ΝΟΣΗΛΕΥΤΙΚΗ 2010, 49(1): 11–25 • NOSILEFTIKI 2010, 49(1): 11–25

ΓΕΝΙΚΟ ΑΡΘΡΟ – GENERAL ARTICLE

Αποχαιρετώντας τις Τιμές p και Καλωσορίζοντας τα Διαστήματα Εμπιστοσύνης στην Ανάλυση Δεδομένων

Πέτρος Γαλάνης

Confidence Intervals in Data Analysis

Abstract at the end of the article

Νοσηλευτής ΠΕ, MSc Δημόσιας Υγείας, Διδάκτωρ Πανεπιστημίου Αθηνών, Εργαστήριο Οργάνωσης και Αξιολόγησης Υπηρεσιών Υγείας, Τμήμα Νοσηλευτικής, Εθνικό και Καποδιστριακό Πανεπιστήμιο Αθηνών, Αθήνα

Υποβλήθηκε: 3.9.2009
Επανυποβλήθηκε: 27.10.2009
Εγκρίθηκε: 14.12.2009

Υπεύθυνος αλληλογραφίας:
Πέτρος Γαλάνης
Δίκης 14
157 73 Αθήνα
Τηλ.: 210 77 81 044, 6944 387 354
e-mail: pegalan@nurs.uoa.gr

Η διαπίστωση της ύπαρξης ή όχι στατιστικά σημαντικών σχέσεων μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης στηρίζεται στη σύγκριση των τιμών p, που προκύπτουν από τους διάφορους στατιστικούς ελέγχους, με μια αυθαίρετα επιλεγμένη τιμή από τους ερευνητές, η οποία είναι γνωστή ως τιμή α και συνήθως είναι ίση με 0,05. Τα τελευταία 30 χρόνια, ορθά έχει ασκηθεί δριμύτατη κριτική στη χρήση των τιμών p για την εξαγωγή συμπερασμάτων στις επιστήμες υγείας. Η μαθηματική και η εννοιολογική προσέγγιση των ελέγχων υποθέσεων και των τιμών p αποτέλεσε ένα σημαντικό βήμα, αλλά η ερμηνεία τους και η πρακτική τους εφαρμογή αντιμετωπίζουν ακόμη και σήμερα σοβαρά προβλήματα. Κάτω από ιδανικές συνθήκες, η ανάλυση των δεδομένων και η παρουσίαση των αποτελεσμάτων μιας μελέτης θα πρέπει να περιλαμβάνουν τα μέτρα σχέσης –ή σπανιότερα τα μέτρα συχνότητας– και τα αντίστοιχα διαστήματα εμπιστοσύνης, μέσω των οποίων δηλώνεται η ακρίβεια της μέτρησης. Αυτές οι δύο αναγκαίες πληροφορίες παρέχονται μέσω της διαδικασίας της στατιστικής εκτίμησης. Το διάστημα εμπιστοσύνης είναι ένα εύρος τιμών γύρω από ένα μέτρο σχέσης που υπολογίζεται σε μια μελέτη και δείχνει το βαθμό στατιστικής ακρίβειας της εκτίμησης. Ένα ευρύ διάστημα εμπιστοσύνης υποδηλώνει μικρότερη ακρίβεια, ενώ ένα στενότερο υποδηλώνει μεγαλύτερη ακρίβεια. Η τιμή p, έπειτα από τη σύγκρισή της με την τιμή α, απλά δηλώνει την ύπαρξη ή όχι στατιστικής σημαντικότητας χωρίς να καθιστά σαφές εάν η ενδεικτική κατηγορία του μελετώμενου προσδιοριστή αυξάνει ή μειώνει τη συχνότητα εμφάνισης της έκβασης. Για το λόγο αυτό πρέπει να αναφέρεται τόσο το μέτρο σχέσης, που δηλώνει το μέγεθος της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, όσο και το αντίστοιχο διάστημα εμπιστοσύνης που δηλώνει την ακρίβεια της μέτρησης. Η εγκυρότητα και η αξιοπιστία των αποτελεσμάτων μιας μελέτης δεν πρέπει να καθορίζονται από την εύρεση στατιστικά σημαντικών σχέσεων, αλλά από τον ερευνητικό σχεδιασμό και τον περιορισμό των σφαλμάτων. Πάντως, ακόμη και αν δεν απο-

μακρυνθούν άμεσα οι τιμές p από την παρουσίαση των αποτελεσμάτων μιας μελέτης, θα ήταν εξαιρετικά επωφελής η παράθεση των τιμών των μέτρων σχέσης και των αντίστοιχων διαστημάτων εμπιστοσύνης. Εξάλλου, η γνώση ενός διαστήματος εμπιστοσύνης συνεπάγεται έμμεσα και τη γνώση της ύπαρξης ή όχι στατιστικής σημαντικότητας, οπότε οι τιμές p είναι ουσιαστικά «άχρηστες» στην περίπτωση που παρουσιάζονται τα διαστήματα εμπιστοσύνης.

Λέξεις ευρετηρίου: *Ανάλυση δεδομένων, διάστημα εμπιστοσύνης, έλεγχος υπόθεσης, στατιστική σημαντικότητα, τιμή p*

Εισαγωγή

Οι στατιστικοί έλεγχοι υποθέσεων και οι τιμές p που εξάγονται, διαδραματίζουν σήμερα πρωταρχικό ρόλο στην ανάλυση δεδομένων* που αφορούν στις επιστήμες υγείας. Αρκετοί ερευνητές θεωρούν ανώφελο να υποβάλλουν προς κρίση ερευνητικές εργασίες, στις οποίες δεν υπάρχουν έλεγχοι της στατιστικής σημαντικότητας. Η στάση τους αυτή εντούτοις δικαιολογείται σε σημαντικό βαθμό από το γεγονός ότι πολλά περιοδικά, δυστυχώς, θεωρούν αναγκαίους τους ελέγχους υποθέσεων για την εξαγωγή ασφαλών συμπερασμάτων. Η διαπίστωση της ύπαρξης ή όχι στατιστικά σημαντικών σχέσεων μεταξύ προσδιοριστή** και συχνότητας εμφάνισης της έκβασης***στηρίζεται στη σύγκριση των τιμών p, που προκύπτουν από τους διάφορους στατιστικούς ελέγχους, με μια αυθαίρετα επιλεγμένη τιμή από τους ερευνητές, που είναι γνωστή ως τιμή α και συνήθως είναι ίση με 0,05.

Τα τελευταία 30 χρόνια, έχει ασκηθεί δριμύτατη κριτική στη χρήση των ελέγχων υποθέσεων και των τιμών p για την εξαγωγή συμπερασμάτων στις επιστήμες υγείας. Χαρακτηριστικά αναφέρεται ότι το 1997 η International Committee of Medical Journal Editors[4] εξέδωσε οδηγίες για τη δημοσίευση μελετών, τις οποίες υιοθέτησαν >500 περιοδικά παγκόσμια (μεταξύ των οποίων το *Annals of Internal Medicine*, το *Lancet*, το *British Medical Journal*, το New *England Journal of Medicine*, το *Canadian Medical Association Journal* κ.ά.). Στις οδηγίες αυτές συστήνεται η αποφυγή της χρήσης των ελέγχων υποθέσεων και των τιμών p για την εξαγωγή συμπερασμάτων και προτείνεται, αντί των τιμών p, να χρησιμοποιούνται τα διαστήματα εμπιστοσύνης, που υποδηλώνουν ταυτόχρονα τόσο το μέγεθος της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης όσο και το μέγεθος του τυχαίου σφάλματος. Ο Kenneth Rothman, ως διευθυντής σύνταξης ενός από τα πιο έγκριτα περιοδικά Επιδημιολογίας (του "*Epidemiology*"), είναι κατηγορηματικός στο γεγονός ότι δεν πρέπει να δημοσιεύονται μελέτες τα συμπεράσματα των οποίων στηρίζονται στην ύπαρξη ή όχι στατιστικής σημαντικότητας.[5] Επιπλέον, μολονότι πολλά περιοδικά συνιστούν τη χρήση των τιμών p για την εξαγωγή συμπερασμάτων, ο Rothman τάσσεται σαφώς εναντίον τους θεωρώντας αναγκαία την παρουσίαση των διαστημάτων εμπιστοσύνης στα αποτελέσματα μιας μελέτης. Επισημαίνει ότι η εξαγωγή συμπερασμάτων δεν πρέπει να στηρίζεται στην ύπαρξη ή όχι στατιστικής σημαντικότητας για ένα ή περισσότερους προσδιοριστές, αλλά στην εύρεση πιθανών παραγόντων (όπως π.χ., η ύπαρξη συστηματικών σφαλμάτων ή συγχυτικών παραγόντων) που μπορούν να συμβάλλουν στην ερμηνεία των αποτελεσμάτων μιας μελέτης.

---

* *Επισημαίνεται ότι ο όρος «ανάλυση δεδομένων» (data analysis) είναι λαθεμένος, δεδομένου ότι (α) δεν υπάρχουν, σύμφωνα με τη σύγχρονη επιστημολογία, «καθαρά δεδομένα» στις εμπειρικές (πραγματολογικές) επιστήμες που δεν είναι θεωρητικά φορτισμένα και (β) τα «δεδομένα» αυτά δεν αναλύονται, αλλά συντίθενται. Εάν δηλαδή διατηρηθεί ο όρος «δεδομένα», τότε ο όρος «ανάλυση δεδομένων» θα πρέπει να αντικατασταθεί από τον όρο «σύνθεση δεδομένων» (data synthesis).*[1]

** *Παράγοντας κινδύνου (risk factor) ή έκθεση (exposure) ή προσδιοριστής (determinant), όπως τελικά επικράτησε να λέγεται σήμερα, είναι το χαρακτηριστικό (συγγενές, περιβαλλοντικό ή συμπεριφοράς) των ατόμων από το οποίο εξαρτάται (σχετίζεται ή συναρτάται) η συχνότητα εμφάνισης της μελετώμενης έκβασης.*[2,3] *Ο προσδιοριστής της συχνότητας εμφάνισης μιας έκβασης περιλαμβάνει δύο κατηγορίες, την ενδεικτική κατηγορία (index category) και την κατηγορία αναφοράς (reference category). Προσδιοριστής, π.χ., της συχνότητας εμφάνισης της νεφρικής νόσου δεν είναι η αρτηριακή υπέρταση, αλλά η αρτηριακή πίεση. Η αρτηριακή υπέρταση είναι μια κατηγορία και συνήθως η ενδεικτική κατηγορία του προσδιοριστή, στην οποία μελετάται η συχνότητα εμφάνισης της νεφρικής νόσου σε σχέση πάντοτε με τη συχνότητα εμφάνισής της στην κατηγορία αναφοράς, στην προκειμένη περίπτωση στην κατηγορία των ατόμων που δεν έχουν αρτηριακή υπέρταση.*

*** *Έκβαση είναι το αποτέλεσμα ή, αλλιώς, η κατάληξη μιας διαδικασίας. Στην αιτιογνωστική επιδημιολογία, η έκβαση (outcome) χρησιμοποιείται για να δηλώσει την εμφάνιση της πάθησης.*[3] *Η έκφραση «σχετίζεται με την έκβαση» σημαίνει σχέση με τη συχνότητα εμφάνισης της έκβασης και όχι με την έκβαση καθαυτή. Στην προγνωστική Επιδημιολογία, τα μελετώμενα άτομα πάσχουν ήδη από μια συγκεκριμένη νόσο οπότε η έκβαση χρησιμοποιείται για να δηλώσει το πέρας της νόσου (π.χ. την ίαση, το θάνατο, την εμφάνιση καταλοίπων κ.ά.).*

Τιμές p

Ο Sir Ronald Aylmer Fisher* ήταν ο πρώτος ερευνητής που καθόρισε με σαφήνεια την έννοια της τιμής p (p value), ενώ παρείχε και τις απαιτούμενες μαθηματικές ισότητες για τον υπολογισμό της τιμής αυτής σ' ένα σημαντικό αριθμό περιπτώσεων.[6] Σημειώνεται ότι ο Fisher αρνούνταν την ερμηνεία της τιμής p ως σχετικής συχνότητας, σε αντίθεση με την πλειοψηφία των ερευνητών ακόμη και σήμερα, χωρίς όμως να έχει προβεί σε πειστικές απαντήσεις σε ορισμένα σημαντικά ερωτήματα, όπως:[7,8]

- Εάν η τιμή p δεν ερμηνευτεί ως σχετική συχνότητα εμφάνισης όλων των πιθανών εκβάσεων ενός πειράματος τύχης, τότε πώς μπορεί να ερμηνευτεί η αριθμητική της τιμή;
- Πώς θα μπορούσε να συνδυαστεί η τιμή p με άλλες πληροφορίες;
- Πώς θα μπορούσε να χρησιμοποιηθεί η τιμή p στην επαγωγική εξήγηση;
- Πώς θα μπορούσε να απορριφθεί η μηδενική υπόθεση, χωρίς την αποδοχή μιας εναλλακτικής;

Προτού γίνει εκτενής αναφορά στην έννοια της τιμής p, θεωρείται σκόπιμο να αναφερθεί, για ιστορικούς και όχι μόνο λόγους, το κλασικό παράδειγμα της ρίψης ενός νομίσματος.[8] Αναλυτικότερα, στο παράδειγμα αυτό η μηδενική υπόθεση (null hypothesis) είναι ότι το νόμισμα είναι «αμερόληπτο» (fair). «Αμερόληπτο» ονομάζεται το συμμετρικό και ομοιογενές νόμισμα στο οποίο, σε κάθε ρίψη, η πιθανότητα εμφάνισης «κεφαλής» ισούται με την πιθανότητα εμφάνισης «γραμμάτων». Όσες ρίψεις και αν πραγματοποιηθούν είναι ανεξάρτητες μεταξύ τους και σε κάθε ρίψη η πιθανότητα εμφάνισης «κεφαλής», όπως βέβαια και η πιθανότητα εμφάνισης «γραμμάτων», είναι ίση με το 1/2. Για τον έλεγχο της μηδενικής υπόθεσης πραγματοποιούνται 20 ρίψεις και καταγράφονται τα αποτελέσματα. Η ανάλυση των αποτελεσμάτων, σύμφωνα με το Fisher, πραγματοποιείται σε τέσσερα βήματα:

- Το πρώτο βήμα είναι η καταγραφή των πιθανών αποτελεσμάτων (δυνατών περιπτώσεων ή απλά ενδεχομένων) του πειράματος τύχης.** Η ρίψη του νομίσματος μπορεί να οδηγήσει στην εμφάνιση «κεφαλής» ή «γραμμάτων». Πραγματοποιούνται συνολικά 20 ρίψεις. Επομένως, οι πιθανοί συνδυασμοί είναι 2.[20] Στο συγκεκριμένο παράδειγμα, η τυχαία μεταβλητή είναι ο συνολικός αριθμός εμφανίσεων «κεφαλής» στις 20 ρίψεις. Η τυχαία αυτή μεταβλητή μπορεί να λάβει τιμές από 0 (καμιά εμφάνιση «κεφαλής») έως και 20 (εμφάνιση «κεφαλής» και στις 20 ρίψεις).
- Το δεύτερο βήμα είναι ο υπολογισμός της πιθανότητας κάθε αποτελέσματος της τυχαίας μεταβλητής, με την προϋπόθεση ότι η μηδενική υπόθεση είναι αληθής. Κάθε τυχαία μεταβλητή έχει μια αντίστοιχη κατανομή πιθανότητας. Μια *κατανομή πιθανότητας* (probability distribution) εφαρμόζει τη θεωρία των πιθανοτήτων για να περιγράψει τη συμπεριφορά μιας τυχαίας μεταβλητής. Στο συγκεκριμένο παράδειγμα, η κατανομή πιθανότητας της τυχαίας μεταβλητής προσδιορίζει όλα τα πιθανά αποτελέσματα της τυχαίας μεταβλητής, καθώς και την πιθανότητα να συμβεί το καθένα από αυτά. Η τυχαία μεταβλητή του παραδείγματος ακολουθεί τη διωνυμική κατανομή, καθώς είναι μια διχοτόμος τυχαία μεταβλητή που μπορεί να λάβει μία από δύο πιθανές τιμές. Εάν με p συμβολιστεί η πιθανότητα εμφάνισης «κεφαλής» και με q η πιθανότητα εμφάνισης «γραμμάτων», τότε η πιθανότητα να εμφανιστεί «κεφαλή» X φορές σε n ρίψεις δίνεται από την εξής ισότητα:

$$\mathrm{P}\,(\mathrm{X}=\times) = \binom{n}{x} p^x q^{n-x} \qquad (1)$$

Με δεδομένο ότι ισχύει η μηδενική υπόθεση, προκύπτει $p = q = 1/2$, ενώ n=20, καθώς πραγματοποιούνται συνολικά 20 ρίψεις. Έτσι, η ισότητα 1 γίνεται ως εξής:

$$\mathrm{P}\,(\mathrm{X}=\times) = \binom{20}{x} \frac{1}{2}^{x} \left(\frac{1}{2}\right)^{20-x}$$
$$\mathrm{P}\,(\mathrm{X}=\times) = \binom{20}{x} \left(\frac{1}{2}\right)^{20} \qquad (2)$$

Με $\binom{n}{x}$ συμβολίζεται ο συνδυασμός n αντικειμένων επιλεγμένων x τη φορά. Αντιπροσωπεύει τον αριθμό των τρόπων με τους οποίους x αντικείμενα μπορούν να επιλεγούν από ένα σύνολο n αντικειμένων όταν δεν έχει σημασία η σειρά τους. Με βάση την ισότητα (2) μπο-

* *Ο Sir Ronald Aylmer Fisher (1890–1962) ήταν Άγγλος μαθηματικός, βιολόγος και γενετιστής με τεράστια συνεισφορά στην ανάπτυξη και των τριών αυτών επιστημονικών πεδίων και δίκαια θεωρείται ως θεμελιωτής της σύγχρονης Στατιστικής και ως ένας από τους πλέον άξιους διαδόχους του Δαρβίνου. Ασχολήθηκε σε βάθος με όλα σχεδόν τα σημαντικά ζητήματα της Στατιστικής, εισάγοντας τις θεωρίες της ανάλυσης διασποράς, της μέγιστης πιθανοφάνειας, της τυχαιοποίησης, της σύγχυσης, της πολυμεταβλητής ανάλυσης, των μη παραμετρικών μεθόδων κ.ά. Μολονότι η οικογενειακή του ζωή χαρακτηρίστηκε από ηρεμία και ευτυχία, καθώς παντρεύτηκε σε ηλικία 27 ετών, αποκτώντας 9 παιδιά, η επαγγελματική του πορεία ήταν ιδιαίτερα έντονη με συνεχείς διαμάχες και συγκρούσεις, κυρίως με τους Karl Pearson, Egon Pearson και Jersy Neyman.*

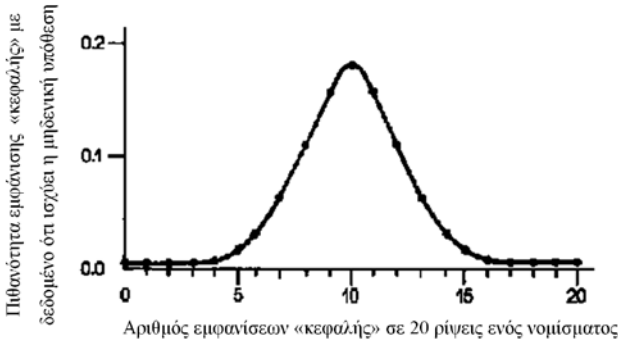
** *Πείραμα τύχης είναι το πείραμα εκείνο που, μολονότι εκτελείται κάτω από τις ίδιες συνθήκες, δεν οδηγεί πάντοτε στο ίδιο αποτέλεσμα.*

ρούν να υπολογιστούν πλέον οι πιθανότητες εμφάνισης οποιουδήποτε αριθμού «κεφαλής» από 0–20, με την προϋπόθεση ότι ισχύει η μηδενική υπόθεση. Οι πιθανότητες αυτές φαίνονται στον πίνακα 1 και απεικονίζονται διαγραμματικά στην εικόνα 1.

- Το τρίτο βήμα είναι η καταγραφή όλων των αποτελεσμάτων που θα μπορούσαν να συμβούν και τα οποία, με δεδομένο ότι ισχύει η μηδενική υπόθεση, έχουν μια πιθανότητα μικρότερη ή ίση από την πιθανότητα του αποτελέσματος που όντως προέκυψε. Εάν, π.χ., στις 20 ρίψεις εμφανιστεί «κεφαλή» 4 φορές, τότε η πιθανότητα να συμβεί το αποτέλεσμα αυτό, σύμφωνα με τον πίνακα 1, είναι 0,0046. Τα αποτελέσματα που έχουν πιθανότητα ≤0,0046 είναι εκείνα για τα οποία $x = 4,3,2,1,0$ και $x = 16,17,18,19,20$. Εφόσον τα ενδεχόμενα αυτά είναι αμοιβαία αποκλειόμενα, τότε η πιθανότητα να προκύψει ένα από αυτά ισούται με το άθροισμα των επιμέρους πιθανοτήτων, οπότε:
  $P^* = 2\times(0,0046+0,0011+2\times10^{-4}+1,9\times10^{-5}+9\times10^{-7})$
  $P^* = 0,012$
- Το τέταρτο βήμα αποτελεί ουσιαστικά μια παραδοχή, σύμφωνα με την οποία η μηδενική υπόθεση θα πρέπει να απορρίπτεται όταν $P^* \leq 0,05$. Ορισμένοι ερευνητές, ωστόσο, προτείνουν η μηδενική υπόθεση να απορρίπτεται όταν $P^* \leq 0,01$ ή $P^* \leq 0,001$. Η τιμή 0,05 που πρότεινε ο Fisher είναι το προκαθορισμένο επίπεδο στατιστικής σημαντικότητας, το οποίο ορίζεται από τους ερευνητές και είναι γνωστό ως τιμή α. Εάν το πείραμα οδηγήσει σε $P^* \leq \alpha$, τότε το αποτέλεσμα ονομάζεται στατιστικά σημαντικό σε επίπεδο σημαντικότητας ίσο με α και η μηδενική υπόθεση απορρίπτεται.

Πίνακας 1. Πιθανότητες εμφάνισης «κεφαλής» Χ φορές σε ένα πείραμα τύχης στο οποίο πραγματοποιούνται 20 ρίψεις ενός «αμερόληπτου» νομίσματος.

Αριθμός εμφανίσεων κεφαλής (x)	Πιθανότητα εμφάνισης (P)
0	9×10^{-7}
1	$1,9\times10^{-5}$
2	2×10^{-4}
3	0,0011
4	0,0046
5	0,0148
6	0,0370
7	0,0739
8	0,1201
9	0,1602
10	0,1762
11	0,1602
12	0,1201
13	0,0739
14	0,0370
15	0,0148
16	0,0046
17	0,0011
18	2×10^{-4}
19	$1,9\times10^{-5}$
20	9×10^{-7}



Εικόνα 1. Διαγραμματική απεικόνιση των πιθανοτήτων εμφάνισης «κεφαλής» Χ φορές σε ένα πείραμα τύχης στο οποίο πραγματοποιούνται 20 ρίψεις ενός «αμερόληπτου» νομίσματος.

Στο συγκεκριμένο παράδειγμα, εμφανίστηκε 4 φορές «κεφαλή» στις 20 ρίψεις, οπότε το P^* που προκύπτει είναι ίσο με 0,012. Εφόσον το P^* είναι μικρότερο από την τιμή α απορρίπτεται η μηδενική υπόθεση στο επίπεδο σημαντικότητας ίσο με 0,05. Εάν όμως εμφανίζονταν 6 φορές «κεφαλή» σε 20 ρίψεις, τότε το υπολογιζόμενο P^* θα ήταν 0,115. Στην περίπτωση αυτή, η μηδενική υπόθεση δεν θα απορριπτόταν στο επίπεδο σημαντικότητας ίσο με 0,05.

Το παραπάνω παράδειγμα είναι ιδιαίτερα διευκρινιστικό όσον αφορά στην έννοια του παρατηρούμενου επιπέδου στατιστικής σημαντικότητας ή, απλούστερα, της τιμής p. Στην πράξη, ωστόσο, η διαδικασία αυτή είναι περισσότερο σύνθετη. Σε μια μελέτη, π.χ., διερευνάται αν υπάρχει διαφορά στο μέσο επίπεδο νοημοσύνης ανάμεσα σε μια ομάδα αγοριών και σε μια ομάδα κοριτσιών. Το πρόβλημα στην περίπτωση αυτή είναι ότι ο ερευνητής δε γνωρίζει τις κατανομές του επιπέδου νοημοσύνης στους πληθυσμούς των δύο ομάδων. Σύμφωνα με τη μηδενική υπόθεση, δεν υπάρχει διαφορά στο μέσο επίπεδο νοημοσύνης μεταξύ των πληθυσμών αγοριών και κοριτσιών. Στο παράδειγμα με τη ρίψη του νομίσματος δεν υπάρχει ιδιαίτερη δυσκολία, καθώς η μελετώμενη τυχαία μεταβλητή ακολουθεί τη διωνυμική κατανομή. Δεν συμβαίνει όμως το ίδιο και στην προαναφερθείσα μελέτη, καθώς

δεν είναι γνωστή η κατανομή πιθανότητας του επιπέδου νοημοσύνης στους πληθυσμούς αγοριών και κοριτσιών. Το γεγονός αυτό αποτελεί σημαντικό πρόβλημα, καθώς οι έλεγχοι σημαντικότητας μπορούν να πραγματοποιηθούν μόνο όταν οι κατανομές πιθανότητας των μεταβλητών είναι καθορισμένες και γνωστές, κάτι το οποίο συμβαίνει σπάνια. Όταν, ωστόσο, τα μελετώμενα «δείγματα» είναι αρκετά μεγάλα, τότε μπορεί να χρησιμοποιηθεί ο στατιστικός έλεγχος t (Student's t-test), οπότε διατυπώνεται η κατάλληλη μηδενική υπόθεση και υπολογίζεται η τιμή της ποσότητας t με βάση τα δεδομένα της μελέτης. Σε άλλες περιπτώσεις, εξάλλου, μπορεί να χρησιμοποιηθεί ο έλεγχος z, ο έλεγχος x^2 κ.ά.

Η τιμή p προτάθηκε από τον Fisher ως ένα μέτρο του μεγέθους της ένδειξης* –ή, αλλιώς, της πληροφορίας– που προέρχεται από τα δεδομένα μιας μελέτης, στηριζόμενοι στην ερμηνεία της πιθανότητας ως σχετικής συχνότητας γεγονότων.[7] Ο Fisher επεδίωκε την εύρεση μιας αντικειμενικής ποσοτικής μεθόδου για την πραγματοποίηση του επαγωγικού διαλογισμού,** έτσι ώστε να προκύψουν αξιόπιστα συμπεράσματα για το φυσικό κόσμο (πραγματικότητα) με βάση τις εμπειρικές παρατηρήσεις (δεδομένα μιας συγκεκριμένης μελέτης).[9]

Ο Fisher δεν ήταν ο πρώτος ερευνητής που χρησιμοποίησε την τιμή p,[10,11] αλλά ήταν ο πρώτος που καθόρισε τη λογική της εφαρμογής της, ενώ παρείχε και τις απαιτούμενες μαθηματικές ισότητες για τον υπολογισμό της τιμής p σ' ένα σημαντικό αριθμό περιπτώσεων. Ο Fisher όρισε την τιμή p, όπως ακριβώς ορίζεται και σήμερα. Πιο συγκεκριμένα, η τιμή p υπολογίζεται με την προϋπόθεση ότι ισχύει η μηδενική υπόθεση σύμφωνα με την οποία –στην πλειονότητα των περιπτώσεων– δεν υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης.[12] Η τιμή p είναι η πιθανότητα, με δεδομένο ότι η μηδενική υπόθεση είναι αληθής, να προκύψει ένα αποτέλεσμα τόσο ακραίο ή πιο ακραίο από αυτό που πραγματικά παρατηρήθηκε σε μια συγκεκριμένη μελέτη. Σε μια μελέτη, π.χ., διερεύνησης της σχέσης μεταξύ της καπνισματικής συνήθειας και της συστολικής αρτηριακής πίεσης βρέθηκε ότι η μέση πίεση στους καπνιστές ήταν μεγαλύτερη κατά 20 mmHg σε σχέση με τους μη καπνιστές. Η μηδενική υπόθεση ήταν ότι η μέση πίεση στον πληθυσμό των καπνιστών είναι ίση με τη μέση πίεση στον πληθυσμό των μη καπνιστών. Ανάλογα με την κατανομή πιθανότητας της συστολικής αρτηριακής πίεσης στους δύο πληθυσμούς (καπνιστών και μη) χρησιμοποιείται το κατάλληλο στατιστικό μοντέλο για τον έλεγχο της μηδενικής υπόθεσης, οπότε προκύπτει μια τιμή p. Η εν λόγω τιμή p είναι η πιθανότητα, με δεδομένο ότι δεν υπάρχει σχέση μεταξύ καπνισματικής συνήθειας και συστολικής αρτηριακής πίεσης (με δεδομένο δηλαδή ότι η διαφορά στις μέσες τιμές των δύο πληθυσμών είναι ίση με μηδέν), να προκύψει (σε μια οποιαδήποτε παρόμοια μελέτη) διαφορά στις μέσες τιμές ≥20 mmHg. Τονίζεται ότι η τιμή p δεν είναι η πιθανότητα ότι η μηδενική υπόθεση είναι αληθής, αλλά πως υπολογίζεται με την προϋπόθεση ότι η μηδενική υπόθεση είναι αληθής.

Ο Fisher πρότεινε την τιμή p ως ένα «ανεπίσημο» μέτρο της ασυμφωνίας μεταξύ των δεδομένων μιας μελέτης και της μηδενικής υπόθεσης.[6,14] Η τιμή p σε καμιά περίπτωση δεν αποτελεί συστατικό του τυπικού διαλογισμού και χρησιμοποιείται λανθασμένα από τους περισσότερους επιστήμονες υγείας για την εξαγωγή συμπερασμάτων σχετικά με την ύπαρξη ή όχι σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Σύμφωνα με το Fisher, η τιμή p δεν πρέπει να ερμηνεύεται ως η υποθετική σχετική συχνότητα του σφάλματος έπειτα από πολλές επαναλήψεις του πειράματος τύχης. Στις επιστήμες υγείας, το πείραμα τύχης είναι αντίστοιχο με τη μελέτη διερεύνησης της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Ο Fisher επισήμανε ότι η τιμή p αποτελεί μέτρο της ένδειξης που παρέχει μια μελέτη (ή ένα πείραμα), εκφράζοντας παράλληλα την αξιοπιστία της μηδενικής υπόθεσης σε σχέση με τα δεδομένα της συγκεκριμένης μελέτης. Έτσι, εάν η τιμή p που προκύπτει από τα δεδομένα μιας μελέτης είναι μικρότερη από το προκαθορισμένο (από τους ερευνητές) επίπεδο στατιστικής σημαντικότητας (ή, αλλιώς, τιμή α), τότε «απορρίπτεται» η μηδενική υπόθεση με την έννοια όμως ότι υπάρχει σημαντική ασυμφωνία ανάμεσα στη μηδενική υπόθεση και τα δεδομένα της συγκεκριμένης μελέτης και όχι ότι υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Όσο μικρότερη είναι η τιμή p τόσο μεγαλύτερη είναι η ασυμφωνία ανάμε-

---

** Σημασιολογικά, η έννοια ένδειξη είναι συνυφασμένη με την πληροφορία.[13] Όπως και η πληροφορία, έτσι και η ένδειξη, μπορεί να οριστεί ως η συλλογή δεδομένων, τα οποία, εφόσον συλλεχθούν με τον κατάλληλο τρόπο, στον κατάλληλο χρόνο και χρησιμοποιηθούν στο κατάλληλο πλαίσιο, βελτιώνουν τη γνώση εκείνου που λαμβάνει την απόφαση, κατά τέτοιον τρόπο, ώστε να τον καθιστούν ικανότερο να λαμβάνει τις βέλτιστες αποφάσεις. Οι ενδείξεις δημιουργούνται έπειτα από την αναζήτηση στη σχετική βιβλιογραφία και την κριτική της αξιολόγηση και εφόσον έχουν αποδειχθεί έγκυρες, σημαντικές και εφαρμόσιμες – με βάση συγκεκριμένα κριτήρια– χρησιμοποιούνται για τη λήψη της βέλτιστης δυνατής απόφασης σ' ένα συγκεκριμένο πάσχοντα.*

*** Ο επαγωγικός διαλογισμός καλείται να απαντήσει στην ερώτηση «με βάση την ένδειξη –ή πληροφορία ή αποτέλεσμα– που προέρχεται από μια συγκεκριμένη μελέτη, ποια υπόθεση είναι περισσότερο πιθανή;». Ουσιαστικά, πρόκειται για τον υπολογισμό μιας πιθανοφάνειας στον οποίο κατέληξε πριν περίπου 300 χρόνια ο Thomas Bayes.*

σα στη μηδενική υπόθεση και τα δεδομένα μιας μελέτης. Αυτή η διαδικασία της αυθαίρετης επιλογής ενός ορίου (τιμής α) για την απόρριψη ή όχι της μηδενικής υπόθεσης, είναι γνωστή ως *«έλεγχος σημαντικότητας»* (significance test) και πρέπει να διακρίνεται από τον *«έλεγχο υπόθεσης»* (hypothesis test) που προτάθηκε από τους Jerzy Neyman* και Egon Pearson** και θα αναλυθεί στη συνέχεια. Ο Fisher τόνιζε ότι εάν χρησιμοποιηθεί μια τιμή α για τον έλεγχο της στατιστικής σημαντικότητας και την απόρριψη ή όχι της μηδενικής υπόθεσης, τότε η επιλογή της τιμής α πρέπει να γίνεται με ιδιαίτερη περίσκεψη, λαμβάνοντας σοβαρά υπόψη και την προϋπάρχουσα γνώση σχετικά με τη σχέση που διερευνάται.[14]

Σημειώνεται ότι στον έλεγχο σημαντικότητας που συμπεριλαμβάνεται στη θεωρία του Fisher γίνεται αναφορά μόνο στη μηδενική υπόθεση, χωρίς να λαμβάνεται υπόψη η εναλλακτική υπόθεση, ενώ στον έλεγχο υπόθεσης των Neyman και Pearson περιλαμβάνονται τόσο η μηδενική όσο και η εναλλακτική υπόθεση.

Σε αρκετές περιπτώσεις, οι επιστήμονες υγείας παρερμηνεύουν την τιμή p, καταλήγοντας σε παραπλανητικά συμπεράσματα.[15–18] Αρκετοί θεωρούν ότι μια τιμή p=0,05 σημαίνει ότι η πιθανότητα να ισχύει η μηδενική υπόθεση είναι μόλις 5%. Πρόκειται για μια τελείως λανθασμένη αντίληψη, καθώς η τιμή p δεν είναι η πιθανότητα να ισχύει η μηδενική υπόθεση, αλλά υπολογίζεται με δεδομένο ότι ισχύει η μηδενική υπόθεση. Το λάθος αυτό σχετίζεται άμεσα και με τη λανθασμένη άποψη ότι με βάση τα δεδομένα μιας μελέτης μπορεί να υπολογιστεί η πιθανότητα να είναι αληθής μια υπόθεση. Η πιθανότητα να ισχύει η μηδενική υπόθεση με βάση την ένδειξη που παρέχεται από μια μελέτη μπορεί να υπολογιστεί μόνο με την εφαρμογή του θεωρήματος του Bayes και όχι βεβαίως με τη χρήση των τιμών p ή των ελέγχων υποθέσεων.[16,20]

Ορισμένοι ερευνητές στράφηκαν εξαρχής εναντίον της θεωρίας του Fisher σχετικά με την τιμή p, θεωρώντας ότι στερείται τόσο λογικής βάσης όσο και πρακτικής χρησιμότητας.[21,22] Το πιο ισχυρό σημείο της κριτικής τους ήταν ότι η τιμή p αποτελεί απλά ένα μέτρο της ένδειξης που παρέχει μια μελέτη χωρίς να λαμβάνεται υπόψη το μέγεθος της σχέσης που προκύπτει από τα δεδομένα της μελέτης. Μια σχέση μικρού μεγέθους σε μια μελέτη με μεγάλο μέγεθος «δείγματος» μπορεί να οδηγήσει στην ίδια τιμή p με μια σχέση μεγάλου μεγέθους σε μια μελέτη με μικρό μέγεθος «δείγματος». Για το λόγο αυτό, τα τελευταία έτη καταβάλλεται συστηματική προσπάθεια ευρύτερης χρησιμοποίησης των διαστημάτων εμπιστοσύνης έναντι των τιμών p.[12,23]

Έλεγχοι υποθέσεων

Μηδενική και εναλλακτική υπόθεση

Οι μαθηματικοί Jerzy Neyman και Egon Pearson προσπάθησαν να επιλύσουν τα προβλήματα που παρουσίαζε η θεωρία του Fisher σχετικά με την τιμή p και τον έλεγχο σημαντικότητας, εισάγοντας την έννοια της εναλλακτικής υπόθεσης και του σφάλματος τύπου II.[24] Η θεωρία των Neyman και Pearson έγινε ευρέως γνωστή με τον όρο «έλεγχος υπόθεσης». Σ' έναν έλεγχο υπόθεσης, οι ερευνητές πρέπει να καθορίσουν τη μηδενική και την εναλλακτική υπόθεση, καθώς επίσης και τις τιμές που θα λάβουν τα σφάλματα τύπου I και II.[25] Συνήθως, η μηδενική υπόθεση –γνωστή και ως υπόθεση της μη διαφοράς– υποστηρίζει ότι δεν υπάρχει σχέση ανάμεσα στον προσδιοριστή και τη συχνότητα εμφάνισης της έκβασης και διατυπώνεται με σκοπό να αναιρεθεί. Η συμπληρωματική της μηδενικής υπόθεσης ονομάζεται εναλλακτική υπόθεση (alternative hypothesis) και σύμφωνα με αυτή υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Στην περίπτωση αυτή, ο έλεγχος υπόθεσης ονομάζεται έλεγχος διπλής κατεύθυνσης (two-sides test). Με βάση το στατιστικό έλεγχο που πραγματοποιείται, η μηδενική υπόθεση είτε απορρίπτεται είτε όχι. Εάν η τιμή p που προκύπτει με βάση τα δεδομένα μιας μελέτης είναι μικρότερη από την τιμή α, τότε απορρίπτεται η μηδενική υπόθεση, ενώ εάν η τιμή p είναι μεγαλύτερη από την τιμή α, τότε δεν απορρίπτεται η μηδενική υπόθεση. Σε καμιά όμως περίπτωση τα δεδομένα μιας μελέτης δεν μπορούν

---

* *Ο πολωνικής καταγωγής Jersy Neyman (1894–1981) γεννήθηκε στη Ρωσία και πέθανε στις ΗΠΑ, όντας ένας από τους θεμελιωτές της σύγχρονης Στατιστικής. Πραγματοποίησε τις βασικές του σπουδές στη Ρωσία και την Πολωνία, ενώ το χρονικό διάστημα 1925–1927 σπούδασε μαθηματικά σε πανεπιστήμια στο Λονδίνο και στο Παρίσι, όπου και συνδέθηκε με βαθιά φιλία με τον Egon Pearson. Η συνεργασία μεταξύ Neyman και Pearson ήταν κομβικής σημασίας στην ιστορία της Στατιστικής, καθώς ανέπτυξαν τη θεωρία του ελέγχου στατιστικών υποθέσεων. Αξίζει να σημειωθεί ότι ο Neyman μόλις το 1937, και ενώ βρισκόταν ακόμη στην Πολωνία, ανέπτυξε τη θεωρία της εκτίμησης των διαστημάτων εμπιστοσύνης.[19] Το 1938 αποδέχθηκε τη θέση του καθηγητή μαθηματικών στο Πανεπιστήμιο της Καλιφόρνια, στο Berkeley, δημιουργώντας ένα από τα κορυφαία κέντρα διδασκαλίας της Στατιστικής παγκόσμια και διοργανώνοντας παράλληλα συνέδρια με τη συμμετοχή κορυφαίων ερευνητών.*

** *Ο Άγγλος Egon Sharp Pearson (1895–1980) ήταν ο μοναδικός υιός του Karl Pearson και ακολουθώντας τα χνάρια του πατέρα του υπήρξε ένας εξαιρετικός στατιστικός. Διετέλεσε καθηγητής Στατιστικής στο University College, στο Λονδίνο, ενώ εξαιρετικά σημαντική ήταν η συνεισφορά του στην καθιέρωση του περιοδικού "Biometrika" ως ένα από τα πλέον έγκριτα περιοδικά Στατιστικής. Τα επιστημονικά ενδιαφέροντα του Pearson επικεντρώθηκαν στους ελέγχους στατιστικών υποθέσεων και στην εφαρμογή των στατιστικών μεθόδων στη βιομηχανία και ιδιαίτερα στην κατασκευή μοντέλων.*

να προσφέρουν απόδειξη ότι η μηδενική υπόθεση είναι αληθής. Αν δεν απορριφθεί η μηδενική υπόθεση, τότε ισχύει ότι τα δεδομένα της μελέτης στα οποία στηρίζεται ο έλεγχος υπόθεσης δεν επαρκούν για την απόρριψή της. Τονίζεται και πάλι ότι ο στατιστικός έλεγχος των υποθέσεων δεν οδηγεί στην απόδειξη της μηδενικής υπόθεσης, αλλά απλά παρέχει την πληροφορία εάν τα δεδομένα μιας μελέτης στηρίζουν την υπόθεση αυτή.

Εάν είναι γνωστό πριν από τη διεξαγωγή μιας μελέτης, ότι υπάρχει θετική ή αρνητική σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, τότε ο έλεγχος υπόθεσης μπορεί να είναι μονής κατεύθυνσης (one-side). Στην περίπτωση αυτή, εάν, π.χ., διερευνάται η σχέση μεταξύ καπνισματικής συνήθειας και συχνότητας εμφάνισης καρκίνου του πνεύμονα, τότε η μηδενική υπόθεση είναι ότι το κάπνισμα μειώνει τη συχνότητα εμφάνισης του καρκίνου, ενώ η εναλλακτική υπόθεση είναι ότι το κάπνισμα αυξάνει τη συχνότητα αυτή.

Σε μια μελέτη διερεύνησης της σχέσης μεταξύ της χρήσης κινητών τηλεφώνων (προσδιοριστής) και της συχνότητας εμφάνισης όγκου στον εγκέφαλο (έκβαση), ο έλεγχος υπόθεσης διπλής κατεύθυνσης περιλαμβάνει τη μηδενική υπόθεση ότι δεν υπάρχει σχέση μεταξύ της χρήσης κινητών τηλεφώνων και της συχνότητας εμφάνισης όγκου στον εγκέφαλο και την εναλλακτική υπόθεση ότι υπάρχει σχέση μεταξύ προσδιοριστή και έκβασης, χωρίς όμως να καθορίζεται αν η συγκεκριμένη σχέση είναι θετική ή αρνητική. Εάν πραγματοποιηθεί έλεγχος υπόθεσης μονής κατεύθυνσης, τότε σύμφωνα με τη μηδενική υπόθεση η χρήση κινητών μειώνει τη συχνότητα εμφάνισης όγκου στον εγκέφαλο, ενώ σύμφωνα με την εναλλακτική υπόθεση η χρήση κινητών αυξάνει τη συχνότητα εμφάνισης όγκου στον εγκέφαλο, καθώς θεωρείται ότι η συχνότητα εμφάνισης όγκου στον εγκέφαλο είναι μεγαλύτερη σ' εκείνους που χρησιμοποιούν κινητά. Ανάλογα με τα δεδομένα της μελέτης πραγματοποιείται ο κατάλληλος στατιστικός έλεγχος και απορρίπτεται ή όχι η μηδενική υπόθεση. Αν απορριφθεί η μηδενική υπόθεση, τότε σύμφωνα με τα δεδομένα της μελέτης αυτής υπάρχει σχέση μεταξύ της χρήσης κινητών και της συχνότητας εμφάνισης όγκου στον εγκέφαλο. Αντίθετα, αν δεν απορριφθεί η μηδενική υπόθεση, τότε δε σημαίνει ότι δεν υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Σημαίνει απλά ότι τα δεδομένα της μελέτης αυτής δεν πρόσφεραν ένδειξη για την απόρριψη της μηδενικής υπόθεσης. Είναι πιθανό τα δεδομένα μιας άλλης μελέτης να οδηγήσουν σε απόρριψη της μηδενικής υπόθεσης.

Συχνά δεν αναφέρεται καν η ακριβής τιμή p, παρά μόνο αν η τιμή αυτή είναι μεγαλύτερη ή μικρότερη από την τιμή α.[12,23] Δεν υπάρχει κάποιος λόγος βέβαια ένα συνεχές μέτρο, όπως είναι η τιμή p, να μετατρέπεται σε διχοτόμο, γιατί έτσι χάνεται πολύτιμη πληροφορία. Για το λόγο αυτό, πρέπει να αναφέρεται η ακριβής τιμή p και όχι απλά αν είναι μεγαλύτερη ή μικρότερη από την τιμή α.

Παρά το γεγονός ότι, ακόμη και σήμερα, πολλοί επιστήμονες υγείας χρησιμοποιούν τις τιμές p για την εξαγωγή συμπερασμάτων, είναι φανερό ότι οι τιμές p αποτελούν απλά ένα συνεχές μέτρο της συμβατότητας μεταξύ της μηδενικής υπόθεσης και των δεδομένων μιας μελέτης.[12,23] Αυθαίρετο κριτήριο της συμβατότητας αυτής αποτελεί η τιμή α που καθορίζεται από τους ερευνητές. Έτσι, αν η τιμή α καθοριστεί ως 0,05, τότε οποιαδήποτε τιμή $p < 0,05$ υποδηλώνει μικρή συμβατότητα της μηδενικής υπόθεσης με τα δεδομένα μιας μελέτης. Όσο μικρότερη δηλαδή είναι η τιμή p τόσο μικρότερη είναι η συμβατότητα μεταξύ της μηδενικής υπόθεσης και των δεδομένων μιας μελέτης.

Συμπερασματικά, η χρήση των τιμών p για την εξαγωγή ασφαλών συμπερασμάτων σχετικά με τον έλεγχο υποθέσεων είναι εξαιρετικά επισφαλής και πρέπει να εγκαταλειφθεί. Η τιμή p ενός στατιστικού ελέγχου δεν πρέπει να αποτελεί κριτήριο για το αν η μηδενική υπόθεση είναι αληθής ή όχι, καθώς απαιτείται η διεξαγωγή πολλών μελετών και η σύγκριση των αποτελεσμάτων που προκύπτουν. Η τιμή p είναι απλά ένα μέτρο της συμβατότητας μεταξύ της μηδενικής υπόθεσης και των δεδομένων μιας μελέτης και χρησιμοποιείται για την απόρριψη ή μη της μηδενικής υπόθεσης. Βέβαια, η μη απόρριψη της μηδενικής υπόθεσης δε σημαίνει ότι είναι αληθής, αλλά ότι τα δεδομένα μιας συγκεκριμένης μελέτης δεν προσφέρουν αρκετή απόδειξη για την απόρριψή της.

Η τιμή p εξάλλου δεν εκφράζει το μέγεθος της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης και την ακρίβεια (απουσία τυχαίου σφάλματος) της μέτρησης που πραγματοποιείται σε μια μελέτη. Για την καλύτερη ερμηνεία των αποτελεσμάτων μιας μελέτης είναι αναγκαίο να καθορίζεται με σαφήνεια τόσο το μέγεθος της σχέσης μεταξύ προσδιοριστή και έκβασης όσο και η ακρίβεια της μελέτης κάτι που επιτυγχάνεται με τη διαδικασία της εκτίμησης, όπως θα αναλυθεί ακολούθως.

Σφάλματα τύπου I και II

Όπως προαναφέρθηκε, ο έλεγχος υποθέσεων προϋποθέτει τον καθορισμό της μηδενικής και της εναλλακτικής υπόθεσης, αλλά και τον προσδιορισμό του σφάλματος τύπου I και II (πίν. 2).[1,3,12,26,27]

Το σφάλμα τύπου I (type I error) ή, αλλιώς, *σφάλμα απόρριψης* (rejection error) ή σφάλμα α (α error) συμβαίνει όταν απορρίπτεται η μηδενική υπόθεση, ενώ είναι

Πίνακας 2. Σύγκριση των αποτελεσμάτων ενός ελέγχου υπόθεσης και της πραγματικότητας που αφορά στον πληθυσμό.

		Πραγματικότητα	
		Αληθής μηδενική υπόθεση	**Ψευδής μηδενική υπόθεση**
Αποτελέσματα ελέγχου υπόθεσης	Απόρριψη μηδενικής υπόθεσης	Σφάλμα τύπου I (ψευδώς θετικά αποτελέσματα)	Ισχύς (αληθώς θετικά αποτελέσματα)
	Μη απόρριψη μηδενικής υπόθεσης	Σωστό (αληθώς αρνητικά αποτελέσματα)	Σφάλμα τύπου II (ψευδώς αρνητικά αποτελέσματα)

αληθής. Ουσιαστικά, το σφάλμα τύπου I είναι το ποσοστό των ψευδώς θετικών αποτελεσμάτων των ελέγχων υποθέσεων, όπου λανθασμένα συμπεραίνεται ότι υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, καθώς στην πραγματικότητα η σχέση αυτή δεν υφίσταται. Η πιθανότητα να διαπραχθεί ένα σφάλμα τύπου I καθορίζεται από το προκαθορισμένο επίπεδο σημαντικότητας του ελέγχου υπόθεσης που είναι η τιμή α. Εάν πραγματοποιηθούν κατ' επανάληψη έλεγχοι υπόθεσης διατηρώντας το επίπεδο σημαντικότητας στο 0,05, θα απορριφθεί λανθασμένα η μηδενική υπόθεση, ενώ ισχύει, 5 φορές στους 100 ελέγχους. Σε ορισμένες περιπτώσεις επιλέγεται η τιμή α να είναι ίση με 0,01, οπότε απορρίπτεται λανθασμένα η μηδενική υπόθεση, ενώ είναι αληθής, μόνο μία φορά στους 100 ελέγχους υπόθεσης.

Το σφάλμα τύπου II (type II error) ή, αλλιώς, *σφάλμα αποδοχής* (acceptance error) ή σφάλμα β (β error) συμβαίνει όταν δεν απορρίπτεται η μηδενική υπόθεση, ενώ είναι λανθασμένη. Εάν, π.χ., β=0,10, τότε η πιθανότητα μη απόρριψης της μηδενικής, ενώ είναι λανθασμένη, είναι 0,10 ή 10%. Ουσιαστικά, το σφάλμα τύπου II είναι το ποσοστό των ψευδώς αρνητικών αποτελεσμάτων των ελέγχων υποθέσεων, όπου λανθασμένα συμπεραίνεται ότι δεν υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, καθώς στην πραγματικότητα η σχέση αυτή υφίσταται.

Εάν το β είναι η πιθανότητα να διαπραχθεί ένα σφάλμα τύπου II, τότε 1 – β είναι η *στατιστική ισχύς* (statistical power) του ελέγχου υπόθεσης. Η ισχύς είναι η πιθανότητα να απορριφθεί η μηδενική υπόθεση, ενώ είναι λανθασμένη ή, αλλιώς, είναι η πιθανότητα να αποφευχθεί ένα σφάλμα τύπου II. Η ισχύς μπορεί, επίσης, να θεωρηθεί ως η πιθανότητα μια συγκεκριμένη μελέτη να διακρίνει μια απόκλιση από τη μηδενική υπόθεση δεδομένου ότι αυτή υπάρχει. Ουσιαστικά, η ισχύς είναι το ποσοστό των αληθώς θετικών αποτελεσμάτων των ελέγχων υποθέσεων, όπου σωστά συμπεραίνεται ότι υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, καθώς στην πραγματικότητα υφίσταται η συγκεκριμένη σχέση.

Σε γενικές γραμμές, ο στόχος των ερευνητών είναι ο σχεδιασμός ελέγχων υποθέσεων που έχουν υψηλή ισχύ. Δεν αρκεί να υπάρχει μικρή πιθανότητα να απορριφθεί η μηδενική υπόθεση όταν είναι αληθής. Πρέπει να υπάρχει μεγάλη πιθανότητα να απορριφθεί η μηδενική υπόθεση όταν είναι λανθασμένη. Ένας τρόπος να αυξηθεί η ισχύς ενός ελέγχου είναι να αυξηθεί η τιμή α. Αύξηση της τιμής α προκαλεί μείωση του σφάλματος τύπου II, αλλά συγχρόνως αυξάνεται το σφάλμα τύπου I. Αντίστροφα, μείωση της τιμής α προκαλεί μείωση του σφάλματος τύπου I και αύξηση του σφάλματος τύπου II.

Κριτική

Στον έλεγχο υπόθεσης των Neyman και Pearson δεν υπάρχει κάποιο μέτρο της ένδειξης που παρέχεται από τα δεδομένα μιας μελέτης, καθώς η τιμή p που προκύπτει χρησιμοποιείται απλά για την απόρριψη ή όχι της μηδενικής υπόθεσης. Η διαφορά αυτή με τον έλεγχο σημαντικότητας του Fisher είναι εξαιρετικά σημαντική, καθώς η θεωρία των Neyman και Pearson απορρίπτει ουσιαστικά κάθε προσπάθεια επαγωγικού διαλογισμού, γεγονός που επιβεβαιώνεται άλλωστε και από τους ίδιους τους ερευνητές.[24] Πάντως, και στις δύο μεθόδους, η τιμή p υπολογίζεται με την προϋπόθεση ότι ισχύει η μηδενική υπόθεση.

Είναι αξιοσημείωτο το γεγονός ότι οι Neyman και Pearson στην προσπάθειά τους να περιοριστεί όσο το δυνατόν περισσότερο η χρήση της τιμής p οδήγησαν στην ακριβώς αντίθετη κατεύθυνση, καθιστώντας ουσιαστικά την τιμή p κριτήριο για τη λήψη αποφάσεων έπειτα από τη σύγκρισή της με την τιμή α.[7] Η μαθηματική και η εννοιολογική προσέγγιση των ελέγχων υποθέσεων αποτέλεσε ένα σημαντικό βήμα, αλλά η ερμηνεία τους και η πρακτική τους εφαρμογή αντιμετωπίζουν ακόμη και σήμερα σοβαρά προβλήματα. Οι Neyman και Pearson δε χρησιμοποίησαν κάποιο μέτρο της ένδειξης που παρέχουν τα δεδομένα μιας μελέτης. Έτσι, απέρριψαν ουσιαστικά την

εφαρμογή του επαγωγικού διαλογισμού σε κάθε μελέτη ξεχωριστά για την εξαγωγή συμπερασμάτων και χρησιμοποίησαν παραγωγικές μεθόδους για να περιορίσουν το μέγεθος του σφάλματος έπειτα από την επανάληψη της ίδιας μελέτης. Σύμφωνα με τους Neyman και Pearson, κανένας έλεγχος δεν μπορεί να παρέχει αξιόπιστη ένδειξη της αλήθειας ή του ψεύδους μιας υπόθεσης και γι' αυτό πρέπει να αναζητηθούν οι αρχές εκείνες με βάση τις οποίες θα λαμβάνονται οι αποφάσεις, που μακροπρόθεσμα θα οδηγήσουν σε μικρό μέγεθος σφάλματος.[28] Οι έλεγχοι υποθέσεων είναι ισοδύναμοι ουσιαστικά, με ένα σύστημα δικαιοσύνης, που δεν επικεντρώνεται στο αν ένας συγκεκριμένος κατηγορούμενος βρεθεί ένοχος ή αθώος (αντίστοιχα, στις επιστήμες υγείας, εάν μια υπόθεση σε μια συγκεκριμένη μελέτη βρεθεί αληθής ή ψευδής), αλλά επιδιώκει να περιορίσει, μακροπρόθεσμα, όσο το δυνατόν περισσότερο τις εσφαλμένες ετυμηγορίες. Ο περιορισμός του σφάλματος μακροπρόθεσμα είναι θεμιτός και πρέπει να επιδιώκεται πάντοτε, αλλά όπως σε κάθε δικαστική διαμάχη το αίσθημα της δικαιοσύνης υπαγορεύει τη σωστή ετυμηγορία για ένα συγκεκριμένο κατηγορούμενο, έτσι ακριβώς και στις επιστήμες υγείας απαιτείται συστηματική προσπάθεια για την εξαγωγή σωστών συμπερασμάτων με βάση την ένδειξη που παρέχεται από κάθε μελέτη ξεχωριστά.

Οι Neyman και Pearson τόνιζαν ότι πρέπει να εγκαταλειφθεί η προσπάθεια για την εύρεση της ένδειξης που παρέχει μια μελέτη και ότι η λήψη αποφάσεων πρέπει να στηρίζεται στην εύρεση ή όχι στατιστικά σημαντικών σχέσεων έπειτα από τη σύγκριση της τιμής p, που προκύπτει από την ανάλυση των δεδομένων μιας μελέτης, με την τιμή α που προκαθορίζεται από τους ερευνητές. Χαρακτηριστικό του πόσο λανθασμένη είναι η προσέγγιση αυτή στη λήψη αποφάσεων είναι και το γεγονός ότι ο ίδιος ο Fisher, που εισήγαγε ουσιαστικά την έννοια της τιμής p, ήταν τελείως αντίθετος με τον (παραγωγικό) τρόπο που χρησιμοποιήθηκε η τιμή p στους ελέγχους υποθέσεων από τους Neyman και Pearson.[14]

Είναι σαφές ότι ακόμη και σήμερα ένας σημαντικός αριθμός περιοδικών που αφορούν στις επιστήμες υγείας, καθώς και η πλειοψηφία των επιστημόνων υγείας υιοθετούν την άποψη αυτή των Neyman και Pearson για τη λήψη αποφάσεων με βάση τη στατιστική σημαντικότητα που προκύπτει από την εφαρμογή των διαφόρων στατιστικών ελέγχων.

Προκαλεί αρκετά ερωτηματικά το γεγονός ότι η εξαγωγή συμπερασμάτων στις επιστήμες υγείας εξακολουθεί δυστυχώς ακόμη και σήμερα να στηρίζεται στους ελέγχους υποθέσεων, μιας παραγωγικής μεθόδου που δε συμβάλλει ουσιαστικά στην αύξηση του πληροφοριακού περιεχομένου για το φυσικό κόσμο, καθώς δε λαμβάνεται υπόψη η ένδειξη που παρέχεται από κάθε μελέτη ξεχωριστά, αλλά επιδιώκεται η μείωση του σφάλματος μακροπρόθεσμα έπειτα από τη διεξαγωγή ενός μεγάλου αριθμού μελετών όσο το δυνατόν πιο όμοιες μεταξύ τους. Η τιμή p που προκύπτει από την ανάλυση των δεδομένων μιας μελέτης χρησιμοποιείται στους ελέγχους υποθέσεων για τη διαπίστωση της ύπαρξης ή όχι στατιστικής σημαντικότητας, μολονότι –σύμφωνα και με το Fisher– η τιμή p αποτελεί απλά ένα μέτρο της ασυμφωνίας ανάμεσα στη μηδενική υπόθεση και τα δεδομένα μιας μελέτης.

Γιατί άραγε η επιμονή αυτή σε μια θεωρία που μεθοδολογικά έχει αποδειχθεί αβάσιμη για την εξαγωγή ασφαλών συμπερασμάτων; Η απάντηση στο ερώτημα αυτό φαίνεται ότι είναι η απλότητα –ή, ίσως, η απλοϊκότητα– με την οποία, μέσω της εφαρμογής των τιμών p, διαπιστώνεται η ύπαρξη ή όχι σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Ιδιαίτερα σήμερα, χάρη στα εξαιρετικά εύχρηστα στατιστικά προγράμματα είναι σχετικά απλή και γρήγορη η ανάλυση (ή καλύτερα η σύνθεση) των δεδομένων μιας μελέτης ακόμη και αν είναι αναγκαία η εφαρμογή εξαιρετικά σύνθετων στατιστικών δοκιμασιών. Με τον τρόπο αυτό, μέσα σε μικρό χρονικό διάστημα μπορούν να πραγματοποιηθούν πολυάριθμοι στατιστικοί έλεγχοι υποθέσεων και να «διαπιστωθεί» άμεσα η ύπαρξη ή όχι σχέσης μεταξύ προσδιοριστή και έκβασης.

Είναι γεγονός ότι η πλειοψηφία των επιστημόνων υγείας σήμερα όχι μόνο έχει εξοικειωθεί με την ευκολία με την οποία ερμηνεύονται τα αποτελέσματα των ελέγχων υποθέσεων, αλλά δυστυχώς φαίνεται να «βολεύεται» κιόλας με την προσέγγιση αυτή, καθώς σχετικά εύκολα και άμεσα διατυπώνονται «εντυπωσιακές» σχέσεις μεταξύ προσδιοριστή και έκβασης με βάση τα δεδομένα μίας και μόνο μελέτης, χωρίς να λαμβάνονται σοβαρά υπόψη τόσο η προϋπάρχουσα ένδειξη όσο και οι βιολογικοί μηχανισμοί. Προς την κατεύθυνση αυτή, εξάλλου, έχει συμβάλλει και η πολιτική των περισσότερων περιοδικών που αφορούν στις επιστήμες υγείας, καθώς αποτελεί κοινό μυστικό ότι είναι περισσότερο πιθανό να δημοσιευτεί μια μελέτη που κατέληξε σε στατιστικά σημαντική σχέση παρά μια μελέτη που κατέληξε σε μη στατιστικά σημαντική σχέση. Οι έλεγχοι υποθέσεων λανθασμένα θεωρούνται από πολλούς ως μια «αντικειμενική» ποσοτική μεθοδολογία με την οποία μπορούν να εξαχθούν αξιόπιστα και «επιστημονικά» συμπεράσματα, οδηγώντας στη λήψη αποφάσεων. Το ερώτημα βέβαια δεν θα έπρεπε να είναι εάν μια μελέτη κατέληξε σε στατιστικά σημαντική σχέση,

αλλά εάν ελήφθησαν υπόψη τα συστηματικά σφάλματα και οι συγχυτικοί παράγοντες, έτσι ώστε το μόνο σφάλμα που υπεισέρχεται στην εκτίμηση του μέτρου σχέσης –το οποίο υποδηλώνει το μέγεθος της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης– να είναι το τυχαίο σφάλμα. Είναι ενθαρρυντικό πάντως το γεγονός ότι τα τελευταία χρόνια έχει αποδώσει, έστω και εν μέρει, καρπούς η προσπάθεια αυτή με την παρουσίαση των διαστημάτων εμπιστοσύνης στα αποτελέσματα μιας μελέτης.[12,26,29–32]

Στατιστική εκτίμηση

Σημειακή εκτίμηση

Λαμβάνοντας υπόψη ότι η επιστημονική έρευνα στηρίζεται σε μετρήσεις, η ανάλυση των δεδομένων μιας μελέτης μπορεί να θεωρηθεί περισσότερο ως ένα πρόβλημα μέτρησης και λιγότερο ως ένα πρόβλημα λήψης απόφασης.[12,23] Η μέτρηση εξάλλου ενός μεγέθους απαιτεί λεπτομερέστερη στατιστική προσέγγιση από ένα στατιστικό έλεγχο που οδηγεί στον υπολογισμό μιας τιμής p, μέσω της οποίας απλά απορρίπτεται ή όχι μια υπόθεση.

Σε μια μελέτη υπολογίζεται ένα μέτρο συχνότητας (όπως, π.χ., η επίπτωση-πυκνότητα, η επίπτωση-ποσοστό, ο επιπολασμός κ.ά.) ή, συνηθέστερα, ένα μέτρο σχέσης (όπως, π.χ., η διαφορά ή ο λόγος των μέτρων συχνότητας, ο συντελεστής παλινδρόμησης κ.ά.).[1,3,33–35] Τα μέτρα συχνότητας χρησιμοποιούνται για την ποσοτικοποίηση της συχνότητας εμφάνισης των διαφόρων εκβάσεων, ενώ τα μέτρα σχέσης για την εκτίμηση του μεγέθους της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Κάτω από ιδανικές συνθήκες, η ανάλυση των δεδομένων και η παρουσίαση των αποτελεσμάτων μιας μελέτης πρέπει να περιλαμβάνει τα μέτρα σχέσης (ή σπανιότερα τα μέτρα συχνότητας) και τα αντίστοιχα διαστήματα εμπιστοσύνης, μέσω των οποίων δηλώνεται η ακρίβεια της μέτρησης. Οι δύο αυτές απαραίτητες πληροφορίες παρέχονται μέσω της διαδικασίας της στατιστικής εκτίμησης (statistical estimation).

Σε μια αιτιολογική επιδημιολογική μελέτη που διεξάγεται με σκοπό την εκτίμηση της σχέσης μεταξύ της χρήσης κινητών τηλεφώνων και της συχνότητας εμφάνισης όγκου στον εγκέφαλο βρέθηκε ότι ο λόγος των μέτρων συχνότητας στους εκτεθειμένους σε σχέση με τους μη εκτεθειμένους ισούται με 2,5. Είναι σαφές ότι η τιμή αυτή δεν αποτελεί την πραγματική τιμή, αλλά μια εκτίμηση της πραγματικής τιμής. Όταν η εκτίμηση παρουσιάζεται ως μια και μοναδική τιμή, τότε ονομάζεται σημειακή εκτίμηση, με τη μοναδική αυτή τιμή να είναι γνωστή ως *σημειακή εκτιμήτρια* (point estimator). Στη συγκεκριμένη μελέτη, η τιμή 2,5 είναι η σημειακή εκτιμήτρια της πραγματικής τιμής του λόγου των μέτρων συχνότητας, η οποία ποσοτικοποιεί τη σχέση ανάμεσα στη χρήση κινητών και τη συχνότητα εμφάνισης όγκου στον εγκέφαλο. Με τον τρόπο αυτόν, τα δεδομένα μιας μελέτης χρησιμοποιούνται για την εκτίμηση ενός μέτρου σχέσης. Ο υπολογισμός των μέτρων σχέσης σε μια μελέτη επηρεάζεται από τυχαία και συστηματικά σφάλματα, καθώς και από τους συγχυτικούς παράγοντες. Η σημειακή εκτιμήτρια ενός μέτρου σχέσης είναι ίδια με την πραγματική τιμή του συγκεκριμένου μέτρου σχέσης μόνο στην περίπτωση που έχουν εξουδετερωθεί τα σφάλματα και οι συγχυτικοί παράγοντες. Για τον καθορισμό εξάλλου της ακρίβειας της σημειακής εκτίμησης χρησιμοποιούνται τα διαστήματα εμπιστοσύνης (confidence intervals, intervals estimation), που παρέχουν ένα εύρος τιμών γύρω από τη σημειακή εκτιμήτρια.

Διαστήματα εμπιστοσύνης

Η σημειακή εκτιμήτρια ενός μέτρου σχέσης –ή ενός μέτρου συχνότητας– αποτελεί μία μόνο τιμή και δεν μπορεί να εκφράσει τη στατιστική μεταβλητότητα, ή, διαφορετικά, το τυχαίο σφάλμα που υπεισέρχεται στην εκτίμηση.[12,23] Εάν το μέγεθος του «δείγματος» σε μια μελέτη είναι μεγάλο, τότε η εκτίμηση ενός μέτρου σχέσης είναι σχετικά ακριβής και το τυχαίο σφάλμα που υπεισέρχεται στην εκτίμηση είναι μικρό. Εάν, ωστόσο, το μέγεθος του «δείγματος» είναι μικρό, τότε η εκτίμηση είναι λιγότερο ακριβής και το τυχαίο σφάλμα είναι μεγαλύτερο. Για τη διαπίστωση του μεγέθους του τυχαίου σφάλματος σε μια εκτίμηση χρησιμοποιείται το διάστημα εμπιστοσύνης. Το επιλεγόμενο διάστημα εμπιστοσύνης είναι αυθαίρετο και καθορίζεται από τους ερευνητές. Το επίπεδο εμπιστοσύνης ισούται με 1 – α, όπου α είναι το προκαθορισμένο επίπεδο στατιστικής σημαντικότητας, η τιμή του οποίου ορίζεται αυθαίρετα από τους ερευνητές. Ένα διάστημα εμπιστοσύνης μπορεί να κυμαίνεται από 0–100%. Στην πλειοψηφία των περιπτώσεων, λαμβάνεται το 95% διάστημα εμπιστοσύνης και σπανιότερα το 90% ή το 99% διάστημα εμπιστοσύνης. Το διάστημα εμπιστοσύνης* είναι ένα εύρος τιμών γύρω από μια σημειακή εκτιμήτρια που δείχνει το βαθμό στατιστικής ακρίβειας της εκτίμησης. Το επίπεδο εμπιστοσύνης λαμβάνεται αυθαίρετα, αλλά για

* *Το διάστημα εμπιστοσύνης για την παράμετρο κατασκευάζεται από ένα συγκεκριμένο σύνολο δεδομένων. Επιπλέον, σχεδιάζεται κατά τέτοιον τρόπο που να εξασφαλίζει ότι σε ένα ορισμένο ποσοστό παρόμοιων εφαρμογών (ανάμεσα σ' έναν μεγάλο αριθμό αυτών) θα περιέχει την πραγματική τιμή της παραμέτρου.*

κάθε επίπεδο εμπιστοσύνης, το εύρος του διαστήματος εκφράζει την ακρίβεια της μέτρησης. Ένα ευρύ διάστημα εμπιστοσύνης υποδηλώνει μικρότερη ακρίβεια, ενώ ένα στενότερο υποδηλώνει μεγαλύτερη ακρίβεια. Το ανώτερο όριο (upper limit) και το κατώτερο όριο (lower limit) του διαστήματος εμπιστοσύνης είναι τα *όρια εμπιστοσύνης* (confidence limits).

Ένα 95% διάστημα εμπιστοσύνης σημαίνει ότι εάν επιλέγονταν τυχαία 100 «δείγματα» (με την πραγματοποίηση αντίστοιχα 100 μελετών) από τον πληθυσμό και χρησιμοποιούνταν για τον υπολογισμό 100 διαστημάτων εμπιστοσύνης για ένα μέτρο σχέσης, τότε τα 95 από τα 100 διαστήματα εμπιστοσύνης θα περιείχαν την πραγματική τιμή του μέτρου σχέσης για το συγκεκριμένο πληθυσμό, ενώ τα 5 δεν θα την περιείχαν. Τονίζεται ότι 95% διάστημα εμπιστοσύνης δε σημαίνει ότι τα 95% όρια εμπιστοσύνης που έχουν προκύψει από μια μελέτη περιέχουν την πραγματική τιμή του μέτρου σχέσης με 95% πιθανότητα. Επιπλέον, η συλλογή και η ανάλυση των δεδομένων και για τα 100 «δείγματα» πρέπει να γίνει με τον ίδιο ακριβώς τρόπο. Έτσι, το μοναδικό στοιχείο που θα διαφέρει, στις 100 αυτές παρόμοιες επαναλήψεις της μελέτης, είναι η στατιστική μεταβλητότητα ή, διαφορετικά, η τύχη. Στην περίπτωση αυτή βέβαια πρέπει (α) η μεταβλητότητα των δεδομένων να περιγράφεται με τη χρήση του κατάλληλου στατιστικού μοντέλου, καθώς και (β) να απουσιάζουν τα συστηματικά σφάλματα και οι συγχυτικοί παράγοντες. Μόνο έτσι μπορεί το διάστημα εμπιστοσύνης να χρησιμοποιηθεί για να εκφράσει τη στατιστική μεταβλητότητα ή, αλλιώς, το τυχαίο σφάλμα μιας εκτίμησης. Οι παραπάνω προϋποθέσεις βέβαια είναι πρακτικά αδύνατον να επιτευχθούν, ακόμη και σε πολύ καλά σχεδιασμένες κλινικές δοκιμές. Ιδιαίτερα στην περίπτωση των μη πειραματικών μελετών, τα διαστήματα εμπιστοσύνης παρέχουν απλά μια πρόχειρη εκτίμηση της στατιστικής μεταβλητότητας που υπάρχει στα δεδομένα μιας μελέτης. Είναι προτιμότερο πάντως τα διαστήματα εμπιστοσύνης να μη χρησιμοποιούνται ως ένα απόλυτο μέτρο της στατιστικής μεταβλητότητας, δηλαδή του τυχαίου σφάλματος, αλλά ως ένα γενικό μέτρο της ποσότητας του σφάλματος στα δεδομένα μιας μελέτης.

Διαστήματα εμπιστοσύνης έναντι τιμών p

Τονίζεται ότι τα διαστήματα εμπιστοσύνης δεν πρέπει να χρησιμοποιούνται ως υποκατάστατα των τιμών p για την απόρριψη ή όχι της μηδενικής υπόθεσης. Πιο συγκεκριμένα, όπως προαναφέρθηκε, με τη μηδενική υπόθεση υποστηρίζεται ότι δεν υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Για τον έλεγχο της υπόθεσης αυτής, επιλέγεται το κατάλληλο στατιστικό μοντέλο, ανάλογα με τα δεδομένα μιας μελέτης, και υπολογίζεται η αντίστοιχη τιμή p, που στη συνέχεια συγκρίνεται με την τιμή α, οπότε απορρίπτεται ή όχι η μηδενική υπόθεση.

Δυστυχώς, ορισμένοι χρησιμοποιούν τα διαστήματα εμπιστοσύνης ως υποκατάστατα των τιμών p για την απόρριψη ή όχι της μηδενικής υπόθεσης. Αναλυτικότερα, εάν το μέτρο σχέσης που υπολογίζεται είναι η διαφορά των μέτρων συχνότητας στους εκτεθειμένους και στους μη εκτεθειμένους, τότε εξετάζεται αν το διάστημα εμπιστοσύνης του μέτρου σχέσης περιλαμβάνει την τιμή μηδέν. Στην περίπτωση που το διάστημα εμπιστοσύνης περιλαμβάνει την τιμή μηδέν, τότε δεν απορρίπτεται η μηδενική υπόθεση, ενώ, αντίθετα, αν το διάστημα εμπιστοσύνης δεν περιλαμβάνει την τιμή μηδέν, τότε απορρίπτεται η μηδενική υπόθεση. Εάν, π.χ., το μέτρο σχέσης που υπολογίζεται σε μια μελέτη είναι η μέση διαφορά της συστολικής αρτηριακής πίεσης μεταξύ καπνιστών και μη καπνιστών και η μηδενική υπόθεση είναι ότι η μέση διαφορά της πίεσης στους πληθυσμούς των καπνιστών και των μη καπνιστών είναι ίση με μηδέν, τότε στην περίπτωση που το διάστημα εμπιστοσύνης της μέσης διαφοράς περιέχει την τιμή μηδέν (π.χ., διάστημα εμπιστοσύνης ίσο με -5 έως 15) δεν απορρίπτεται η μηδενική υπόθεση.

Εάν το μέτρο σχέσης που υπολογίζεται είναι ο λόγος των μέτρων συχνότητας στους εκτεθειμένους σε σχέση με τους μη εκτεθειμένους, τότε εξετάζεται αν το διάστημα εμπιστοσύνης περιλαμβάνει τη μονάδα. Στην περίπτωση που το διάστημα εμπιστοσύνης περιλαμβάνει τη μονάδα, τότε δεν απορρίπτεται η μηδενική υπόθεση, ενώ, αντίθετα, αν το διάστημα εμπιστοσύνης δεν περιλαμβάνει τη μονάδα, τότε απορρίπτεται η μηδενική υπόθεση.

Έτσι, στην περίπτωση που το μέτρο σχέσης είναι η διαφορά των μέτρων συχνότητας, η «κρίσιμη» τιμή που εξετάζεται είναι το μηδέν, ενώ αν το μέτρο σχέσης είναι ο λόγος των μέτρων συχνότητας, τότε η «κρίσιμη» τιμή είναι η μονάδα. Η ερμηνεία αυτή είναι ισοδύναμη με τη διεξαγωγή ενός στατιστικού ελέγχου της υπόθεσης ότι δεν υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Επειδή οι έλεγχοι υποθέσεων προσφέρουν λιγότερες πληροφορίες απ' ό,τι τα διαστήματα εμπιστοσύνης και συχνά παρερμηνεύονται, συστήνεται η χρήση των διαστημάτων εμπιστοσύνης για την εκτίμηση τόσο του μεγέθους της σχέσης μεταξύ προσδιοριστή και έκβασης όσο και της ακρίβειας μιας μελέτης.

Είναι σαφές ότι τα διαστήματα εμπιστοσύνης δεν πρέπει να χρησιμοποιούνται ως υποκατάστατα των τιμών p. Εάν

συμβεί κάτι τέτοιο, τότε χάνονται τα πλεονεκτήματα που έχει η χρήση των διαστημάτων εμπιστοσύνης έναντι των τιμών p. Η τιμή p που προκύπτει από ένα στατιστικό έλεγχο αποτελεί ουσιαστικά ένα μέτρο της συμβατότητας μεταξύ της μηδενικής υπόθεσης και των δεδομένων μιας μελέτης. Τα διαστήματα εμπιστοσύνης προσφέρουν επιπλέον μια εκτίμηση του μεγέθους της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, ενώ δηλώνουν και την ακρίβεια της σημειακής αυτής εκτίμησης. Δεν προκαλεί έκπληξη το γεγονός ότι τα διαστήματα εμπιστοσύνης είναι χρησιμότερα και προσφέρουν περισσότερες πληροφορίες, λαμβάνοντας υπόψη ότι τα όρια εμπιστοσύνης είναι δύο αριθμοί, ενώ η τιμή p είναι ένας αριθμός. Η κατάσταση γίνεται ακόμη δυσκολότερη, όταν σε αρκετές μελέτες δεν αναφέρεται καθόλου η ακριβής τιμή p παρά μόνο αν είναι μεγαλύτερη ή μικρότερη από την τιμή α. Στην περίπτωση αυτή, ένα συνεχές μέτρο, όπως είναι η τιμή p, μετατρέπεται σε διχοτόμο, οπότε χάνεται πολύτιμη πληροφορία. Ακόμη πάντως και στην περίπτωση που τα διαστήματα εμπιστοσύνης χρησιμοποιούνται στην παρουσίαση των αποτελεσμάτων μιας μελέτης, δεν θα πρέπει το ενδιαφέρον να εστιάζεται αποκλειστικά στα όρια εμπιστοσύνης, αλλά να λαμβάνεται σοβαρά υπόψη και το εύρος των διαστημάτων εμπιστοσύνης. Σε κάθε περίπτωση βέβαια θα πρέπει να υπολογίζονται τα μέτρα σχέσης και τα αντίστοιχα διαστήματα εμπιστοσύνης, καθώς και η εξαγωγή συμπερασμάτων να μη στηρίζεται στις τιμές p.

Στον πίνακα 3 φαίνονται τα αποτελέσματα μιας υποθετικής αιτιολογικής μελέτης με κλειστούς* πληθυσμούς. Πραγματοποιώντας τον κατάλληλο στατιστικό έλεγχο, που είναι ο έλεγχος x^2, προκύπτει ότι η τιμή p=0,3, οπότε σε επίπεδο στατιστικής σημαντικότητας ίσο με 0,05, δεν απορρίπτεται η μηδενική υπόθεση. Επομένως, λαμβάνοντας υπόψη την τιμή p προκύπτει ότι δεν υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της πάθησης. Ο λόγος των επιπτώσεων-ποσοστών,** ωστόσο, είναι ίσος με 3,2, οπότε προκύπτει ότι η συχνότητα εμφάνισης της νόσου είναι 3 φορές μεγαλύτερη στους εκτεθειμένους σε σχέση με τους μη εκτεθειμένους. Επομένως, η σημειακή εκτιμήτρια του μέτρου σχέσης υποδηλώνει την ύπαρξη σχέσης μεταξύ προσδιοριστή και έκβασης. Το 95% διάστημα εμπιστοσύνης για την τιμή 3,2 του λόγου των επιπτώσεων-ποσοστών είναι ίσο με 0,3 – 31,3. Το γεγονός ότι το 95% διάστημα εμπιστοσύνης περιλαμβάνει την «κρίσιμη» τιμή, δηλαδή τη μονάδα, σημαίνει ότι δεν υπάρχει στατιστικά σημαντική σχέση σε επίπεδο σημαντικότητας ίσο με 1 – 0,95 = 0,05. Με βάση τα όρια του διαστήματος εμπιστοσύνης προκύπτει το συμπέρασμα ότι υπάρχει μέτρια έως ισχυρή σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, παρά το γεγονός ότι ο στατιστικός έλεγχος δηλώνει απουσία σχέσης. Όσο πιο στενό εξάλλου είναι το διάστημα εμπιστοσύνης τόσο μεγαλύτερη είναι και η ακρίβεια της εκτίμησης, τόσο μικρότερο δηλαδή είναι το τυχαίο σφάλμα της μελέτης. Αντίθετα, όσο πιο ευρύ είναι το διάστημα εμπιστοσύνης τόσο μικρότερη είναι και η ακρίβεια της εκτίμησης, τόσο μεγαλύτερο δηλαδή είναι το τυχαίο σφάλμα της μελέτης.

Σύνοψη

Είναι σαφές ότι η χρήση των τιμών p και των ελέγχων υποθέσεων για τη διαπίστωση της ύπαρξης ή όχι σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης είναι λανθασμένη και δεν μπορεί να οδηγήσει στην εξαγωγή ασφαλών συμπερασμάτων. Είναι χαρακτηριστική εξάλλου η έντονη διαμάχη ανάμεσα στο Fisher, που εισήγαγε ουσιαστικά την έννοια των τιμών p, καθώς και στους Neyman και Pearson, τους θεμελιωτές της θεωρίας των ελέγχων υποθέσεων. Ο Fisher ήταν τελείως αντίθετος με τη χρήση των τιμών p για τη διαπίστωση της ύπαρξης σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, επισημαίνοντας ότι η τιμή p αποτελεί μέτρο της ένδειξης που παρέχει μια μελέτη,

** Κλειστός (closed) πληθυσμός είναι ένα «κλειστό» σύνολο ατόμων, όπου η ιδιότητα του μέλους καθορίζεται από ένα συμβάν (event) σε μια συγκεκριμένη τοποχρονική περιοχή.[33] Ο χαρακτηρισμός «κλειστός» σημαίνει ότι «απαγορεύεται» η έξοδος των μελών από τον πληθυσμό αυτό. Η ιδιότητα του μέλους ενός κλειστού πληθυσμού δε χάνεται ούτε με το θάνατο του μέλους. Τυπικότερο παράδειγμα κλειστών πληθυσμών είναι οι δύο σειρές πασχόντων που προκύπτουν μετά από την τυχαιοποίηση και που υποβάλλονται στις δύο συγκρινόμενες θεραπευτικές αγωγές (κλινικές δοκιμές). Ανοικτός (open) πληθυσμός είναι ο πληθυσμός μιας πόλης ή μιας χώρας, οι νοσηλευόμενοι σε ένα νοσοκομείο, τα μέλη μιας ασφαλιστικής εταιρείας κ.ά. Τα μέλη του ανοικτού πληθυσμού εναλλάσσονται στο χρόνο. Η ιδιότητα αυτή δικαιολογεί και το χαρακτηρισμό του ως ανοικτού. Εκείνο που βαθύτερα χαρακτηρίζει τον ανοικτό πληθυσμό και αποτελεί το κύριο στοιχείο του ορισμού του, είναι ότι η ιδιότητα του μέλους προσδιορίζεται από μια κατάσταση (state) και διαρκεί όσο διαρκεί η εν λόγω κατάσταση. Ένα άτομο είναι μέλος του ανοικτού πληθυσμού της Αθήνας όσο ζει στην Αθήνα και για το χρονικό διάστημα που ζει σε αυτή. Χάνει όμως την ιδιότητα του μέλους εφόσον απομακρυνθεί από αυτή.*

*** Η επίπτωση-ποσοστό (incidence-proportion) αποτελεί μέτρο συχνότητας εμφάνισης των νόσων και είναι το ποσοστό των προσωπο-στιγμών, στην αρχή της παρακολούθησης, που εμφάνισε τις περιπτώσεις νόσου κατά τη διάρκεια μιας ορισμένης χρονικής περιόδου.[33] Οι περιπτώσεις νόσου, της οποίας μελετάται η συχνότητα, είναι συμβάντα που παρατηρούνται κατά τη διάρκεια μιας περιόδου παρακολούθησης και ονομάζονται συμβάντα περιόδου. Η επίπτωση-ποσοστό εφαρμόζεται μόνο σε κλειστούς πληθυσμούς και εφόσον ο αριθμός των συμβάντων περιόδου δεν είναι σχετικά μεγάλος.*

Πίνακας 3. Υποθετικά δεδομένα μιας αιτιολογικής μελέτης με κλειστούς πληθυσμούς.

	Προσδιοριστής	
	Ενδεικτική κατηγορία (εκτεθειμένοι)	**Κατηγορία αναφοράς (μη εκτεθειμένοι)**
Περιπτώσεις πάθησης	3	1
Μη περιπτώσεις πάθησης	135	145

εκφράζοντας παράλληλα την αξιοπιστία της μηδενικής υπόθεσης σε σχέση με τα δεδομένα της συγκεκριμένης μελέτης. Είναι αξιοσημείωτο το γεγονός ότι οι Neyman και Pearson στην προσπάθειά τους να περιορίσουν όσο το δυνατόν περισσότερο τη χρήση της τιμής p οδήγησαν στην ακριβώς αντίθετη κατεύθυνση, καθιστώντας ουσιαστικά την τιμή p κριτήριο για τη λήψη αποφάσεων, έπειτα από τη σύγκρισή της με την τιμή α.

Είναι επιβεβλημένο, τουλάχιστον ως πρώτο βήμα, τα αποτελέσματα μιας μελέτης να μην παρουσιάζονται με τη μορφή των τιμών p και τη διαπίστωση της ύπαρξης ή όχι στατιστικά σημαντικών σχέσεων, αλλά με τη μορφή των τιμών των μέτρων σχέσης και των αντίστοιχων διαστημάτων εμπιστοσύνης. Εάν, π.χ., σε μια κλινική δοκιμή διερεύνησης της σχέσης μεταξύ δύο θεραπευτικών παρεμβάσεων για την αντιμετώπιση του καρκίνου του μαστού και της θνητότητας βρεθεί τιμή p=0,001, τότε δεν είναι σαφές εάν η πρώτη παρέμβαση υπερέχει της δεύτερης ή το αντίθετο. Η τιμή p, έπειτα από τη σύγκρισή της με την τιμή α, απλά δηλώνει την ύπαρξη ή όχι στατιστικής σημαντικότητας χωρίς να καθιστά σαφές εάν η ενδεικτική κατηγορία του μελετώμενου προσδιοριστή αυξάνει ή μειώνει τη συχνότητα εμφάνισης της έκβασης. Για το λόγο αυτό πρέπει να αναφέρεται τόσο το μέτρο σχέσης, που δηλώνει το μέγεθος της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, όσο και το αντίστοιχο διάστημα εμπιστοσύνης που δηλώνει την ακρίβεια της μέτρησης. Εάν στο προαναφερθέν παράδειγμα βρεθεί ότι ο λόγος θνητοτήτων είναι ίσος με 4, τότε είναι σαφές ότι η θνητότητα από καρκίνο του μαστού στη μια ομάδα είναι 4 φορές μεγαλύτερη σε σχέση με τη δεύτερη ομάδα. Επιπλέον, η παράθεση του διαστήματος εμπιστοσύνης του μέτρου σχέσης παρέχει τη δυνατότητα εκτίμησης της ακρίβειας της μέτρησης, καθώς όσο μικρότερο είναι το εύρος ενός διαστήματος εμπιστοσύνης τόσο μεγαλύτερη είναι η ακρίβεια της μέτρησης.

Και τα διαστήματα εμπιστοσύνης, ωστόσο, εμφανίζουν σημαντικά μειονεκτήματα, με σημαντικότερο το γεγονός ότι δε συνδυάζουν την ένδειξη που προέρχεται από μια συγκεκριμένη μελέτη με την ένδειξη που προέρχεται από το σύνολο των προγενέστερων μελετών. Το μειονέκτημα αυτό των διαστημάτων εμπιστοσύνης αντιμετωπίζεται με την εφαρμογή Μπεϋζιανών μεθόδων και πιο συγκεκριμένα με τον υπολογισμό του παράγοντα Bayes. Ο επιθυμητός αυτός επαγωγικός τρόπος σκέψης στις επιστήμες υγείας παρέχει τη δυνατότητα να συνδυαστεί η ένδειξη που προέρχεται από προγενέστερες μελέτες –μέσω της εκ των προτέρων πιθανότητας της μηδενικής υπόθεσης να είναι αληθής– με την ένδειξη που προκύπτει από μια συγκεκριμένη μελέτη –μέσω του υπολογισμού του παράγοντα Bayes– για τον υπολογισμό της εκ των υστέρων πιθανότητας της μηδενικής υπόθεσης να είναι αληθής (μέσω της εφαρμογής του θεωρήματος του Bayes).[7,9,16,36–41]

Συνιστάται ανεπιφύλακτα στους συγγραφείς ερευνητικών εργασιών η αποφυγή των τιμών p και η χρησιμοποίηση των τιμών των μέτρων σχέσης και των αντίστοιχων διαστημάτων εμπιστοσύνης στην παρουσίαση των αποτελεσμάτων μιας μελέτης, ως το πρώτο βήμα εξαγωγής ασφαλών συμπερασμάτων, με το δεύτερο να είναι η εφαρμογή των Μπεϋζιανών μεθόδων. Οι τιμές p σε καμιά περίπτωση δεν πρέπει να χρησιμοποιούνται για τη διαπίστωση της ύπαρξης ή όχι σχέσεων μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Ο ρόλος των ερευνητών εξάλλου δεν είναι να εξάγουν συμπεράσματα και να καταλήγουν στη διαπίστωση σχέσεων, αλλά να παρουσιάζουν αναλυτικά τον ερευνητικό σχεδιασμό* μιας μελέτης, καθώς και τη στατιστική μεθοδολογία που χρησιμοποιείται, παρέχοντας τη δυνατότητα στους αναγνώστες να κρίνουν το μέγεθος και την αξιοπιστία της ένδειξης που παρέχουν τα δεδομένα μιας μελέτης. Το ερώτημα δεν θα έπρεπε να είναι εάν μια μελέτη κατέληξε σε στατιστικά σημαντική σχέση, αλλά εάν ελήφθησαν υπόψη τα συστηματικά σφάλματα και οι συγχυτές, έτσι ώστε το μόνο σφάλμα που υπεισέρχεται στην εκτίμηση του μέτρου σχέσης –το οποίο υποδηλώνει το μέγεθος της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης– να είναι το τυχαίο σφάλμα.

Είναι σαφές ότι η εγκυρότητα και η αξιοπιστία των

* *Ο ερευνητικός σχεδιασμός περιλαμβάνει το σχεδιασμό του αντικειμένου και της μεθόδου κάθε μελέτης.*[3,35] *Με τον όρο αντικείμενο νοείται το τελικό αποτέλεσμα μιας μελέτης και πιο συγκεκριμένα, το είδος και η ποσότητα της εμπειρικής πληροφορίας που αυτή παρέχει. Με τον όρο μέθοδος νοείται ο τρόπος προσέγγισης του τελικού αποτελέσματος, δηλαδή η διεργασία απόκτησης της εμπειρικής πληροφορίας.*

αποτελεσμάτων μιας μελέτης δεν πρέπει να καθορίζονται από την εύρεση στατιστικά σημαντικών σχέσεων, αλλά από τον ερευνητικό σχεδιασμό και τον περιορισμό των σφαλμάτων. Για το λόγο αυτό, οι προσπάθειες των ερευνητών πρέπει να επικεντρώνονται στο σχεδιασμό και τη διεξαγωγή μελετών με το μικρότερο δυνατό σφάλμα και όχι στη μανιώδη αναζήτηση στατιστικά σημαντικών σχέσεων που δυστυχώς είναι γεγονός ότι αυξάνουν σημαντικά την πιθανότητα δημοσίευσης μιας μελέτης. Είναι οξύμωρο το γεγονός ότι τα τελευταία 30 έτη, μολονότι έχουν καταβληθεί σημαντικές προσπάθειες για την περιθωριοποίηση των τιμών p και την υιοθέτηση των διαστημάτων εμπιστοσύνης, τα περισσότερα περιοδικά που αφορούν στις επιστήμες υγείας δεν έχουν κινηθεί προς την κατεύθυνση αυτή. Πάντως, ακόμη και αν δεν απομακρυνθούν άμεσα οι τιμές p από την παρουσίαση των αποτελεσμάτων μιας μελέτης, θα ήταν εξαιρετικά επωφελής η παράθεση των τιμών των μέτρων σχέσης και των αντίστοιχων διαστημάτων εμπιστοσύνης. Όπως προαναφέρθηκε, εξάλλου, η γνώση ενός διαστήματος εμπιστοσύνης συνεπάγεται έμμεσα και τη γνώση της ύπαρξης ή όχι στατιστικής σημαντικότητας, οπότε οι τιμές p είναι ουσιαστικά «άχρηστες» στην περίπτωση που παρουσιάζονται τα διαστήματα εμπιστοσύνης.

ABSTRACT

Saying Farewell to p Values and Welcoming Confidence Intervals in Data Analysis

Petros Galanis

MSc in Public Health, PhD, RN, Center for Health Services Management and Evaluation, Faculty of Nursing, National and Kapodistrian University of Athens, Athens, Greece

The finding or not of statistically significant relationships between a determinant and the frequency of occurrence of an outcome relies on the comparison of p values, which are extracted from various statistical tests, using an arbitrarily selected value, determined by the researchers, known as the alpha level. The alpha level usually used is 0.05. Over the last 30 years, the application of p values for the derivation of conclusions in health sciences has been rightly criticized. The mathematical and conceptual approach of hypotheses tests and p values was an important step in statistical analysis, but there are inherent problems in their interpretation and practical application. Ideally, the data analysis and presentation of the results of a study should include the measures of association (or rarely the measures of frequency) and the corresponding confidence intervals (CIs) that indicate the precision of measurement. Statistical estimation can provide this information. The CI is a range of values around a measure of association, computed in a study, and it displays the degree of statistical precision of the estimation. A wide CI indicates lower precision, while a narrow CI higher precision. The p value, after the comparison with the alpha level, indicates only the existence or not of statistical significance without making clear whether the index category of the determinant under study increases or decreases the frequency of occurrence of outcome. For this reason, the measure of association should be reported that indicates the magnitude of the relation between the determinant and the frequency of occurrence of the outcome, along with the corresponding CI, that indicates the precision of measurement. The validity and credibility of the results of a study should not be determined by the finding of a statistically significant relationship alone, but by the research design and the reduction of errors. Although p values should not be eliminated from the presentation of the study results, it would be beneficial to introduce the measures of association and the corresponding CIs. The knowledge of a CI includes indirectly the knowledge of the presence or absence of statistical significance, rendering p values unnecessary when CIs are used. ***NOSILEFTIKI* 2010, 49 (1): 11-25.**

Key-words: *confidence interval, data analysis, hypothesis test, p value, statistical significance*

✉ **Corresponding Author:** Petros Galanis, 14 Dikis street, GR-157 73 Athens, Greece, tel.: +30 210 77 81 044, +30 6944 387 354, e-mail: pegalan@nurs.uoa.gr

Βιβλιογραφία

1. Σπάρος ΛΔ, Γαλάνης Π. *Δοκίμια επιδημιολογίας*. Εκδόσεις Παρισιάνου, Αθήνα, 2006
2. Miettinen OS. *Theoretical epidemiology. Principles of occurrence research in medicine.* John Wiley & Sons, New York, 1985
3. Γαλάνης ΠΑ, Σπάρος ΛΔ. *Εγχειρίδιο επιδημιολογίας.* Εκδόσεις ΒΗΤΑ, Αθήνα, 2010
4. Anonymous. Uniform requirements for manuscripts submitted

to biomedical journals. International Committee of Medical Journal Editors. *N Engl J Med* 1997, 36:309–315
5. Rothman KJ. Writing for epidemiology. *Epidemiology* 1998, 9:333–337
6. Fisher RA. *Statistical methods for research workers.* 13th ed. Hafner, New York, 1958
7. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med* 1999, 130:995–1004
8. Howson C, Urbach P. *Scientific reasoning: The Bayesian approach.* 2nd ed. Open Court Publishing Company, Chicago, 1993
9. Goodman SN. P values, hypothesis tests and likelihood: Implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993, 137:485–496
10. Stigler SM. *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press, Cambridge, 1986
11. David HA. First (?) occurrence of common terms in mathematical statistics. *American Statistician* 1995, 49:121–133
12. Rothman KJ, Greenland S, Lash TL. *Modern epidemiology.* 3rd ed. Lippincott Williams & Wilkins, Philadelphia, 2008
13. Sackett DL, Straus SE, Richardson SW, Rosenberg W, Haynes BR. *Επί ενδείξεων βασιζόμενη ιατρική. Πώς να ασκείται και να διδάσκεται την EBI* (ελληνική μετάφραση Ε. Ανευλαβή). Εκδόσεις Πασχαλίδης, Αθήνα, 2002:21–55
14. Fisher RA. *Statistical methods and scientific inference.* 3rd ed. Macmillan, New York, 1973
15. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987, 257:2459–2463
16. Diamond GA, Forrester JS. Clinical trials and statistical verdicts: Probable grounds for appeal. *Ann Intern Med* 1983, 98:385–394
17. Lilford RJ, Braunholtz D. The statistical basis of public policy: A paradigm shift is overdue. *BMJ* 1996, 313:603–607
18. Freeman PR. The role of p-values in analysing trial results. *Stat Med* 1993, 12:1443–1552
19. Neyman J. Outline of a theory of statistical estimation based on the classical theory of probability. *Phil Trans R Soc Lond A* 1937, 236:333–380
20. Brophy JM, Joseph L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA* 1995, 273:871–875
21. Berkson J. Tests of significance considered as evidence. *JASA* 1942, 37:325–335
22. Pearson E. "Student" as a statistician. *Biometrika* 1938, 38:210–250
23. Rothman KJ. *Epidemiology: An introduction.* Oxford University Press, New York, 2002:113–129
24. Neyman J, Pearson E. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 1928, 20:175–240
25. Lehmann EL. *Testing statistical hypotheses.* 2nd ed. Wiley, New York, 1986
26. Γαλάνης Π, Σπάρος ΛΔ. Στατιστικά μοντέλα για την ανάλυση των επιδημιολογικών δεδομένων. *Αρχ Ελλ Ιατρ* 2006, 23:404–417
27. Greenberg RS, Daniels SR, Flanders DW, Eley WJ, Boring JR. *Medical epidemiology.* Prentice-Hall International, London, 1993
28. Neyman J, Pearson E. On the problem of the most efficient tests of statistical hypotheses. *Phil Trans R Soc Lond A* 1933, 231:289–337
29. Rothman KJ. Significance questing. *Ann Intern Med* 1986, 105:445–447
30. Barnett ML, Mathisen A. Tyranny of the p-value: The conflict between statistical significance and common sense. *J Dent Res* 1997, 76:534–536
31. Bailar JC 3rd, Mosteller F. Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations. *Ann Intern Med* 1988, 108:266–273
32. Γαλάνης Π, Σπάρος ΛΔ. Ανάλυση δεδομένων: Μη μπαγιεσιανή προσέγγιση. *Αρχ Ελλ Ιατρ* 2005, 22:377–391
33. Γαλάνης Π, Σπάρος ΛΔ. Μέτρα συχνότητας των νοσημάτων. *Αρχ Ελλ Ιατρ* 2005, 22:178–191
34. Γαλάνης Π, Σπάρος ΛΔ. Η έννοια του αποδοτέου κλάσματος στην εφαρμοσμένη ιατρική έρευνα. *Αρχ Ελλ Ιατρ* 2005, 22:157–169
35. Σπάρος ΛΔ, Γαλάνης Π, Ζάχος Ι, Τσιλίδης Κ. *Επιδημιολογία Ι.* Εκδόσεις ΒΗΤΑ, Αθήνα, 2004
36. Goodman SN. Toward evidence-based medical statistics. 2: The Bayes factor. *Ann Intern Med* 1999, 130:1005–1013
37. Miettinen OS. Up from "false positives" in genetic -and other- epidemiology. *Eur J Epidemiol* 2009, 24:1–5
38. Goodman SN, Royall R. Evidence and scientific research. *Am J Public Health* 1988, 78:1568–1574
39. Freedman L. Bayesian statistical methods. *BMJ* 1996, 313:569–570
40. Etzioni RD, Kadane JB. Bayesian statistical methods in public health and medicine. *Annu Rev Public Health* 1995, 16:23–41
41. Kadane JB. Prime time for Bayes. *Control Clin Trials* 1995, 16:313–318