ARCHIVES OF HELLENIC MEDICINE: ISSN 11-05-3992

ΑΡΧΕΙΑ ΕΛΛΗΝΙΚΗΣ ΙΑΤΡΙΚΗΣ 2010, 27(4):691–707
ARCHIVES OF HELLENIC MEDICINE 2010, 27(4):691–707

Η λανθασμένη εφαρμογή των τιμών P και του ελέγχου των υποθέσεων στη βιοϊατρική έρευνα

Π. Γαλάνης

Εργαστήριο Οργάνωσης και Αξιολόγησης Υπηρεσιών Υγείας, Τμήμα Νοσηλευτικής, Πανεπιστήμιο Αθηνών, Αθήνα

The wrong application of P values and hypotheses test in biomedical research

Abstract at the end of the article

Λέξεις ευρετηρίου

Διαλογισμός
Διάστημα εμπιστοσύνης
Έλεγχος υπόθεσης
Εξήγηση
Παράγοντας Bayes
Τιμή P

Υποβλήθηκε 8.9.2009
Εγκρίθηκε 16.9.2009

1. ΕΙΣΑΓΩΓΗ

Η ανάλυση των δεδομένων* και η ερμηνεία των αποτελεσμάτων που αφορούν στη βιοϊατρική έρευνα εξακολουθεί ακόμη και σήμερα να αντιμετωπίζει ένα εξαιρετικά σημαντικό πρόβλημα. Πιο συγκεκριμένα, τόσο η κατανόηση των βιολογικών μηχανισμών μεταξύ προσδιοριστή** και έκβασης*** όσο και τα αποτελέσματα προηγούμενων μελετών συναφών με την εκάστοτε επιστημονική υπόθεση δυστυχώς δεν λαμβάνονται ιδιαίτερα υπόψη στην ερμηνεία των αποτελεσμάτων μιας συγκεκριμένης μελέτης, μειώνοντας έτσι την αξιοπιστία των συμπερασμάτων.[1] Στην πλειονότητα των περιπτώσεων σήμερα, η εξαγωγή συμπερασμάτων στηρίζεται στους ελέγχους των υποθέσεων και στις τιμές P χωρίς να λαμβάνεται υπόψη η ένδειξη που προέρχεται από προγενέστερες μελέτες και ιδιαιτέρως τυχαιοποιημένες ελεγχόμενες δοκιμές (randomized controlled trials).

Το 1992, στο Πανεπιστήμιο McMaster του Καναδά, η ερευνητική ομάδα του David Sackett υιοθέτησε τον όρο *«Ιατρική βασιζόμενη σε ενδείξεις»***** (evidence-based

* 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, τα οποία δεν είναι θεωρητικά φορτισμένα και (β) τα «δεδομένα» αυτά δεν αναλύονται, αλλά συντίθενται. Εάν δηλαδή διατηρηθεί ο όρος «δεδομένα», τότε ο όρος «ανάλυση δεδομένων» πρέπει να αντικατασταθεί από τον όρο «σύνθεση δεδομένων» (data synthesis).[2]

** 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.[3,4] 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 πάθησης 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 πάθησης 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.[4] Η έκφραση «σχετίζεται με την έκβαση» σημαίνει σχέση με τη συχνότητα εμφάνισης της έκβασης και όχι με την έκβαση καθεαυτή. Στην προγνωστική επιδημιολογία, τα μελετώμενα άτομα πάσχουν ήδη από μια συγκεκριμένη πάθηση, οπότε η έκβαση χρησιμοποιείται για να δηλώσει το πέρας της πάθησης (π.χ. την ίαση, το θάνατο, την εμφάνιση καταλοίπων κ.ά.).

**** Η Ιατρική βασιζόμενη σε ενδείξεις συνιστά τη σύνθεση των βέλτιστων ενδείξεων της έρευνας με την κλινική εμπειρία και τις αξίες του πάσχοντα.[6] Ο συνδυασμός των τριών αυτών στοιχείων οδηγεί στη λήψη κλινικών αποφάσεων με ορθολογικό τρόπο, βελτιστοποιώντας έτσι τις κλινικές εκβάσεις και την ποιότητα ζωής των πασχόντων.

medicine) για να δηλώσει την ανάγκη να καταφεύγουν οι επιστήμονες υγείας στις υπάρχουσες μελέτες σχετικά με μια συγκεκριμένη επιστημονική υπόθεση, αποβλέποντας έτσι στη λήψη κλινικών αποφάσεων με ορθολογικό τρόπο.[5] Έκτοτε, ο αριθμός των άρθρων που αφορούν στην Ιατρική βασιζόμενη σε ενδείξεις έχει αυξηθεί σημαντικά (από μία δημοσίευση το 1992 σε περίπου 1.000 το 1998), ενώ έχουν δημιουργηθεί και περιοδικά (όπως π.χ. το *Evidence-Based Medicine*, το *Journal of Evidence-Based Health*, το *Evidence-Based Cardiovascular Medicine*, το *Evidence-Based Mental Health*, το *Evidence-Based Nursing* κ.ά.) που εστιάζονται σχεδόν αποκλειστικά προς την κατεύθυνση αυτή. Εντούτοις, στις περισσότερες περιπτώσεις, οι μέθοδοι του στατιστικού διαλογισμού που χρησιμοποιούνται δεν στηρίζονται στις ενδείξεις, με αποτέλεσμα να δημιουργείται σύγχυση. Πιο συγκεκριμένα, οι στατιστικές μέθοδοι που χρησιμοποιούνται στη βιοϊατρική έρευνα καταλήγουν στον υπολογισμό μιας πιθανότητας (που καλείται τιμή P), η οποία χρησιμοποιείται για να «διαπιστωθεί» αν υπάρχει ή όχι σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης χωρίς όμως να λαμβάνονται υπόψη τόσο η προϋπάρχουσα ένδειξη* όσο και οι υπάρχοντες βιολογικοί μηχανισμοί. Με τον τρόπο αυτόν περιορίζεται σημαντικά η αξιοπιστία των συμπερασμάτων μιας μελέτης και παράλληλα είναι ανέφικτο να συνδυαστεί η ένδειξη που προκύπτει από μια συγκεκριμένη μελέτη με την ένδειξη που προέρχεται από το σύνολο των μελετών που έχουν ήδη διερευνήσει την ίδια επιστημονική υπόθεση.

Η χρήση των ελέγχων των υποθέσεων και των τιμών P μπορεί να οδηγήσει σε επισφαλή συμπεράσματα και γι' αυτό συστήνεται η εγκατάλειψή τους και η υιοθέτηση μπεϋζιανών μεθόδων και ειδικότερα ο υπολογισμός ενός μέτρου που είναι γνωστό ως *παράγοντας Bayes* (Bayes factor). Η υιοθέτηση της μπεϋζιανής μεθοδολογίας είναι εξαιρετικής σημασίας, καθώς δίνει τη δυνατότητα συνδυασμού της ένδειξης που προέρχεται από μια συγκεκριμένη μελέτη με την προϋπάρχουσα ένδειξη, οδηγώντας έτσι τους επιστήμονες υγείας στην ορθολογική λήψη αποφάσεων με βάση τις ενδείξεις.[7] Ο μπεϋζιανός διαλογισμός** συχνά παρουσιάζεται ως μια μέθοδος εκτίμησης της μεταβολής του *βαθμού πεποίθησης* (degree of belief) ενός επιστήμονα υγείας σχετικά με μια συγκεκριμένη επιστημονική υπόθεση με βάση την ένδειξη (ή, αλλιώς, την πληροφορία ή το αποτέλεσμα) που προέρχεται από μια συγκεκριμένη μελέτη. Μολονότι η μπεϋζιανή μεθοδολογία έχει αναπτυχθεί σημαντικά τα τελευταία 30 χρόνια, η πλειοψηφία των ερευνητών αρνείται πεισματικά την εφαρμογή της (προς όφελος των ελέγχων των υποθέσεων και των τιμών P) θεωρώντας ότι αποτελεί υποκειμενική μέθοδο ανάλυσης των δεδομένων. Στο άρθρο αυτό θα αναλυθούν οι τιμές P και οι έλεγχοι των υποθέσεων, καθώς και τα μειονεκτήματα που παρουσιάζει η εφαρμογή τους στη βιοϊατρική έρευνα.

2. ΕΠΙΣΤΗΜΟΝΙΚΗ ΕΞΗΓΗΣΗ

Προτού γίνει εκτενής αναφορά στα μειονεκτήματα που παρουσιάζουν οι τιμές P και οι έλεγχοι των υποθέσεων, απαραίτητη είναι η κατανόηση των διαφόρων μορφών της επιστημονικής εξήγησης. Οι δύο κεντρικές λειτουργίες της επιστημονικής γνώσης είναι η εξήγηση (explanation) και η πρόβλεψη (prediction) ή πρόρρηση των φαινομένων της πραγματικότητας. Και οι δύο αποτελούν επιστημονικές συστηματοποιήσεις*** με τις οποίες εφαρμόζονται οι επιστημονικές υποθέσεις ή νόμοι**** και θεωρίες***** σε συγκεκριμένες περιπτώσεις, αποκαλύπτοντας με τον τρόπο αυτόν τις διαρθρωτικές σχέσεις μεταξύ των φαινομένων.[8]

Πολλές επιστημονικές εξηγήσεις επιζητούνται μέσω ερωτημάτων του τύπου «γιατί;» σε μια προσπάθεια να

* Σημασιολογικά, η έννοια ένδειξη είναι συνυφασμένη με την πληροφορία.[6] 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 σε ένα συγκεκριμένο πάσχοντα. Η ένδειξη εκφράζει τη γνώση (ως αποτέλεσμα συνήθως των παρατηρήσεων) που είναι διαθέσιμη στους ερευνητές και η οποία μπορεί να χρησιμοποιηθεί για τον καθορισμό του βαθμού επικύρωσης (degree of confirmation) μιας υπόθεσης ή μιας εκτίμησης. Η υπόθεση (hypothesis), εξάλλου, αφορά σε άγνωστα γεγονότα (όπως π.χ. μια πρόβλεψη ή ένα νόμο) και κρίνεται με βάση την υπάρχουσα ένδειξη.

** Διαλογισμός (inference) είναι η νοητική διαδικασία ή η μέθοδος με την οποία ο νους καταστρώνει ένα επιχείρημα. Το επιχείρημα αποτελεί την αρτιότερη λογική κατασκευή και είναι μια σειρά αλληλένδετων κρίσεων-προτάσεων που σχηματίζεται για να κάνει φανερή (να «αποδείξει») την αλήθεια μιας απόφανσης. Απαιτείται να γίνεται διάκριση ανάμεσα στο διαλογισμό και το συλλογισμό (sylogism), που αποτελεί ορισμένο είδος διαλογισμού.

*** Οι επιστημονικές συστηματοποιήσεις αποτελούν τα λογικά εκείνα σχήματα που επιτρέπουν την έγκυρη συναγωγή παρόντων, παρελθόντων και μελλόντων περιστατικών.[8]

**** Τα απλούστερα στοιχεία της επιστημονικής γλώσσας είναι υποθέσεις ή νόμοι, δηλαδή προτάσεις με γενικό υποθετικό χαρακτήρα που περιέχουν το διαψεύσιμο ισχυρισμό ότι φαινόμενα της πραγματικότητας έχουν ορισμένες ιδιότητες ή σχέσεις.[8]

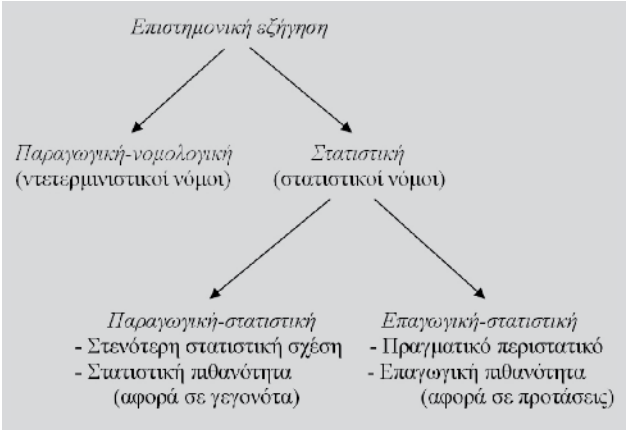
***** Οι θεωρίες είναι σύνολα, λογικά συνδεδεμένων μεταξύ τους, υποθέσεων που συστηματοποιούν, ενοποιούν και εξηγούν μια ορισμένη περιοχή της πραγματικότητας.[8] Με την κατασκευή και την ακριβή διατύπωση επιστημονικών θεωριών επιδιώκεται η σύλληψη της εσωτερικής ενότητας των φαινομένων και η εξήγηση ευρύτερων περιοχών της πραγματικότητας.

κατανοηθούν διάφορα φαινόμενα. Οι επιστημονικές εξηγήσεις είναι συχνά απαντήσεις σε ερωτήσεις τέτοιου τύπου, αλλά διαφέρουν ουσιωδώς από άλλες μορφές εξήγησης. Επισημαίνεται ότι, στην επιστημονική εξήγηση, το αντικείμενο της εξηγητικής προσπάθειας ή, αλλιώς, το *εξηγητέο* (explanandum) μπορεί να είναι είτε επιμέρους περιστατικό είτε επιστημονικός νόμος.[9]

Η επιστημονική εξήγηση των πραγματικών περιστατικών έχει ως αντικείμενο ένα φαινόμενο της πραγματικότητας που περιγράφεται από μια ατομική πρόταση βάσης.[10] Το ερώτημα στο οποίο επιχειρείται με την επιστημονική εξήγηση να δοθεί μια απάντηση είναι «γιατί συνέβη το περιστατικό που περιγράφει η ατομική πρόταση βάσης;». Με τον τρόπο αυτόν, η επιστημονική εξήγηση αποσκοπεί στο να γίνει κατανοητό ένα επιμέρους συμβάν ή κάποιο γενικό γεγονός, καταφεύγοντας σε άλλα επιμέρους ή και γενικά γεγονότα που λαμβάνονται από έναν ή περισσότερους κλάδους της εμπειρικής επιστήμης.[9]

Εξήγηση ενός φαινομένου είναι η υπαγωγή του σε επιστημονικούς νόμους.[8] Απαρχές της ιδέας αυτής υπάρχουν ήδη στον Αριστοτέλη, ενώ οι David Hume, Immanuel Kant και John Stuart Mill συνέβαλαν αποφασιστικά προς την κατεύθυνση αυτή. Το πλέον αποφασιστικό βήμα πάντως έγινε από τον Karl Popper[11] και λίγο αργότερα από τους Carl Hempel και Paul Oppenheim.[12] Το σχήμα της επιστημονικής εξήγησης που στηρίζεται σε ντετερμινιστικούς νόμους* έγινε γνωστό ως το P-H-O σχήμα από τα αρχικά των ονομάτων των τριών αυτών συγγραφέων (Popper, Hempel και Oppenheim) που συντέλεσαν στην ακριβέστερη μορφοποίηση του λογικού σχήματος της εξήγησης.

Η επιστημονική εξήγηση διακρίνεται στην παραγωγική-νομολογική εξήγηση και στη στατιστική εξήγηση (παραγωγική-στατιστική και επαγωγική-στατιστική εξήγηση) (εικ. 1).



Εικόνα 1. Μορφές επιστημονικής εξήγησης.

2.1. Παραγωγική-νομολογική εξήγηση

Η βασική ιδέα σε αυτή τη μορφή εξήγησης συνίσταται στο ότι η επιστημονική εξήγηση έχει τη λογική μορφή του επιχειρήματος αποτελούμενη από τις προκείμενες (εξηγούν) και το συμπέρασμα (εξηγητέο). Το παρακάτω P-H-O σχήμα αποτελεί τη βάση της παραγωγικής-νομολογικής εξήγησης:[8]

$$\begin{array}{l} L_1, L_2 \ldots . L_k \\ C_1, C_2 \ldots . C_m \\ \hline E \end{array} \quad \begin{array}{l} \\ \text{Εξηγούν} \\ \text{Εξηγητέο} \end{array}$$

Στο παραπάνω σχήμα, με $L_1, L_2 \ldots L_k$ συμβολίζονται οι επιστημονικοί νόμοι, με $C_1, C_2 \ldots C_m$ συμβολίζονται οι προτάσεις που περιέχουν τους όρους εφαρμογής (αρχικές ή διαρκείς συνθήκες) των επιστημονικών νόμων και με E συμβολίζεται η πρόταση που περιγράφει το εξηγητέο περιστατικό. Σύμφωνα με το παραπάνω σχήμα, εξήγηση του E είναι η λογική του παραγωγή από μια σειρά επιστημονικών νόμων και τους αντίστοιχους όρους εφαρμογής τους.[8]

Ένα χαρακτηριστικό παράδειγμα εφαρμογής της παραγωγικής-νομολογικής εξήγησης δίνεται από τον Popper, απαντώντας στο ερώτημα «γιατί μια συγκεκριμένη κλωστή έσπασε, όταν κρεμάστηκε από αυτή ένα ορισμένο βάρος;». Η εξήγηση περιέχεται στη φράση ότι η κλωστή είχε αντοχή 1 kg, ενώ το βάρος που κρεμάστηκε ήταν 2 kg. Εφαρμόζοντας το P-H-O σχήμα προκύπτουν τα εξής:

(L_1) Για κάθε κλωστή με μια χαρακτηριστική δομή S (που καθορίζεται από το υλικό, το πάχος κ.ά.) υπάρχει ένα χαρακτηριστικό βάρος w, ώστε η κλωστή σπάζει, αν επιβαρύνεται με βάρος μεγαλύτερο από w

(L_2) Για κάθε κλωστή με τη δομή S_1, το χαρακτηριστικό βάρος είναι 1 kg

(C_1) Αυτή είναι μια κλωστή με δομή S_1

(C_2) Το βάρος που επιδρά στην κλωστή αυτή είναι 2 kg

(E) Η κλωστή σπάζει

Το σπάσιμο της συγκεκριμένης κλωστής δείχνει ότι η πρόταση που το περιγράφει παράγεται λογικά από επιστη-

* Με κριτήριο τη μορφή τους, οι επιστημονικοί νόμοι διακρίνονται σε ντετερμινιστικούς και στατιστικούς (πιθανολογικούς ή στοχαστικούς).[8] Οι ντετερμινιστικοί περιέχουν έναν ισχυρισμό που αφορά σε όλα τα φαινόμενα μιας ορισμένης κατηγορίας και δεν επιδέχονται εξαιρέσεις, ενώ οι στατιστικοί ισχυρίζονται την ύπαρξη μιας ιδιότητας ή σχέσης που καλύπτει ένα μόνο τμήμα των φαινομένων μιας κατηγορίας. Και στις δύο περιπτώσεις πρόκειται για γενικές υποθετικές προτάσεις που ισχύουν ανεξάρτητα από τόπο και χρόνο, ενώ δεν περιέχουν απλή απαρίθμηση επιμέρους περιπτώσεων ή παρωχημένων γεγονότων.

μονικούς νόμους και τους όρους εφαρμογής τους.

Ένα ακόμη κλασικό παράδειγμα παραγωγικού επιχειρήματος είναι το εξής:

Όλοι οι άνθρωποι είναι θνητοί
Ο Σωκράτης είναι άνθρωπος
——— Εξηγούν
Ο Σωκράτης είναι θνητός ← Εξηγητέο

2.2. Στατιστική εξήγηση

Η στατιστική (στοχαστική ή πιθανολογική) εξήγηση περιέχει στο *εξηγούν* (explanans) μία τουλάχιστον στατιστική υπόθεση και διακρίνεται στην παραγωγική-στατιστική εξήγηση και στην επαγωγική-στατιστική εξήγηση.

2.2.1. Παραγωγική-στατιστική εξήγηση

Στα πλαίσια της εξήγησης αυτής μια στενότερη στατιστική σχέση παράγεται (και εξηγείται) από μια ευρύτερη.[8] Στην περίπτωση αυτή, το εξηγητέο έχει πιθανολογικό χαρακτήρα, ενώ η σχέση του προς το εξηγούν αποτελεί λογική ακολουθία, που σημαίνει ότι το εξηγητέο παράγεται λογικά, έγκυρα και με βεβαιότητα από το εξηγούν. Στο εξηγούν περιλαμβάνονται στατιστικοί και όχι ντετερμινιστικοί νόμοι.

Χαρακτηριστικό παράδειγμα της εξήγησης αυτής αναφέρει ο Hempel, σύμφωνα με το οποίο εξηγητέο είναι η πρόταση «η πιθανότητα να ριφθούν γράμματα έπειτα από μια σειρά κορώνες είναι ½». Για να εξηγηθεί η σχέση που περιγράφει η πρόταση αυτή, χρησιμοποιούνται στο εξηγούν οι παρακάτω δύο υποθέσεις:

(α) Η πιθανότητα να ριφθούν γράμματα με ένα κανονικό νόμισμα είναι ½.

(β) Οι διάφορες ρίψεις στατιστικά είναι ανεξάρτητα μεταξύ τους γεγονότα.

2.2.2. Επαγωγική-στατιστική εξήγηση

Στην περίπτωση αυτή, εξηγείται ένα συγκεκριμένο πραγματικό περιστατικό και όχι μια στενότερη στατιστική σχέση, ενώ το γενικό σχήμα που εφαρμόζεται είναι το παρακάτω:[8]

$$\frac{P(G, F) \text{ είναι κοντά στο } 1}{F_i} \qquad \frac{P(G, F) \text{ είναι κοντά στο } 0}{F_i}$$

Κάνει πολύ πιθανό (είναι σχεδόν βέβαιο)

$$G_i \qquad G_i$$

Στην πρώτη σειρά, υπάρχει ένας στατιστικός νόμος που περιέχει τον ισχυρισμό ότι η πιθανότητα να είναι ένα γεγονός του είδους F και γεγονός του είδους G είναι πολύ μεγάλη ή πολύ μικρή (κοντά στο ένα ή το μηδέν, αντίστοιχα). Στη δεύτερη σειρά υπάρχει ο όρος εφαρμογής: Το γεγονός i ανήκει στην κατηγορία F. Το συμπέρασμα είναι ότι με μεγάλη πιθανότητα (χωρίς να είναι απολύτως βέβαιο), το γεγονός i ανήκει (ή δεν ανήκει) στην κατηγορία G. Η διπλή γραμμή υποδηλώνει ότι η σχέση ανάμεσα στο εξηγούν και το εξηγητέο δεν είναι παραγωγική αλλά «επαγωγική», καθώς οι στατιστικοί νόμοι και οι όροι εφαρμογής προσφέρουν επαγωγική μόνο στήριξη στο εξηγητέο.

Στο χαρακτηριστικό παράδειγμα της ερυθράς που αναφέρει ο Hempel ισχύουν τα εξής:

Η πιθανότητα να προσβληθεί από ερυθρά όποιος εκτίθεται σ' αυτή είναι μεγάλη
Ο Α έχει εκτεθεί στην ερυθρά
——— Κάνει πολύ πιθανό (είναι σχεδόν βέβαιο)
Ο Α προσβάλλεται από ερυθρά

Με άλλη διατύπωση, αν όλοι σχεδόν οι άνθρωποι που εκτίθενται στην ερυθρά προσβάλλονται από την πάθηση αυτή και ο Α έχει εκτεθεί στην ερυθρά, τότε είναι σχεδόν βέβαιο ότι ο Α θα προσβληθεί από την πάθηση.

Επισημαίνεται ότι η επαγωγική (ή λογική) πιθανότητα που στηρίζει το εξηγητέο διαφέρει εννοιολογικά από τη στατιστική πιθανότητα, καθώς η πρώτη αφορά σε σχέση μεταξύ προτάσεων και χρησιμοποιείται στην επαγωγική-στατιστική εξήγηση, ενώ η δεύτερη αφορά σε σχέση μεταξύ γεγονότων και χρησιμοποιείται στην παραγωγική-στατιστική εξήγηση. Η στατιστική πιθανότητα ερμηνεύεται ως σχετική συχνότητα των γεγονότων που περιγράφονται στη στατιστική υπόθεση, ενώ η επαγωγική πιθανότητα συνδέει το εξηγούν με το εξηγητέο και αποτελεί μέτρο της δύναμης της επαγωγικής στήριξης ή του βαθμού της λογικής αξιοπιστίας που προσφέρει το εξηγούν στο εξηγητέο.[13]

2.3. Εξήγηση στη βιοϊατρική έρευνα και στη στατιστική

Η παραγωγική εξήγηση θεωρείται έγκυρη με την έννοια ότι το εξηγητέο είναι πάντοτε αληθές εφόσον οι προκείμενες στο εξηγούν είναι αληθείς. Ωστόσο, το μεγάλο μειονέκτημα της παραγωγικής εξήγησης είναι ότι δεν παρέχει τη δυνατότητα επέκτασης της γνώσης πέρα από εκείνα που είναι ήδη γνωστά στις υποθέσεις οι οποίες χρησιμοποιούνται.[7]

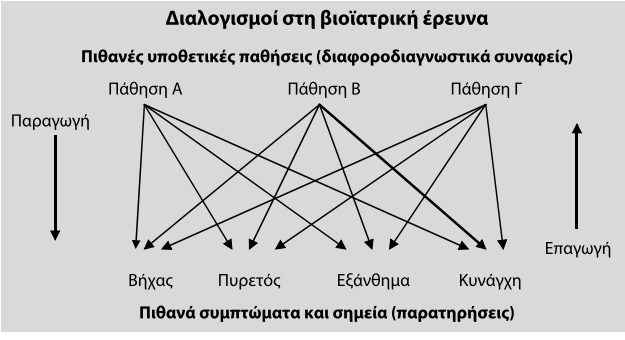
Η επαγωγική εξήγηση λειτουργεί ακριβώς προς την αντίθετη κατεύθυνση. Με βάση τις εμπειρικές παρατηρήσεις, επιλέγεται η πιο πειστική υπόθεση. Η έννοια της ένδειξης, όπως χρησιμοποιείται στη βιοϊατρική έρευνα, αναφέρεται στην επαγωγική εξήγηση, καθώς με βάση τις παρατηρήσεις επάγονται οι υποθέσεις, με την αλήθεια του συμπεράσματος βεβαίως να είναι πιθανή και όχι αναγκαία. Το πλεονέκτημα της επαγωγής είναι ότι τα συμπεράσματα αναφορικά με τις υποθέσεις είναι ευρύτερα σε σχέση με τις εμπειρικές παρατηρήσεις στις οποίες στηρίζονται.[7] Έτσι, η επαγωγή χρησιμοποιείται για να δημιουργηθούν νέες υποθέσεις και να αυξηθεί το πληροφοριακό περιεχόμενο σχετικά με το φυσικό κόσμο. Το πρόβλημα της επαγωγής, εξάλλου, είναι ότι τα συμπεράσματα που προκύπτουν για το φυσικό κόσμο δεν είναι βέβαια.[14]

Στην κλινική πράξη, η εφαρμογή της παραγωγικής εξήγησης είναι σχετικά απλή, καθώς στηρίζεται στην καταγραφή της συχνότητας των συμπτωμάτων ή και των σημείων (εμπειρικές παρατηρήσεις) με δεδομένη την ύπαρξη μιας συγκεκριμένης πάθησης (εικ. 2).[7] Η επαγωγική εξήγηση ωστόσο είναι αρκετά πιο δύσκολη, καθώς αφορά στη διαφορική διάγνωση και πιο συγκεκριμένα στην εύρεση της *πιθανοφάνειας* (likelihood) των διαφοροδιαγνωστικών παθήσεων με δεδομένα τα συμπτώματα, τα σημεία και τα αποτελέσματα των εργαστηριακών εξετάσεων ενός πάσχοντα. Η παραγωγική εξήγηση θεωρείται περισσότερο αντικειμενική από την επαγωγική, αλλά είναι λιγότερο χρήσιμη στην καθημερινή κλινική πράξη.

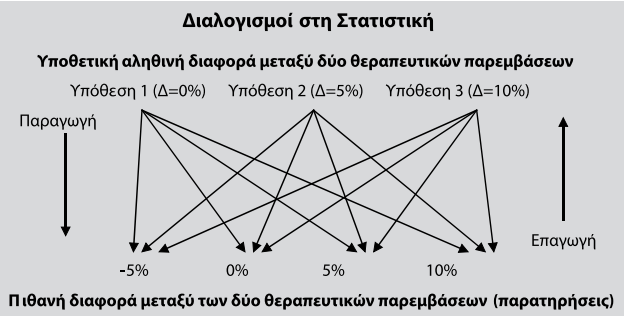
Με αντίστοιχο τρόπο εφαρμόζονται η παραγωγική και η στατιστική εξήγηση και στη στατιστική. Αναλυτικότερα, με δεδομένη την υπόθεση ότι δύο θεραπευτικές παρεμβάσεις έχουν την ίδια αποτελεσματικότητα, είναι απλό να υπολογιστεί με παραγωγικό τρόπο η συχνότητα όλων των πιθανών αποτελεσμάτων που μπορούν να προκύψουν (εικ. 3).[7] Το πρόβλημα ανακύπτει στην περίπτωση που χρειάζεται να απαντηθεί το πλέον σημαντικό επαγωγικό ερώτημα «πόσο πιθανό είναι οι δύο θεραπευτικές παρεμβάσεις να είναι πράγματι εξίσου αποτελεσματικές;», με δεδομένο το αποτέλεσμα μιας συγκεκριμένης κλινικής δοκιμής.

Κατά τη διάρκεια του 20ού αιώνα, οι πλέον σημαντικοί φιλόσοφοι προσπάθησαν να επιλύσουν το πρόβλημα της επαγωγής χωρίς όμως να καταλήξουν σε κατάλληλα μοντέλα ή, αλλιώς, υποδείγματα, με την αδυναμία τους αυτή να φανερώνει ότι δεν υπάρχει μεθοδολογική λύση στο πρόβλημα της ύπαρξης σφάλματος στην επιστημονική γνώση.[7] Ενδεικτικά, ο Popper[11] ήταν τελείως αντίθετος με την επαγωγή χρησιμοποιώντας μόνο τον παραγωγικό τρόπο, ενώ ο Rudolf Carnap[15,16] κινήθηκε προς την αντίθετη κατεύθυνση, προσπαθώντας να καταστήσει την επαγωγική εξήγηση εξίσου λογική με την παραγωγική.

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 από 300 χρόνια περίπου ο Thomas Bayes.* Το 1763, δύο χρόνια μετά από το θάνατο του Bayes, ο στενός του φίλος Richard Price δημοσίευσε το θεώρημα του Bayes, το οποίο



Εικόνα 2. Η διαδικασία της παραγωγικής και της επαγωγικής εξήγησης στη βιοϊατρική έρευνα (απλοποιημένη σχέση). Στην περίπτωση της παραγωγής καταγράφεται η συχνότητα των συμπτωμάτων ή και των σημείων με δεδομένη την ύπαρξη μιας συγκεκριμένης πάθησης. Η επαγωγική εξήγηση αφορά στη διαφορική διάγνωση και πιο συγκεκριμένα στην εύρεση της πιθανοφάνειας των διαφοροδιαγνωστικών παθήσεων με δεδομένα τα συμπτώματα, τα σημεία και τα αποτελέσματα των εργαστηριακών εξετάσεων ενός πάσχοντα.



Εικόνα 3. Η διαδικασία της παραγωγικής και της επαγωγικής εξήγησης στη στατιστική. Με Δ συμβολίζεται η αληθινή διαφορά μεταξύ των δύο θεραπευτικών παρεμβάσεων. Στην περίπτωση της παραγωγής είναι δεδομένη η αληθινή διαφορά ανάμεσα στις δύο παρεμβάσεις, οπότε είναι απλό να υπολογιστεί με παραγωγικό τρόπο η συχνότητα όλων των πιθανών αποτελεσμάτων που μπορούν να προκύψουν. Στην περίπτωση της επαγωγής είναι δεδομένο το αποτέλεσμα μιας συγκεκριμένης κλινικής δοκιμής και αναζητείται απάντηση στο ερώτημα «πόσο πιθανό είναι η αληθινή διαφορά ανάμεσα στις δύο θεραπευτικές παρεμβάσεις να είναι 0%, 5%, 10% κ.λπ.;».

* Ο Άγγλος μαθηματικός και πρεσβυτεριανός αιδεσιμότατος Thomas Bayes (1702–1761) σπούδασε μαθηματικά, λογική και θεολογία, ενώ το θεώρημά του βρέθηκε στα γραπτά του, μετά το θάνατό του.

σχετίζει την εκ των προτέρων πιθανότητα μιας υπόθεσης με την ένδειξη που προέρχεται από μια συγκεκριμένη μελέτη, έτσι ώστε να υπολογιστεί η εκ των υστέρων πιθανότητα της υπόθεσης.[17,18] Από μαθηματική άποψη, το θεώρημα του Bayes δεν εμφανίζει ιδιαίτερα προβλήματα και έχει χρησιμοποιηθεί ευρύτατα στα τυχερά παιχνίδια και στις διαγνωστικές δοκιμασίες στις επιστήμες υγείας. Εντούτοις, από μεθοδολογική άποψη δέχθηκε σημαντική κριτική, κυρίως αναφορικά με την ανάγκη καθορισμού της εκ των προτέρων πιθανότητας μιας υπόθεσης, που θεωρήθηκε υποκειμενική εκτίμηση.[14,16] Αυτός φαίνεται ότι είναι ο σημαντικότερος λόγος για τον οποίο οι επιστήμονες υγείας θεωρούν τον μπεϋζιανό διαλογισμό ευάλωτο στην υποκειμενική κρίση, υιοθετώντας στην πλειοψηφία των περιπτώσεων τις τιμές P και τους ελέγχους των υποθέσεων.

3. ΤΙΜΕΣ P

Η αντίληψη των ερευνητών ότι το θεώρημα του Bayes μειονεκτεί στην υποκειμενική εκτίμηση της εκ των προτέρων πιθανότητας της μηδενικής υπόθεσης είχε ως αποτέλεσμα κατά το χρονικό διάστημα 1920–1940 να αναπτυχθούν εναλλακτικές προσεγγίσεις στο στατιστικό διαλογισμό. Πιο συγκεκριμένα, υιοθετήθηκε η παραγωγική-στατιστική εξήγηση με τον υπολογισμό στατιστικών πιθανοτήτων με βάση μαθηματικές ισότητες που περιέγραφαν (κάτω από ορισμένες υποθέσεις) τη συχνότητα εμφάνισης όλων των πιθανών εκβάσεων ενός πειράματος τύχης (ή μιας μελέτης) εφόσον πραγματοποιούνταν πολλές επαναλήψεις του συγκεκριμένου πειράματος.[16] Ο κύριος εκφραστής της προσέγγισης αυτής, της πιθανότητας ως σχετικής συχνότητας των γεγονότων, ήταν ο Sir Ronald Aylmer Fisher,* ο οποίος μάλιστα εισήγαγε και την ιδέα της τιμής P (P value).[19] Σημειώνεται πάντως ότι ο Fisher απέρριπτε την ερμηνεία της τιμής P ως σχετικής συχνότητας, σε αντίθεση με την πλειοψηφία των ερευνητών ακόμη και σήμερα, χωρίς όμως να έχει δώσει πειστικές απαντήσεις σε ορισμένα σημαντικά ερωτήματα, όπως:[1,16]

- 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.[16] Αναλυτικότερα, στο παράδειγμα αυτό η μηδενική υπόθεση (null hypothesis) είναι ότι το νόμισμα είναι «αμερόληπτο» (fair). «Αμερόληπτο» καλείται 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 ½. Για τον έλεγχο της μηδενικής υπόθεσης πραγματοποιούνται 20 ρίψεις και καταγράφονται τα αποτελέσματα. Η ανάλυση των αποτελεσμάτων, σύμφωνα με τον 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». Συνολικά, πραγματοποιούνται 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

* 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 δικαίως 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 εννέα 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.

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:

$$P(X=x)=\binom{n}{x}p^x q^{n-x} \quad (1)$$

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

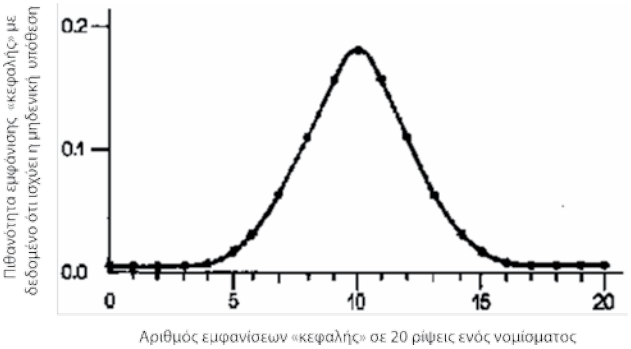
$$P(X=x)=\binom{20}{x}\frac{1}{2}^x\left(\frac{1}{2}\right)^{20-x}$$

$$P(X=x)=\binom{20}{x}\left(\frac{1}{2}\right)^{20} \quad (2)$$

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 μπορούν 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 4.

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

Αριθμός εμφανίσεων κεφαλής (×)	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 4. Διαγραμματική απεικόνιση των πιθανοτήτων εμφάνισης «κεφαλής» × φορές σε ένα πείραμα τύχης, στο οποίο πραγματοποιούνται 20 ρίψεις ενός «αμερόληπτου» νομίσματος.

- 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 προέκυψε πραγματικά. 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:

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

		Πραγματικότητα	
		Αληθής μηδενική υπόθεση	**Ψευδής μηδενική υπόθεση**
Αποτελέσματα ελέγχου της 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)

$$P^{*}=2\times(0,0046+0,0011+2\times10^{-4}+1,9\times10^{-5}+9\times10^{-7})$$

$$P^{*}=0,012$$

- Το τέταρτο βήμα αποτελεί ουσιαστικά μια παραδοχή, σύμφωνα με την οποία η μηδενική υπόθεση πρέπει να απορρίπτεται όταν P* ≤0.05. Ορισμένοι ερευνητές, ωστόσο, προτείνουν η μηδενική υπόθεση να απορρίπτεται όταν P* ≤0.01 ή P* ≤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* ≤α, τότε το αποτέλεσμα καλείται στατιστικά σημαντικό σε επίπεδο σημαντικότητας ίσο με α ενώ απορρίπτεται η μηδενική υπόθεση.

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 απορρίπτονταν στο επίπεδο σημαντικότητας ίσο με 0,05.

Το παραπάνω παράδειγμα είναι ιδιαίτερα διευκρινιστικό όσον αφορά στην έννοια του παρατηρούμενου επιπέδου 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 δεν 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 μόνον 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.[1] Ο Fisher επιδίωκε 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).[20] Επιπλέον, ο Fisher αρνείτο πεισματικά την εφαρμογή του θεωρήματος του Bayes για τη μετατροπή της εκ των προτέρων πιθανότητας μιας υπόθεσης (της πιθανότητας δηλαδή πριν από τη διεξαγωγή μιας συγκεκριμένης μελέτης) σε εκ των υστέρων πιθανότητα της ίδιας υπόθεσης με βάση τα δεδομένα της μελέτης. Η άρνησή του αυτή οφειλόταν στο γεγονός ότι θεωρούσε τις εκ των προτέρων πιθανότητες κατά κανόνα αβέβαιες και υποκειμενικές. Έτσι, πρότεινε τρεις μεθόδους που δεν ελάμβαναν υπόψη την εκ των προτέρων πιθανότητα μιας υπόθεσης: (α) Η πρώτη μέθοδος βασιζόταν στην τιμή P, (β) η δεύτερη βασιζόταν στην έννοια της πιθανοφάνειας και (γ) η τρίτη έμεινε γνωστή με τον όρο «αξιόπιστος διαλογισμός» (fiducial inference), χωρίς όμως να έχει ιδιαίτερη απήχηση, καθώς σε γενικές γραμμές κρίθηκε ανεπαρκής.[20]

Ο Fisher δεν ήταν ο πρώτος ερευνητής που χρησιμοποίησε την τιμή P,[17,21] 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.[22] 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 >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. Αυτή η 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 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.[19,23] Η τιμή P σε καμιά περίπτωση δεν αποτελεί στοιχείο 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 τον Fisher, η τιμή P δεν πρέπει να ερμηνεύεται ως η υποθετική σχετική συχνότητα του σφάλματος έπειτα από πολλές επαναλήψεις του πειράματος τύχης. Στη βιοϊατρική έρευνα, 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 ανάμεσα στη μηδενική υπόθεση και τα δεδομένα μιας μελέτης. Η διαδικασία αυτή, 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 της 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 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.[23]

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 της υπόθεσης των Neyman και Pearson συμπεριλαμβάνονται τόσο η μηδενική όσο και η εναλλακτική υπόθεση.

Σε αρκετές περιπτώσεις, οι επιστήμονες υγείας παρερμηνεύουν την τιμή P, καταλήγοντας σε παραπλανητικά συμπεράσματα.[25–28] Αρκετοί πιστεύουν ότι μια τιμή P ίση με 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 βέβαια με τη χρήση των τιμών P ή των ελέγχων των υποθέσεων.[26,29] Η ένδειξη μιας μελέτης συνίσταται από την εμπειρική τιμή του μέτρου σχέσης (ή σπανιότερα του μέτρου συχνότητας) που υπολογίζεται και το αντίστοιχο διάστημα εμπιστοσύνης (μέσω του υπολογισμού του τυπικού σφάλματος), που αποτελεί μέτρο της ακρίβειας της μέτρησης.

Ορισμένοι ερευνητές στράφηκαν εξαρχής εναντίον της θεωρίας του Fisher σχετικά με την τιμή P, θεωρώντας ότι στερείται τόσο λογικής βάσης όσο και πρακτικής χρησιμότητας.[30,31] Το ισχυρότερο 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, το οποίο 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

* 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.[24] 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), μοναδικός υιός του Karl Pearson, ακολουθώντας το πρότυπο του πατέρα του υπήρξε ένας εξαιρετικός στατιστικός. Διατέλεσε καθηγητής στατιστικής στο University College, στο Λονδίνο, ενώ εξαιρετικά σημαντική ήταν η συνεισφορά του στην καθιέρωση του περιοδικού "Biometrika" ως ενός 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.

μικρό μέγεθος «δείγματος». Για το λόγο αυτόν, τα τελευταία χρόνια καταβάλλεται συστηματική προσπάθεια ευρύτερης χρησιμοποίησης των διαστημάτων εμπιστοσύνης έναντι των τιμών P.[22]

4. ΕΛΕΓΧΟΣ ΤΩΝ ΥΠΟΘΕΣΕΩΝ

4.1. 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.[32] Η θεωρία των Neyman και Pearson έγινε ευρέως γνωστή με τον όρο «έλεγχος της υπόθεσης». Σ' έναν έλεγχο της 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.[33] 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 καλείται εναλλακτική υπόθεση (alternative hypothesis) και σύμφωνα με αυτή υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης. Στην περίπτωση αυτή, ο έλεγχος της υπόθεσης καλείται 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 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 της 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 της 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 της 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 σχέση αυτή είναι θετική ή αρνητική. Εάν πραγματοποιηθεί έλεγχος της 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 σε 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 δεν 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.

4.2. Σφάλματα τύπου I και II

Όπως προαναφέρθηκε, ο έλεγχος των υποθέσεων προϋποθέτει τον καθορισμό της μηδενικής και της εναλλακτικής υπόθεσης, αλλά και τον προσδιορισμό του σφάλματος τύπου I και II (πίν. 2).[2,4,22,34,35]

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 αληθής. Ουσιαστικά, το σφάλμα τύπου I είναι το ποσοστό των ψευδώς θετικών αποτελεσμάτων των ελέγχων των υποθέσεων, όπου λανθασμένα συμπεραίνεται ότι υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, καθώς στην πραγματικότητα δεν υφίσταται η σχέση αυτή. Η πιθανότητα να διαπραχθεί ένα σφάλμα τύπου I καθορίζεται από το προκαθορισμένο επίπεδο σημαντικότητας του ελέγχου της υπόθεσης που είναι η τιμή α. Εάν πραγματοποιηθούν κατ' επανάληψη έλεγχοι

της 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 της 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 των υποθέσεων, όπου λανθασμένα συμπεραίνεται ότι δεν υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, καθώς στην πραγματικότητα υφίσταται η σχέση αυτή.

Εάν το β είναι η πιθανότητα να διαπραχθεί ένα σφάλμα τύπου II, τότε 1 – β είναι η *στατιστική ισχύς* (statistical power) του ελέγχου της 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 των υποθέσεων, όπου σωστά συμπεραίνεται ότι υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης, καθώς στην πραγματικότητα η σχέση αυτή υφίσταται.

Σε γενικές γραμμές, ο στόχος των ερευνητών είναι ο σχεδιασμός ελέγχων των 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. Αντιστρόφως, μείωση της τιμής α προκαλεί μείωση του σφάλματος τύπου I και αύξηση του σφάλματος τύπου II.

4.3. Κριτική

Στον «έλεγχο της 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, κάτι που επιβεβαιώνεται άλλωστε και από τους ίδιους τους ερευνητές.[32] Και στις δύο μεθόδους, πάντως, 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.[1] Η μαθηματική και εννοιολογική προσέγγιση των «ελέγχων των υποθέσεων» αποτέλεσε ένα σημαντικό βήμα, αλλά η ερμηνεία τους και η πρακτική τους εφαρμογή αντιμετωπίζουν ακόμη και σήμερα σοβαρά προβλήματα. Οι Neyman και Pearson δεν χρησιμοποίησαν κάποιο μέτρο της ένδειξης που παρέχουν τα δεδομένα μιας μελέτης ακριβώς για να αποφύγουν οποιαδήποτε συσχέτιση με το θεώρημα του Bayes. 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.[36] Οι έλεγχοι των υποθέσεων είναι ισοδύναμοι ουσιαστικά με ένα νομικό σύστημα, που δεν επικεντρώνεται στο αν ένας συγκεκριμένος κατηγορούμενος βρεθεί ένοχος ή αθώος (αντίστοιχα, στο βιοϊατρικό διαλογισμό, 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 δικαίου υπαγορεύει τη σωστή ετυμηγορία για ένα συγκεκριμένο κατηγορούμενο, έτσι ακριβώς και στη βιοϊατρική έρευνα 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

* Οι αρχές (αξιώματα και αιτήματα) είναι άμεσα βέβαιες και δεν παράγονται από άλλες.

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 των υποθέσεων από τους Neyman και Pearson.[23]

Είναι σαφές ότι ακόμη και σήμερα ένας σημαντικός αριθμός περιοδικών που αφορούν στη βιοϊατρική έρευνα, 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. Τα τελευταία 30 χρόνια έχει ασκηθεί δριμύτατη κριτική στη χρήση των ελέγχων των υποθέσεων για την εξαγωγή συμπερασμάτων στη βιοϊατρική έρευνα. Χαρακτηριστικά αναφέρεται ότι το 1997 η International Committee of Medical Journal Editors[37] εξέδωσε οδηγίες για τη δημοσίευση μελετών τις οποίες υιοθέτησαν >500 περιοδικά παγκοσμίως (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 των 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 έκβασης όσο και το μέγεθος του τυχαίου σφάλματος. Ο Kenneth Rothman, ως διευθυντής σύνταξης ενός από τα πλέον έγκριτα περιοδικά επιδημιολογίας (του *Epidemiology*), είναι κατηγορηματικός στο ότι δεν πρέπει να δημοσιεύονται μελέτες, τα συμπεράσματα των οποίων στηρίζονται στην ύπαρξη ή όχι στατιστικής σημαντικότητας.[38] Επιπλέον, μολονότι πολλά περιοδικά συνιστούν τη χρήση των ελέγχων των υποθέσεων και των τιμών P, ο Rothman τάσσεται σαφώς εναντίον τους θεωρώντας απαραίτητη 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 έναν ή περισσότερους προσδιοριστές, αλλά στην εύρεση πιθανών παραγόντων (π.χ. όπως η ύπαρξη συστηματικών σφαλμάτων ή συγχυτών) που μπορούν να συμβάλλουν στην ερμηνεία των αποτελεσμάτων μιας μελέτης. Τα τελευταία 20 χρόνια, εξάλλου, ορισμένοι από τους κορυφαίους επιδημιολόγους, όπως ο Olli Miettinen,[39] ο Sander Greenland,[40] ο Steven Goodman[1,7,20,41] κ.ά., καταβάλλουν συστηματικές προσπάθειες για την περιθωριοποίηση των ελέγχων των υποθέσεων και την υιοθέτηση της επαγωγικής μεθόδου, που εισήχθη για πρώτη φορά από τον Bayes, με τον υπολογισμό του παράγοντα Bayes και τη χρησιμοποίηση της εκ των προτέρων πιθανότητας μιας υπόθεσης.

Προκαλεί αρκετά ερωτηματικά το γεγονός ότι η εξαγωγή συμπερασμάτων στη βιοϊατρική έρευνα* εξακολουθεί δυστυχώς ακόμη και σήμερα να στηρίζεται στους ελέγχους των υποθέσεων, μιας παραγωγικής μεθόδου που δεν συμβάλλει ουσιαστικά στην αύξηση του πληροφοριακού περιεχομένου για το φυσικό κόσμο, καθώς δεν 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 όμοιων μεταξύ τους. Η τιμή P που προκύπτει από την ανάλυση των δεδομένων μιας μελέτης χρησιμοποιείται στους ελέγχους των υποθέσεων για τη διαπίστωση της ύπαρξης ή όχι στατιστικής σημαντικότητας, μολονότι –σύμφωνα και με τον 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 η απλοϊκότητα ίσως) με την οποία, μέσω της εφαρμογής των ελέγχων των υποθέσεων, 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 απαραίτητη η εφαρμογή εξαιρετικά σύνθετων στατιστικών δοκιμασιών. Με τον τρόπο αυτό, μέσα σε μικρό χρονικό

* Ο όρος «έρευνα» (research) χρησιμοποιείται σήμερα με πολλούς διαφορετικούς τρόπους, χωρίς οι περισσότεροι να σκέπτονται το τι ακριβώς σημαίνει. Ο Peyton Rous (1879–1970), βραβευμένος με Nobel σε ηλικία 87 ετών για την ανακάλυψή του (σε ηλικία 31 ετών) ότι ορισμένοι τύποι καρκίνου μεταδίδονται στα ζώα μέσω ιών, ήταν ο πρώτος που διέκρινε τον όρο "REsearch" από τον όρο "reSEARCH", με τον πρώτο να δηλώνει την επανειλημμένη αναζήτηση και το δεύτερο την ανακάλυψη. Η έρευνα περιλαμβάνει τέσσερα βήματα: (α) Την παρατήρηση, (β) την αντιστοίχιση της παρατήρησης με την προϋπάρχουσα εμπειρία, (γ) την επινόηση μιας υπόθεσης για την ερμηνεία της παρατήρησης και (δ) τον πειραματικό έλεγχο της υπόθεσης. Το τελευταίο βήμα μπορεί να μην είναι άμεσα δυνατό, με τη θεωρία της σχετικότητας, π.χ., του Albert Einstein να υφίσταται πειραματικό έλεγχο αρκετά χρόνια μετά από τη διατύπωσή της. Ο όρος "biomedical" (βιοϊατρικός), εξάλλου, αποτελεί αμάλγαμα των όρων "biological" (βιολογικός) και "medical" (ιατρικός). Ο όρος "biomedical" χρησιμοποιήθηκε για πρώτη φορά το 1947 από τον Arno Benedict Luckhardt (Αμερικανός Καθηγητής Φυσιολογίας, 1885–1957) στον πρόλογο του βιβλίου "A history of scientific English" του Edmund Andrews. Ο όρος "biomedical" χρησιμοποιήθηκε ευρύτατα μετά το Β΄ Παγκόσμιο Πόλεμο, αναφορικά με τα πειράματα που διεξήχθησαν στον Ειρηνικό Ωκεανό για τον έλεγχο των επιδράσεων σε πειραματόζωα της ακτινοβολίας που εκλύεται από την έκρηξη ατομικών βομβών.

διάστημα μπορούν να πραγματοποιηθούν πολυάριθμοι στατιστικοί έλεγχοι των υποθέσεων (κάτι εξαιρετικά σύνηθες στη γενετική επιδημιολογία, όπου διερευνάται η σχέση ανάμεσα σε χιλιάδες γονίδια και στη συχνότητα εμφάνισης μιας πάθησης) 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 των υποθέσεων, αλλά φαίνεται κιόλας ότι η προσέγγιση αυτή παρέχει την άνεση σχετικά εύκολα να 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 συμβάλει και η πολιτική των περισσότερων περιοδικών που αφορούν στη βιοϊατρική έρευνα, 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 των 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 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 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.[22,34,42–45] Τα *διαστήματα εμπιστοσύνης* (confidence intervals) υποδηλώνουν το μέγεθος του τυχαίου σφάλματος σε μια μελέτη, εφόσον όμως (α) δεν υπάρχουν συστηματικά σφάλματα και συγχυτές και (β) χρησιμοποιείται το κατάλληλο στατιστικό μοντέλο.[22] Όσο μικρότερο είναι το εύρος ενός διαστήματος εμπιστοσύνης* τόσο μικρότερο είναι και το τυχαίο σφάλμα της μέτρησης. Εντούτοις, και τα διαστήματα εμπιστοσύνης παρουσιάζουν σημαντικά μειονεκτήματα, καθώς υπολογίζονται με βάση τις μαθηματικές ισότητες που χρησιμοποιούνται και στους ελέγχους των υποθέσεων, ενώ δεν λαμβάνουν υπόψη τους την προϋπάρχουσα ένδειξη και τους βιολογικούς μηχανισμούς. Πρόκειται και πάλι για εφαρμογή του παραγωγικού διαλογισμού, με το πλεονέκτημα τουλάχιστον ότι δεν εφαρμόζονται για την απόρριψη ή όχι μιας υπόθεσης, αλλά για την εκτίμηση του μεγέθους της σχέσης μεταξύ προσδιοριστή και έκβασης, με την προϋπόθεση ότι στη μέτρηση υπεισέρχεται μόνο το τυχαίο σφάλμα. Η προσπάθεια περιορισμού της χρήσης των ελέγχων των υποθέσεων και της ευρύτερης εφαρμογής των διαστημάτων εμπιστοσύνης αντιμετωπίζεται με διστακτικότητα από τους περισσότερους επιστήμονες υγείας, οι οποίοι λανθασμένα θεωρούν την τιμή P ως ένα ισχυρό μέτρο που τους παρέχει τη δυνατότητα λήψης «αντικειμενικών» αποφάσεων.[46,47]

Τονίζεται πάντως ότι και τα διαστήματα εμπιστοσύνης, μολονότι κινούνται προς τη σωστή κατεύθυνση, δεν επιλύουν όλα τα προβλήματα που παρουσιάζουν οι έλεγχοι των υποθέσεων.[48] Το πλέον σημαντικό μειονέκτημα των διαστημάτων εμπιστοσύνης είναι ότι δεν λαμβάνεται υπόψη η ένδειξη που προέρχεται από προγενέστερες μελέτες σχετικά με μια συγκεκριμένη επιστημονική υπόθεση. Η απάντηση στο πρόβλημα αυτό, εξάλλου, δίνεται με την εφαρμογή των μπεϋζιανών μεθόδων και πιο συγκεκριμένα με τον υπολογισμό του παράγοντα Bayes, που αποτελεί το μπεϋζιανό μέτρο της ένδειξης που προέρχεται από μια μελέτη και το οποίο μεταβάλλει όχι μόνο την παρουσίαση των αποτελεσμάτων αλλά, κυρίως, και τον τρόπο σκέψης (επαγωγικό πλέον) των ερευνητών.[1,7,20,26,39,41,48–50]

Σημειώνεται ότι είναι τελείως λανθασμένη η χρησιμοποίηση των διαστημάτων εμπιστοσύνης για την απόρριψη ή όχι της μηδενικής υπόθεσης, ανάλογα με το αν το διάστημα εμπιστοσύνης περιέχει τη μηδενική τιμή (null value).[22,51] 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. Τονίζεται και πάλι ότι τα διαστήματα εμπιστοσύνης δεν πρέπει να χρησιμοποι-

* Το επιλεγόμενο διάστημα εμπιστοσύνης είναι αυθαίρετο και καθορίζεται από τον ερευνητή, με την τιμή του να κυμαίνεται μεταξύ 0–100%.[45] Στην πλειοψηφία των περιπτώσεων υπολογίζεται το 95% διάστημα εμπιστοσύνης και σπανιότερα το 90% ή το 99% διάστημα εμπιστοσύνης. 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 δεν 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» θα πρέπει να γίνει με τον ίδιο ακριβώς τρόπο.

ούνται ως υποκατάστατα των ελέγχων των υποθέσεων για τη διαπίστωση στατιστικά σημαντικών σχέσεων, αλλά για την εκτίμηση του τυχαίου σφάλματος και του μεγέθους της σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης.

4.4. Πολλαπλοί έλεγχοι των υποθέσεων

Ο Miettinen[39] πρόσφατα επανέφερε στο προσκήνιο ένα εξαιρετικά σημαντικό μειονέκτημα που παρουσιάζει η θεωρία του ελέγχου υποθέσεων και αποτελεί σημείο έντονων αντιπαραθέσεων. Σε αρκετές περιπτώσεις, η ανάλυση των δεδομένων μιας μελέτης μπορεί να περιλαμβάνει έναν πολύ μεγάλο αριθμό ελέγχων των υποθέσεων ανάλογα με τον αριθμό των προσδιοριστών και των εκβάσεων που διερευνώνται. Εάν, π.χ., σε μια μελέτη διερευνώνται 40 διατροφικά σχήματα, 100 διατροφικά συστατικά και 30 παθήσεις, τότε ο ελάχιστος αριθμός ελέγχων των υποθέσεων θα είναι 120.000.[52] Σήμερα, πρόκειται για μια εξαιρετικά συνηθισμένη κατάσταση στη γενετική επιδημιολογία, όπου η αποκωδικοποίηση του ανθρώπινου υλικού οδήγησε στην αναγνώριση περίπου 30.000 γονιδίων, καθένα από τα οποία μπορεί να σχετίζεται με τη συχνότητα εμφάνισης οποιασδήποτε πάθησης.[53] Έτσι, εάν σε μια μελέτη διεξαχθούν πολλαπλοί έλεγχοι των υποθέσεων και η τιμή α διατηρηθεί ίση με 0,05 σε κάθε έλεγχο της υπόθεσης ξεχωριστά, τότε αυξάνεται σημαντικά το ποσοστό των ψευδώς θετικών αποτελεσμάτων των ελέγχων, κάτι που πρέπει να ληφθεί σοβαρά υπόψη πριν από την ανάλυση των δεδομένων.[54] Στην περίπτωση αυτή, εφαρμόζονται ορισμένες στατιστικές δοκιμασίες, μεταξύ των οποίων η συχνότερη είναι η *διόρθωση κατά Bonferroni* (Bonferroni' s correction), σύμφωνα με την οποία εάν κατά την ανάλυση των δεδομένων μιας μελέτης πρόκειται να διεξαχθούν n έλεγχοι, τότε η συνολική τιμή α (για όλους τους ελέγχους της ανάλυσης) διαιρείται με το συνολικό αριθμό των ελέγχων, που είναι ίσος με n.

Εάν, π.χ., η συνολική τιμή α για όλους τους ελέγχους είναι 0,05 και πρόκειται να διεξαχθούν 100 έλεγχοι, τότε η τιμή α για καθέναν από τους ελέγχους αυτούς ξεχωριστά θα είναι ίση με $\frac{\alpha}{n} = \frac{0,05}{100} = 0,0005$, ενώ αν πρόκειται να διεξαχθούν 500 έλεγχοι, τότε η τιμή α για κάθε έλεγχο ξεχωριστά θα είναι ίση με $\frac{\alpha}{n} = \frac{0,05}{500} = 0,0001$. Με τον τρόπο αυτό μειώνεται η τιμή α για κάθε έλεγχο ξεχωριστά, έτσι ώστε η συνολική τιμή α της μελέτης να διατηρηθεί ίση με 0,05. Είναι σαφές ότι εάν δεν εφαρμοστεί η διόρθωση κατά Bonferroni,* τα αποτελέσματα μπορεί να είναι τελείως παραπλανητικά, καθώς μια τιμή P ίση με 0,001 είναι στατιστικά σημαντική (οδηγώντας δυστυχώς στο συμπέρασμα ότι υπάρχει σχέση μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης) χωρίς τη διόρθωση κατά Bonferroni, αλλά δεν είναι στατιστικά σημαντική έπειτα από τη διόρθωση κατά Bonferroni και εφόσον η ανάλυση των δεδομένων της μελέτης περιλαμβάνει τουλάχιστον 50 ελέγχους των υποθέσεων.

Δυστυχώς, αρκετοί θεωρούν ότι το πρόβλημα των πολλαπλών συγκρίσεων επιλύεται με την εφαρμογή επιπλέον στατιστικών μεθόδων, όπως η διόρθωση κατά Bonferroni, γεγονός όμως που δεν ισχύει στην πραγματικότητα. Η μοναδική αξιόπιστη λύση στο πρόβλημα αυτό είναι η εφαρμογή μπεϋζιανών μεθόδων και ειδικότερα ο υπολογισμός του παράγοντα Bayes, καθώς στην περίπτωση αυτή λαμβάνεται υπόψη τόσο η ένδειξη μιας συγκεκριμένης μελέτης όσο και η εκ των προτέρων πιθανότητα μιας υπόθεσης.[22,55–59] Ο Miettinen[39] σημειώνει χαρακτηριστικά ότι η εκ των προτέρων πιθανότητα της μηδενικής υπόθεσης να μην υπάρχει σχέση ανάμεσα σ' ένα αυθαίρετα επιλεγμένο γονίδιο και στη συχνότητα εμφάνισης μιας συγκεκριμένης πάθησης είναι εξαιρετικά μεγάλη, οπότε η εύρεση στατιστικά σημαντικής σχέσης –έπειτα από τη σύγκριση της τιμής P με την τιμή α– είναι εξαιρετικά πιθανό (με πιθανότητα κοντά στο 1) να οφείλεται στην τύχη παρά στην πραγματικότητα. Για το λόγο αυτόν, είναι απαραίτητο να ληφθούν σοβαρά υπόψη, μέσω της εφαρμογής του θεωρήματος του Bayes, τόσο η ένδειξη μιας συγκεκριμένης μελέτης όσο και η εκ των προτέρων πιθανότητα της μηδενικής υπόθεσης (της μη ύπαρξης σχέσης μεταξύ προσδιοριστή και συχνότητας εμφάνισης της έκβασης), έτσι ώστε να υπολογιστεί η εκ των υστέρων πιθανότητα της μηδενικής αυτής υπόθεσης. Τονίζεται και πάλι ότι η εφαρμογή πολλαπλών ελέγχων των υποθέσεων στην ανάλυση των δεδομένων μιας μελέτης –και ιδιαίτερα στην περίπτωση της γενετικής επιδημιολογίας– απαιτεί ιδιαίτερη περίσκεψη για την αποφυγή λανθασμένων συμπερασμάτων.[21,52,55,60–63]

5. ΣΥΝΟΨΗ

Είναι σαφές ότι η χρήση των τιμών P και των ελέγχων των υποθέσεων για τη διαπίστωση (με παραγωγικό τρόπο) 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

* Ο Ιταλός μαθηματικός Carlo Emilio Bonferroni (1892–1960) διατέλεσε από το 1933 καθηγητής μαθηματικών στο πανεπιστήμιο της Φλωρεντίας. Το επιστημονικό του έργο επικεντρώθηκε στη θεωρία των πιθανοτήτων και στη γεωμετρία.

στον Fisher, που εισήγαγε ουσιαστικά την έννοια των τιμών P, και στους Neyman και Pearson, τους θεμελιωτές της θεωρίας των ελέγχων των 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, εκφράζοντας παράλληλα την αξιοπιστία της μηδενικής υπόθεσης σε σχέση με τα δεδομένα της συγκεκριμένης μελέτης. Ο Fisher μάλιστα όρισε την τιμή P όπως ακριβώς ορίζεται και σήμερα, τονίζοντας ότι 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.

Ο Fisher σωστά θεωρούσε ότι η εξαγωγή συμπερασμάτων πρέπει να βασίζεται στην επαγωγική λογική και προς την κατεύθυνση αυτή εισήγαγε μια νέα θεωρία στηριζόμενη στην έννοια της πιθανοφάνειας. Οι Neyman και Pearson, εντούτοις, θεωρούσαν αδύνατη την εφαρμογή του επαγωγικού διαλογισμού για την εξαγωγή συμπερασμάτων και ακριβώς για να αποφύγουν οποιαδήποτε συσχέτιση με το θεώρημα του Bayes δεν 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. 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 ίση με 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, απλώς 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 αυτόν 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 δεν 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 στη βιοϊατρική έρευνα 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).

ABSTRACT

The wrong application of P values and hypotheses test in biomedical research

P. GALANIS

Center for Health Services Management and Evaluation, Department of Nursing, University of Athens, Athens, Greece

Archives of Hellenic Medicine 2010, 27(4):691–707

Data analysis and interpretation of results in biomedical research continues even today to present problems. In particular, in the majority of cases, conclusions are based on hypotheses test and P values without taking into account the biological settings and the evidence from previous studies concerning a specific hypothesis. This form of deductive inference does not permit increase of knowledge about the natural world, and for this reason the abandonment of this type of inference and the adoption of Bayesian (inductive) methods is recommended, and specifically, the calculation of a measure known as the Bayes factor. The P value is a measure of discrepancy between the data derived from a study and the null hypothesis, and represents the probability, assuming that the null hypothesis is true, of obtaining a result as far as, or further than, what was actually obtained in a particular study. The application of hypotheses test aimed at the constraint of P values, has led in the opposite direction, with P value being used as a criterion for decision making. The P value does not constitute a component of typical inference and is used wrongly by most health scientists for the ascertainment of relationship between a determinant and frequency of occurrence of an outcome. The application of hypotheses test rejects, in practice, the application of inductive inference in each study separately for extracting conclusions, since the hypothesis test is a deductive method for restriction of error rate after the iteration of a particular study. It is encouraging that lately the efforts for the restriction of P values and the application of confidence intervals in the presentation of the results of a study have yielded fruit. Confidence intervals have considerable disadvantages, the most important being that they do not relate the evidence from a particular study with that from previous studies. This disadvantage of confidence intervals is overcome using Bayesian methods and especially with the calculation of Bayes factor.

Key words: Bayes factor, Confidence interval, Explanation, Hypothesis test, Inference, P value

Βιβλιογραφία

1. GOODMAN S. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med* 1999, 130:995–1004
2. ΣΠΑΡΟΣ ΛΔ, ΓΑΛΑΝΗΣ Π. *Δοκίμια Επιδημιολογίας.* Εκδόσεις Παρισιάνου, Αθήνα, 2006
3. MIETTINEN OS. *Theoretical epidemiology. Principles of occurrence research in medicine.* John Wiley & Sons, New York, 1985
4. ΓΑΛΑΝΗΣ ΠΑ, ΣΠΑΡΟΣ ΛΔ. *Εγχειρίδιο Επιδημιολογίας.* Εκδόσεις ΒΗΤΑ, Αθήνα, 2010
5. ΣΠΑΡΟΣ Λ, ΔΕΛΗΟΛΑΝΗΣ Ι. Ιατρική βασιζόμενη στη γνώση. *Αρχ Ελλ Ιατρ* 2008, 25:389–395
6. SACKETT DL, STRAUS SE, RICHARDSON SW, ROSENBERG W, HAYNES BR. *Επί ενδείξεων βασιζόμενη Ιατρική. Πώς να ασκείται και να διδάσκεται η EBI* (Ελληνική μετάφραση: Ε. Ανευλαβής). Εκδόσεις Πασχαλίδης, Αθήνα, 2002:21–55
7. GOODMAN S. Toward evidence-based medical statistics. 2: The Bayes Factor. *Ann Intern Med* 1999, 130:1005–1013
8. ΓΕΜΤΟΣ ΠΑ. *Μεθοδολογία των κοινωνικών επιστημών*. Τόμος 2ος, 3η έκδοση. Εκδόσεις Παπαζήση, Αθήνα, 1987:13–33, 199–209, 237–248
9. SALMON MH, EARMAN J, GLYMOUR C, LENNOX JG. *Introduction to the philosophy of science.* Prentice-Hall Inc, New Jersey, 1992
10. ΣΠΑΡΟΣ ΛΔ, ΓΑΛΑΝΗΣ Π, ΖΑΧΟΣ Ι, ΤΣΙΛΙΔΗΣ Κ. *Επιδημιολογία Ι.* Εκδόσεις ΒΗΤΑ, Αθήνα, 2004:120–129
11. POPPER KR. *The logic of scientific discovery.* Basic Books, New York, 1959
12. HEMPEL CG, OPPENHEIM P. Studies in the logic of explanation. *Philosophy Science* 1948, 15:135–175
13. CARNAP R. *An introduction to the philosophy of science*. Dover Publ Inc, New York, 1995
14. HACKING I. *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference.* Cambridge University Press, Cambridge, 1975
15. CARNAP R. *Logical foundations of probability*. Chicago University Press, Chicago, 1950
16. HOWSON C, URBACH P. *Scientific reasoning: The Bayesian approach.* 2nd ed. Open Court Publishing Company, Chicago, 1993
17. STIGLER SM. *The history of statistics: The measurement of uncertainty before 1900.* Harvard University Press, Cambridge, 1986
18. BAYES T. Essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society* 1763, 53:370–418

19. FISHER RA. *Statistical methods for research workers.* 13th ed. Hafner, New York, 1958
20. GOODMAN S. P values, hypothesis tests and likelihood: Implications for epidemiology of a neglected historical debate. *Am J Epidemiol* 1993, 137:485–496
21. DAVID HA. First (?) occurrence of common terms in mathematical statistics. *American Statistician* 1995, 49:121–133
22. ROTHMAN KJ, GREENLAND S, LASH TL. *Modern epidemiology*. 3rd ed. Lippincott Williams & Wilkins, Philadelphia, 2008
23. FISHER RA. *Statistical methods and scientific inference*. 3rd ed. Macmillan, New York, 1973
24. NEYMAN J. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society, Series A* 1937, 236:333–380
25. BROWNER W, NEWMAN T. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. *JAMA* 1987, 257:2459–2463
26. DIAMOND GA, FORRESTER JS. Clinical trials and statistical verdicts: Probable grounds for appeal. *Ann Intern Med* 1983, 98:385–394
27. LILFORD RJ, BRAUNHOLTZ D. For debate: The statistical basis of public policy: A paradigm shift is overdue. *Br Med J* 1996, 313:603–607
28. FREEMAN PR. The role of P-values in analysing trial results. *Stat Med* 1993, 12:1442–1552
29. BROPHY JM, JOSEPH L. Placing trials in context using Bayesian analysis. GUSTO revisited by Reverend Bayes. *JAMA* 1995, 273:871–875
30. BERKSON J. Tests of significance considered as evidence. *JASA* 1942, 37:325–335
31. PEARSON E. "Student" as a statistician. *Biometrika* 1938, 38:210–250
32. NEYMAN J, PEARSON E. On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* 1928, 20:175–240
33. LEHMANN EL. *Testing statistical hypotheses*. 2nd ed. Wiley, New York, 1986
34. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ ΛΔ. Στατιστικά μοντέλα για την ανάλυση των επιδημιολογικών δεδομένων. *Αρχ Ελλ Ιατρ* 2006, 23:404–417
35. GREENBERG RS, DANIELS SR, FLANDERS DW, ELEY WJ, BORING JR. *Medical epidemiology.* Prentice-Hall International, London, 1993
36. NEYMAN J, PEARSON E. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society* 1933, 231:289–337
37. INTERNATIONAL COMMITTEE OF MEDICAL JOURNAL EDITORS. Uniform requirements for manuscripts submitted to biomedical journals. *N Engl J Med* 1997, 36:309–315
38. ROTHMAN KJ. Writing for epidemiology. *Epidemiology* 1998, 9:333–337
39. MIETTINEN OS. Up from "false positives" in genetic –and other– epidemiology. *Eur J Epidemiol* 2009, 24:1–5
40. GREENLAND S. Probability logic and probabilistic induction. *Epidemiology* 1998, 9:322–332
41. GOODMAN SN, ROYALL R. Evidence and scientific research. *Am J Public Health* 1988, 78:1568–1574
42. ROTHMAN KJ. Significance questing (editorial). *Ann Intern Med* 1986, 105:445–447
43. BARNETT ML, MATHISEN A. Tyranny of the P-value: The conflict between statistical significance and common sense (editorial). *J Dent Res* 1997, 76:534–536
44. BAILAR JC 3rd, MOSTELLER F. Guidelines for statistical reporting in articles for medical journals. Amplifications and explanations. *Ann Intern Med* 1988, 108:266–273
45. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ ΛΔ. Ανάλυση δεδομένων: Μη μπαγιεσιανή προσέγγιση. *Αρχ Ελλ Ιατρ* 2005, 22:377–391
46. LANG JM, ROTHMAN KJ, CANN CI. That confounded P-value (editorial). *Epidemiology* 1998, 9:7–8
47. EVANS SJ, MILLS P, DAWSON J. The end of the P value? *Br Heart J* 1988, 60:177–180
48. FREEDMAN L. Bayesian statistical methods (editorial). *Br Med J* 1996, 313:569–570
49. ETZIONI RD, KADANE JB. Bayesian statistical methods in public health and medicine. *Annu Rev Public Health* 1995, 16:23–41
50. KADANE JB. Prime time for Bayes. *Control Clin Trials* 1995, 16:313–318
51. POOLE C. Beyond the confidence interval. *Am J Public Health* 1987, 77:195–199
52. GREENLAND S. Multiple comparisons and association selection in general epidemiology. *Int J Epidemiol* 2008, 17:430–434
53. ORR HA. *The genetic adventurer.* New York Rev Books, New York, 2008:20
54. BOFFETTA P, McLAUGHLIN J, La VECCHIA C, TARONE R, LIPWORTH L, BLOT WZ. False-positive results in cancer epidemiology: A plea for epistemological modesty. *J Natl Cancer Inst* 2008, 100:988–995
55. GREENLAND S, ROBINS JM. Empirical-Bayes adjustments for multiple comparisons are sometimes useful. *Epidemiology* 1991, 2:244–251
56. GREENLAND S. Variable selection and shrinkage in the control of multiple confounders. *Am J Epidemiol* 2008, 167:523–529
57. GREENLAND S. Bayesian methods for epidemiologic research. II. Regression analysis. *Int J Epidemiol* 2007, 36:195–202
58. THOMAS DC, SEMIATYCKI J, DEWAR R, ROBINS J, GOLDBERG M, ARMSTRONG BG. The problem of multiple inference in studies designed to generate hypotheses. *Am J Epidemiol* 1985, 122:1080–1095
59. GREENLAND S. Methods for epidemiologic analyses of multiple exposures: A review and comparative study of maximum likelihood, preliminary testing, and empirical-Bayes regression. *Stat Med* 1993, 12:717–736
60. GOODMAN S. Multiple comparisons, explained. *Am J Epidemiol* 1998, 147:807–812
61. ROTHMAN KJ. No adjustments are needed for multiple comparisons. *Epidemiology* 1990, 1:43–46
62. POOLE C. Multiple comparisons? No problem! (editorial). *Epidemiology* 1991, 2:241–242
63. SAVITZ DA, OLSHAN AF. Multiple comparisons and related issues in epidemiologic research. *Am J Epidemiol* 1995, 142:904–908

Corresponding author:

P. Galanis, 14 Dikis str., GR-157 73 Athens, Greece
e-mail: pegalan@nurs.uoa.gr