta_1X$ έτσι ώστε οποιαδήποτε εκτίμηση και αν προκύπτει για κάποιον κίνδυνο να μην βρίσκεται κάτω από το μηδέν ή πάνω από το ένα. Έστω $\eta= \beta_0+\beta_1X$ Η λογιστική συνάρτηση (logistic function) ορίζεται ως $f(\eta)=\frac{1}{1+\exp(-\eta)}$ και είναι ισοδύναμη με $P(Y=1|X)=\frac{\exp(\eta)}{1+\exp(\eta)}$ που ορίζει το μοντέλο λογιστικής παλινδρόμησης. Το μοντέλο μπορεί να γραφεί ως λόγος σχετικών πιθανοτήτων στη μορφή $\frac{P(Y=1|X)}{1-P(Y=1|X)} = \exp(\beta_0+\beta_1X)$ Παίρνοντας τον λογάριθμο του αριστερού μέλους της εξίσωσης καταλήγουμε στη σχέση:

$$\log[\frac{P(Y=1|X)}{1-P(Y=1|X)}]=\text{logit}(P(Y=1|X))= \beta_0+\beta_1X=\eta$$

Το μοντέλο αυτό συνδέει τον λογάριθμο του λόγου σχετικών πιθανοτήτων εμφάνισης της ασθένειας γραμμικά με μια ανεξάρτητη μεταβλητή. Όταν η εξαρτημένη μεταβλητή y είναι δίτιμη τότε θα πρέπει να ισχύουν οι ακόλουθες υποθέσεις για το μοντέλο λογιστικής παλινδρόμησης:

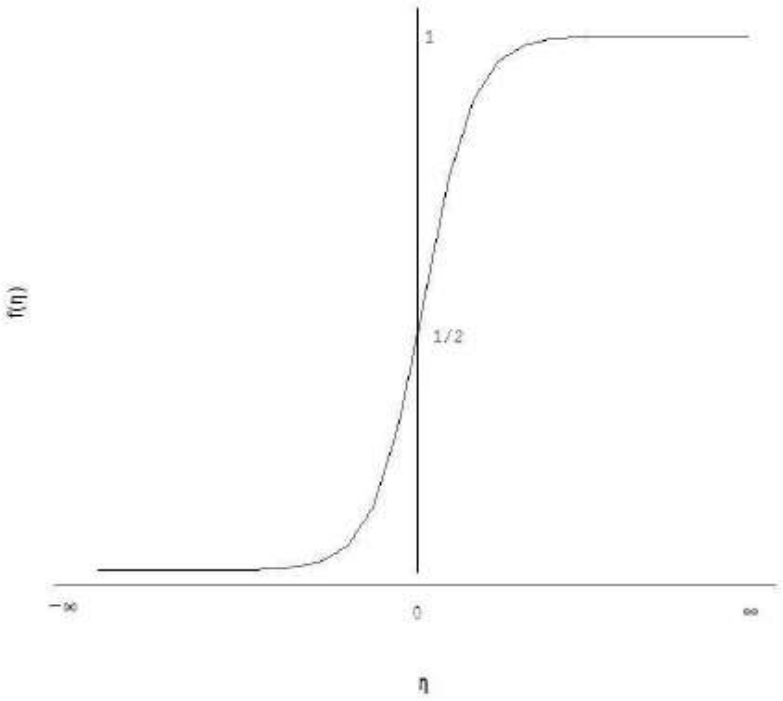
- Το y_i ακολουθεί διωνυμική κατανομή (ο δείκτης i δηλώνει την i-οστη παρατήρηση)
- Ο μέσος $E(Y|X)=P(Y=1|X)$ δίνεται από τη λογιστική συνάρτηση
- Οι τιμές της εξαρτημένης μεταβλητής είναι στατιστικά ανεξάρτητες

Η λογιστική συνάρτηση έχει μερικές πολύ χρήσιμες ιδιότητες. Η συνάρτηση έχει σιγμοειδή μορφή. Όταν το $\eta= -\infty$ τότε η $f(\eta)=0$ ενώ όταν $\eta= \infty$ τότε $f(\eta)=1$. Καθώς προχωράμε από το $\eta= -\infty$ η $f(\eta)$ βρίσκεται κοντά στο 0 ως ότου από κάποιο σημείο και έπειτα αυξάνει με γρήγορο ρυθμό. Προς το δεξί μέρος του διαγράμματος η συνάρτηση πλησιάζει προς τη μονάδα και εκεί ο ρυθμός αύξησης μειώνεται σημαντικά. Η χαρακτηριστική αυτή μορφή της λογιστικής παλινδρόμησης περιγράφει με άριστο τρόπο πολλά επιδημιολογικά φαινόμενα. Για παράδειγμα, ο κίνδυνος εμφάνισης καρδιακής νόσου είναι χαμηλός για τις μικρές ηλικίες και αυξάνει

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ανεπαίσθητα ως ότου ένας άντρας ξεπεράσει ένα κατώφλι, την ηλικία που θεωρείται κρίσιμη για έμφραγμα. Από εκεί και έπειτα ο κίνδυνος αυξάνει δραματικά μέχρι το επόμενο κατώφλι και παραμένει σταθερά υψηλός στα τελευταία ηλικιακά χρόνια. (Ντζούφρας, Ι., Περπέρογλου, Α. 2009)

Μορφή Λογιστικής Συνάρτησης



2.2 ΠΑΡΑΔΕΙΓΜΑ

Στον παρακάτω πίνακα παρουσιάζονται οι συχνότητες εμφάνισης στεφανιαίας νόσου ανά ηλικιακή ομάδα. Η ανάλυση μπορεί να γίνει με την εφαρμογή ενός μοντέλου λογιστικής παλινδρόμησης και τη χρήση του SPSS.

Ηλικία	Νόσος		
	0:Υγιής	1:Ασθενής	Σύνολο
≤40	86	6	92
>40	88	20	108
Σύνολο	174	26	200

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

Η μεταβλητή που δηλώνει την παρουσία (1=ναι) ή όχι (0=όχι) στεφανιαίας νόσου είναι η εξαρτημένη ενώ η μεταβλητή που δηλώνει την ηλικία εισέρχεται ως συμμεταβλητή (covariate) στο μοντέλο. Οι εκτιμήσεις των παραμέτρων υπολογίζονται και παρουσιάζονται στον παρακάτω πίνακα.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	agegroup	1.181	.490	5.820	1	.016	3.258
	Constant	-2.663	.422	39.782	1	.000	.070

a. Variable(s) entered on step 1: agegroup.

Ο συντελεστής B στον πίνακα με το όνομα constant δίνει την εκτίμηση της σταθεράς β_0 του μοντέλου, ενώ ο συντελεστής που αντιστοιχεί στη μεταβλητή agegroup είναι η εκτίμηση για την παράμετρο β_1. Η μεταβλητή agegroup, έχει δύο κατηγορίες, τα άτομα ηλικίας μέχρι 40 ετών (Q=0) και τα άτομα ηλικίας άνω των 40 ετών (Q=1). Με τον τρόπο που έγινε η κωδικοποίηση προκύπτουν από το μοντέλο $\log[\frac{P(Y=1|X)}{1-P(Y=1|X)}]=\text{logit}(P(Y=1|X))= \beta_0+\beta_1X=\eta$, δύο συναρτήσεις. Αν ένα άτομο ανήκει στην πρώτη κατηγορία (κάτω των 40 ετών) τότε ο λογαριθμικός λόγος σχετικών πιθανοτήτων να αναπτύξει στεφανιαία νόσο υπολογίζεται ως:

$$\log \frac{P(Y=1|X=0)}{P(Y=1|X=0)} = \beta_0 + \beta_1X=\beta_0= -2.663$$

ενώ ο αντίστοιχος λόγος για άτομα ηλικίας άνω των 40 ετών είναι:

$$\log \frac{P(Y=1|X=1)}{P(Y=1|X=0)} = \beta_0 + \beta_1X = \beta_0 + \beta_1= -2.663+1.181= -1.482$$

Η διαφορά μεταξύ αυτών των δύο λόγων:

$$\text{logit}\{P(Y=1|X=1)\} – \text{logit}\{P(Y=1|X=0)\}=1.818$$

είναι ο συντελεστής β_1. Ο συντελεστής β_1 δίνει μια εκτίμηση του λογαριθμικού λόγου σχετικών πιθανοτήτων να νοσήσει ένα άτομο ηλικίας άνω των 40 ετών, σε σχέση με ένα άτομο νεώτερης ηλικίας. Με βάση τις εκτιμήσεις που παρουσιάζονται στον πίνακα μπορούν να υπολογιστούν οι πιθανότητες εμφάνισης στεφανιαίας νόσου σε κάποια άτομα, ανάλογα με την ηλικιακή ομάδα στην οποία ανήκουν.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

Έτσι, για ένα άτομο ηλικίας 35 ετών η πιθανότητα εμφάνισης στεφανιαίας νόσου εκτιμάται ως: $P(Y=1|X=0)= \frac{\exp(-2.663)}{1+\exp(-2.663)} = 0.065$ ενώ για ένα άτομο ηλικίας 55 ετών η αντίστοιχη πιθανότητα είναι: $P(Y=1|X=1)= \frac{\exp(-2.663+1.181)}{1+\exp(-2.663+1.181)} =0.185$ (Ντζούφρας, Ι., Περπέρογλου, Α. 2009)

Για να εξετάσουμε αν η εμφάνιση της στεφανιαίας νόσου επηρεάζεται από την ηλικιακή ομάδα των ατόμων μπορούμε να χρησιμοποιήσουμε και τις μεθόδους που αναλύσαμε στο κεφάλαιο 1 (X^2 test). Η ανάλυση των δεδομένων θα γίνει με τη βοήθεια του στατιστικού προγράμματος SPSS από το output του οποίου λαμβάνουμε τον παρακάτω πίνακα.

agegroup * stefaniaia Crosstabulation

			stefaniaia		
			oxi	nai	Total
agegroup	mexri 40	Expected Count	80,0	12,0	92,0
		% within agegroup	93,5%	6,5%	100,0%
	megalytero 40	Expected Count	94,0	14,0	108,0
		% within agegroup	81,5%	18,5%	100,0%
Total		Expected Count	174,0	26,0	200,0
		% within agegroup	87,0%	13,0%	100,0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	6,322[a]	1	,012		
Continuity Correction[b]	5,306	1	,021		
Likelihood Ratio	6,695	1	,010		
Fisher's Exact Test				,019	,009
Linear-by-Linear Association	6,290	1	,012		
N of Valid Cases	200				

a. 0 cells (0,0%) have expected count less than 5. The minimum expected count is 11,96.

b. Computed only for a 2x2 table

Παρατηρούμε ότι οι p-value<0,05 Έτσι γίνεται αντιληπτό ότι η εμφάνιση στεφανιαίας νόσου εξαρτάται από την ηλικιακή ομάδα στην οποία ανήκουν τα άτομα, με επίπεδο σημαντικότητας 5%, γεγονός που επιβεβαιώνεται και με τη μέθοδο της λογιστικής παλινδρόμησης. Συγκρίνοντας τις στατιστικές μεθόδους X^2 και λογιστική παλινδρόμηση, διαπιστώνουμε ότι, για τη σωστή εφαρμογή της λογιστικής παλινδρόμησης, απαιτείται μεγάλο δείγμα προκειμένου να έχουμε αξιόπιστο

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

αποτέλεσμα (Δημητροπουλάκης, 2017), ενώ το στατιστικό κριτήριο X^2 μπορεί να εφαρμοστεί και στην περίπτωση που το δείγμα αποτελείται από μικρό αριθμό παρατηρήσεων. Ένα μειονέκτημα της λογιστικής παλινδρόμησης είναι το ότι παρόλο που δεν απαιτεί οι μεταβλητές να ακολουθούν την κανονική κατανομή, αν αυτές χαρακτηρίζονται από ακραία μη κανονικότητα τότε τα αποτελέσματα του μοντέλου ενδέχεται να μην είναι ικανοποιητικά.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

ΣΥΜΠΕΡΑΣΜΑΤΑ

Η Στατιστική και οι μέθοδοί της έχουν καταλυτική δράση στην εξέλιξη και την πρόοδο των επιστημών και ιδιαίτερα των επιστημών υγείας. Η αλληλεπίδραση Στατιστικής-Ιατρικής παρήγαγε αμοιβαία οφέλη. Οι ιατρικές έρευνες-μελέτες θα πρέπει από την αρχική φάση του σχεδιασμού τους να βασίζονται σε στέρεες στατιστικές αρχές και να επιλέγονται οι κατάλληλες μέθοδοι. Οι στατιστικές μέθοδοι: X^2 του Pearson, το ακριβές test του Fisher, η διόρθωση συνέχειας του Yates και η μέθοδος της Λογιστικής Παλινδρόμησης έχουν συχνή εμφάνιση στις ιατρικές μελέτες που αφορούν ποιοτικές μεταβλητές.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169

Βιβλιογραφία

Δημητροπουλάκης, Π. (2017). Διδακτικές Σημειώσεις, Εισαγωγή στη χρήση του SPSS for Windows

Ζαχαροπούλου, Χ. (2001). *Στατιστική (Μέθοδοι- Εφαρμογές)*. Θεσσαλονίκη: Ζυγός.

Μπατσίδης, Α. (2014). *Διδακτικές Σημειώσεις, Στατιστική Ανάλυση Δεδομένων με το SPSS*. Ιωάννινα: Πανεπιστημιακό Τυπογραφείο Ιωαννίνων.

Μπούτσικας, Μ. (2004*). Σημειώσεις μαθήματος «Στατιστικά Προγράμματα» Τμήμα Στατ. & Ασφ. Επιστήμης*. Πανεπιστήμιο Πειραιώς.

Norusis, M. J. (2011). *Οδηγός Ανάλυσης Δεδομένων με το IBM SPSS 19*. Αθήνα: Κλειδάριθμος.

Ντζούφρας, Ι., Περπέρογλου, Α. (2009). Εισαγωγή στην Βιοστατιστική και την Επιδημιολογία. Αθήνα.

Παπαϊωάννου, Τ., Λουκάς, Σ. (2002). Εισαγωγή στη Στατιστική. Αθήνα: Σταμούλης Αθ.

Τριανταφύλλου, Ι. (2016*). Σημειώσεις για το Στατιστικό Πακέτο IBM SPSS19.0*. Λαμία.

Τριχόπουλος, Δ., Τζώνου, Α., Κατσουγιάννη, Κ. (2001). Βιοστατιστική. Αθήνα: Επιστημονικές Εκδόσεις Παρισιανού.

Τσίπος, Στ., Κωνσταντινίδης, Θ. (2010). *Βασικές Αρχές Βιοστατιστικής Εφαρμογές με χρήση του SPSS*. Αλεξανδρούπολη.

Institutional Repository - Library & Information Centre - University of Thessaly
16/06/2024 17:04:19 EEST - 52.15.60.169