ARCHIVES OF HELLENIC MEDICINE: ISSN 11-05-3992

ΕΦΑΡΜΟΣΜΕΝΗ ΙΑΤΡΙΚΗ ΕΡΕΥΝΑ
APPLIED MEDICAL RESEARCH

Στατιστικές «παγίδες» στη βιοϊατρική έρευνα Η αξία της τιμής p

ΑΡΧΕΙΑ ΕΛΛΗΝΙΚΗΣ ΙΑΤΡΙΚΗΣ 2010, 27(1):113–118
ARCHIVES OF HELLENIC MEDICINE 2010, 27(1):113–118

Δ.Β. Παναγιωτάκος,
Α. Χαϊμάνη,
Μ. Σιταρά

Τμήμα Επιστήμης Διαιτολογίας-Διατροφής, Χαροκόπειο Πανεπιστήμιο, Αθήνα

Statistical "errors" in biomedical research: The value of the p-value

Abstract at the end of the article

Λέξεις ευρετηρίου

Ανάλυση δεδομένων
Διακύμανση
Ιατρική
Μέσος όρος
p-value
Στατιστική

Υποβλήθηκε 19.12.2008
Εγκρίθηκε 12.1.2009

1. ΕΙΣΑΓΩΓΗ

Τα περιγραφικά μέτρα, οι έλεγχοι σημαντικότητας και τα αντίστοιχα σφάλματα p (ή p-value, όπως συνήθως ονομάζονται ακόμα και στην ελληνική ορολογία) έχουν ιδιαιτέρα σημαντικό ρόλο στη λήψη αποφάσεων στην ιατρο-βιολογική έρευνα.[1–4] Στο άρθρο αυτό γίνεται μια προσπάθεια σφαιρικής παρουσίασης όχι μόνο της σημασίας και της ερμηνείας, αλλά και της παρερμηνείας των βασικών περιγραφικών μέτρων, καθώς και του p, το οποίο στην πλειοψηφία των ερευνών αποτελεί το σημαντικότερο κριτήριο στη λήψη αποφάσεων. Επίσης, εξετάζονται εναλλακτικές μέθοδοι και μέτρα για την αξιολόγηση των αποτελεσμάτων των ερευνών.

2. ΠΑΡΕΡΜΗΝΕΙΕΣ ΤΩΝ ΠΕΡΙΓΡΑΦΙΚΩΝ ΣΤΑΤΙΣΤΙΚΩΝ ΜΕΤΡΩΝ

Τα περιγραφικά μέτρα, όπως ο αριθμητικός μέσος και η τυπική απόκλιση, έχουν ιδιαίτερη θέση στην παρουσίαση των ευρημάτων μιας έρευνας. Θα πρέπει όμως να επισημανθεί ότι πολλές φορές ο αναγνώστης μιας επιστημονικής εργασίας λαμβάνει αμφιλεγόμενα μηνύματα από τα στατιστικά αποτελέσματα που παρουσιάζονται.

Ο αριθμητικός μέσος ή μέσος όρος ή αναμενόμενη τιμή είναι το άθροισμα όλων των μετρήσεων ή παρατηρήσεων διαιρεμένο με το πλήθος τους, δηλαδή $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$. Πολλές φορές, η μέση τιμή μπορεί να δώσει μια αρκετά αντιπροσωπευτική εικόνα του δείγματος, συχνά όμως είναι παραπλανητική. Αυτό συμβαίνει γιατί επηρεάζεται από τις ακραίες τιμές, με αποτέλεσμα, όταν το δείγμα είναι ανομοιογενές, η αναμενόμενη τιμή να είναι πλασματική αφού δεν αντιπροσωπεύει την πλειοψηφία των παρατηρήσεων. Για παράδειγμα, μια μεταβλητή με παρατηρήσεις 2, 3, 4, 18, 25, 20 έχει μια μέση τιμή. Δηλαδή, μια τυχαία παρατήρηση αυτής της μεταβλητής αναμένεται να έχει τιμή $\bar{x}=\frac{2+3+4+18+25+20}{6}=\frac{72}{6}=12$. Παρ' όλα αυτά, στο συγκεκριμένο δείγμα καμιά παρατήρηση δεν έχει τιμή 12, επομένως το 12 δεν εκφράζει την αναμενόμενη τιμή, όπως θα έπρεπε. Σε τέτοιες περιπτώσεις χρησιμοποιούνται και άλλα μέτρα θέσης, όπως είναι η διάμεσος και η επικρατούσα τιμή. Διάμεσος είναι η τιμή, η οποία είναι μικρότερη από τις μισές παρατηρήσεις και μεγαλύτερη από τις άλλες μισές και ορίζεται ως τιμή της παρατήρησης με θέση $\frac{n+1}{2}$, μετά από διάταξη κατά αύξουσα ή φθίνουσα σειρά. Επικρατούσα τιμή είναι αυτή με τη μεγαλύτερη συχνότητα

εμφάνισης στο δείγμα. Χρησιμοποιώντας όμως τα μέτρα θέσης δεν μπορεί να σχηματιστεί πάντα σαφής εικόνα για τα δεδομένα. Για το λόγο αυτόν, κρίνεται σκόπιμος ο υπολογισμός των λεγόμενων μέτρων διασποράς, όπως το εύρος, η διασπορά ή η διακύμανση, η τυπική απόκλιση και τα εκατοστημόρια. Το εύρος ορίζεται ως η διαφορά των τιμών της ελάχιστης από τη μέγιστη παρατήρηση. Είναι το απλούστερο αλλά και το λιγότερο πληροφοριακό μέτρο διασποράς, καθώς βασίζεται μόνο στις δύο ακραίες τιμές του δείγματος. Τα εκατοστημόρια έχουν αντίστοιχη έννοια με τη διάμεσο και έτσι το Κ εκατοστημόριο είναι η τιμή της παρατήρησης με σειρά $\frac{(n+1)\times K}{100}$ και δηλώνει ότι το Κ% των παρατηρήσεων είναι μικρότερο από αυτή την τιμή, ενώ το (Κ-1)% μεγαλύτερο. Η διακύμανση ορίζεται ως η μέση τιμή των τετραγώνων των αποκλίσεων των παρατηρήσεων από τη μέση τιμή του δείγματος, δηλαδή $s^2=\frac{\sum_{i=1}^{n}\left(x-\bar{x}\right)^2}{n-1}$. Βασικό μειονέκτημα αυτού του μέτρου είναι ότι εκφράζεται στα τετράγωνα των μονάδων μέτρησης της μεταβλητής, γεγονός που δεν διευκολύνει τυχόν υπολογισμούς. Για το λόγο αυτό χρησιμοποιείται ως μέτρο διασποράς η τυπική ή η σταθερή απόκλιση, που ορίζεται ως η τετραγωνική ρίζα της διακύμανσης, δηλαδή $s=\sqrt{s^2}=\sqrt{\frac{\sum_{i=1}^{n}\left(x-\bar{x}\right)^2}{n-1}}$ και εκφράζεται στις μονάδες του παρατηρούμενου μεγέθους.

Ένα άλλο, αρκετά αξιόπιστο μέτρο, διασποράς είναι ο συντελεστής μεταβλητότητας, που ορίζεται ως ο λόγος της τυπικής απόκλισης προς τη μέση τιμή, δηλαδή $CV=\frac{s}{\bar{x}}$, και εκφράζει τη μεταβλητότητα που υπάρχει στις παρατηρήσεις ενός μεγέθους. Ο συντελεστής μεταβλητότητας είναι καθαρός αριθμός, καθώς η μέση τιμή και η τυπική απόκλιση έχουν τις ίδιες μονάδες μέτρησης, ενώ συχνά πολλαπλασιαζόμενος επί 100 εκφράζεται και ως ποσοστό. Κατά συνέπεια, μπορεί να χρησιμοποιηθεί για συγκρίσεις των διασπορών μεταξύ διαφορετικών μεγεθών. Πολλές φορές όμως χρησιμοποιείται και σε περιπτώσεις σύγκρισης διασποράς συνόλων τιμών, οι οποίες μπορεί να έχουν μετρηθεί στις ίδιες μονάδες μέτρησης αλλά οι μέσες τιμές τους απέχουν πολύ μεταξύ τους. Ένα δείγμα χαρακτηρίζεται ως ομοιογενές αν $CV<0,1$ (ή 10%). Επισημαίνεται ότι ο συντελεστής μεταβλητότητας δεν μπορεί να υπολογιστεί όταν $\bar{x}=0$, ενώ σε περιπτώσεις όπου η μέση τιμή είναι αρνητική, η τελευταία μπορεί να αντικατασταθεί από την απόλυτη τιμή της. Σε ένα γενικότερο πλαίσιο, ο συντελεστής μεταβλητότητας είναι περισσότερο πληροφοριακός για μεταβλητές που έχουν πάντα θετικές ή τουλάχιστον ομόσημες τιμές και πολλές φορές είναι χρήσιμος για τη σύγκριση αποτελεσμάτων διαφορετικών ερευνών που εξετάζουν τη συμπεριφορά του ίδιου χαρακτηριστικού στον ίδιο πληθυσμό ή σε διαφορετικούς πληθυσμούς.

3. ΛΗΨΗ ΑΠΟΦΑΣΕΩΝ ΣΤΗ ΒΙΟΪΑΤΡΙΚΗ ΕΡΕΥΝΑ

Η σύγχρονη Ιατρική των ενδείξεων αποσκοπεί στην εφαρμογή των πληροφοριών που προκύπτουν από τις επιστημονικές έρευνες σε διάφορους τομείς της ιατρικής πρακτικής. Συγκεκριμένα, επιδιώκει να εκτιμήσει την ποιότητα των διαφόρων στοιχείων που σχετίζονται τόσο με τους κινδύνους όσο και με τα οφέλη που προκύπτουν από τα ατομικά χαρακτηριστικά ή τις διάφορες θεραπείες.[1–4] Σύμφωνα με το Κέντρο της Αποδεικτικής Ιατρικής, *«η Αποδεικτική Ιατρική αποτελεί την ευσυνείδητη, σαφή και συνετή χρήση των καλύτερων στοιχείων στη λήψη ορθών αποφάσεων όσον αφορά στη φροντίδα του κάθε ασθενούς»*.[4,5] Ακρογωνιαίος λίθος στην ιατρική έρευνα είναι η ποιότητα αυτών των αποφάσεων. Σύμφωνα με την Ιατρική των ενδείξεων, η έρευνα κατηγοριοποιείται και κατατάσσεται ανάλογα με την ισχύ της έλλειψης διαφόρων τύπων σφαλμάτων. Τα ισχυρότερα στοιχεία για θεραπευτικές επεμβάσεις προέρχονται από τη μετα-ανάλυση των τυχαιοποιημένων, διπλών-τυφλών, ελεγχόμενων κλινικών δοκιμών. Αντίθετα, οι υποθετικές αναφορές και οι απόψεις των ειδικών δεν έχουν ιδιαίτερη αξία. Η Αμερικανική Εταιρεία Ιατρικής των Ενδείξεων[4] κατατάσσει τις επιστημονικές ενδείξεις με την ακόλουθη σειρά: (α) Στοιχεία που αποκτήθηκαν από περισσότερες από μία τυχαιοποιημένες ελεγχόμενες δοκιμές ή μετα-αναλύσεις (επίπεδο Ι), (β) στοιχεία που αποκτήθηκαν από ελεγχόμενες κλινικές δοκιμές χωρίς τυχαιοποίηση (επίπεδο ΙΙ-1) ή στοιχεία που αποκτήθηκαν από προοπτικές επιδημιολογικές μελέτες ή μελέτες ασθενών-μαρτύρων (επίπεδο ΙΙ-2) ή στοιχεία που αποκτήθηκαν από πολλαπλές σειρές δοκιμών με ή χωρίς επέμβαση (επίπεδο ΙΙ-3) και (γ) απόψεις από καταξιωμένους επιστήμονες, οι οποίες βασίζονται στην κλινική τους εμπειρία, τις περιγραφικές μελέτες ή σε αναφορές ειδικών επιτροπών (επίπεδο ΙΙΙ). Η Βρετανική Εθνική Υπηρεσία Υγείας χρησιμοποιεί ένα παρόμοιο σύστημα κατηγοριοποίησης.[4] Κάθε φορά που πρέπει να επιλεχθεί η καταλληλότερη μέθοδος ανάμεσα σε πολλές εναλλακτικές, ο ερευνητής αναλαμβάνει το ρόλο να βοηθήσει σε αυτή την επιλογή βασιζόμενος στο επίπεδο των ενδείξεων. Ειδικότερα, όταν οι αποφάσεις είναι περίπλοκες και απαιτούν προσεκτική μελέτη και συστηματική αναθεώρηση των διαθέσιμων πληροφοριών, η συμβολή του ερευνητή είναι καθοριστική.

Η Ιατρική των ενδείξεων επιδιώκει να εκφράσει τα αποτελέσματα της έρευνας, χρησιμοποιώντας αυστηρές μαθηματικές (στατιστικές) μεθόδους. Τα εργαλεία που

χρησιμοποιούνται από τους ερευνητές περιλαμβάνουν μεταξύ άλλων τους λόγους πιθανοφανειών, διάφορα –μονομεταβλητά ή πολυμεταβλητά– στατιστικά υποδείγματα, το εμβαδόν κάτω από την καμπύλη λειτουργικών χαρακτηριστικών, γνωστή και ως καμπύλη ROC (receiver operator characteristic curve), τη γραφική παράσταση ευαισθησίας-ειδικότητας, και πολλά άλλα. Το p είναι ένας από τους στατιστικούς όρους με την πιο ευρεία χρήση στη λήψη αποφάσεων στις βιοϊατρικές έρευνες και βοηθά τους ερευνητές στην εξαγωγή συμπερασμάτων σχετικά με τη στατιστική σημαντικότητα της έρευνας. Μέχρι σήμερα, οι περισσότεροι ερευνητές βασίζουν τις αποφάσεις τους στην τιμή της πιθανότητας p. Πολλές φορές, όμως, ο όρος p παρερμηνεύεται ή και άλλες φορές υπερεκτιμάται, γεγονός που οδηγεί σε σοβαρά μεθοδολογικά λάθη.[6,7] Σε αυτό το άρθρο, παρουσιάζεται η ερμηνεία του p και κάποιες εναλλακτικές διαθέσιμες επιλογές, οι οποίες κρίνονται καταλληλότερες για την εξαγωγή ορισμένων συμπερασμάτων.

Στη στατιστική επιστήμη, με p ορίζεται η πιθανότητα να προκύψει αποτέλεσμα τουλάχιστον τόσο ακραίο όσο εκείνο που παρατηρήθηκε στο βιολογικό ή στο κλινικό πείραμα ή στην επιδημιολογική έρευνα, με την προϋπόθεση ότι η μηδενική υπόθεση δεν μπορεί να απορριφθεί[1–3,6,7] ή, αλλιώς, το παρατηρούμενο επίπεδο της στατιστικής σημαντικότητας. Η πιθανότητα p συνοδεύει κάθε έλεγχο υποθέσεων και προκύπτει με βάση το στατιστικό κριτήριο που χρησιμοποιείται. Οι έλεγχοι υποθέσεων αποτελούν θεμελιώδη διαδικασία στην επαγωγική στατιστική και θα μπορούσαν να θεωρηθούν ως «μέθοδος» για τη λήψη στατιστικών αποφάσεων χρησιμοποιώντας πειραματικά δεδομένα. Σε αυτό το σημείο θα ήταν σκόπιμο να εισαχθούν κάποιοι όροι που σχετίζονται με τους ελέγχους υποθέσεων. Αρχικά, υπάρχουν πάντα δύο υποθέσεις, η μηδενική (συμβολικά H_0) και η εναλλακτική (συμβολικά H_A). Συνήθως, η μηδενική υπόθεση δηλώνει την έλλειψη συσχέτισης μεταξύ των παραγόντων ή των χαρακτηριστικών που διερευνώνται (τα οποία μετρώνται με τη χρήση τυχαίων μεταβλητών). Παραδείγματα μηδενικών υποθέσεων είναι τα εξής: «ο επιπολασμός των καρδιαγγειακών νοσημάτων έχει την ίδια τιμή μεταξύ των ανδρών και των γυναικών», δηλαδή «δεν υπάρχει συσχέτιση μεταξύ του φύλου και της νόσου». Από την άλλη πλευρά, η εναλλακτική υπόθεση δηλώνει συσχέτιση ανάμεσα στις μεταβλητές που εξετάζονται στο ίδιο παράδειγμα, «ο επιπολασμός των καρδιαγγειακών νοσημάτων διαφέρει μεταξύ των δύο φύλων» (αμφίπλευρος έλεγχος) ή «ο επιπολασμός στους άνδρες είναι μεγαλύτερος από τον επιπολασμό στις γυναίκες ή ο επιπολασμός στις γυναίκες είναι μεγαλύτερος από τον επιπολασμό στους άνδρες» (μονόπλευρος έλεγχος). Το 1950, ο Fisher[8] πρότεινε τους ελέγχους στατιστικής σημαντικότητας ως μέσο εξέτασης της ασυμφωνίας μεταξύ των δεδομένων και της μηδενικής υπόθεσης. Ορισμένοι από τους πιο συχνούς χρησιμοποιούμενους ελέγχους στη βιοϊατρική έρευνα είναι το Z-test, το Student's t-test, το F-test και το x^2.

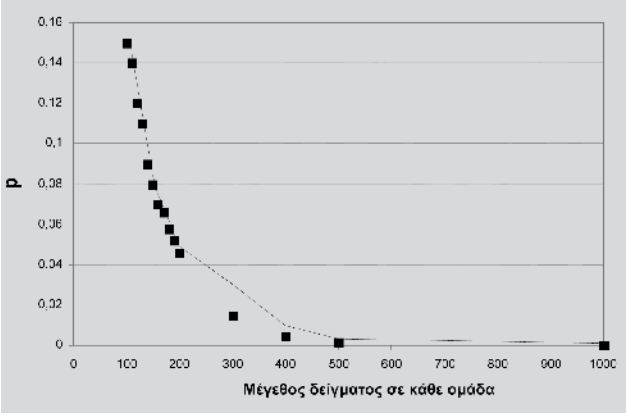
Με βάση την αυστηρή μαθηματική στατιστική, το p είναι μια πιθανότητα η οποία προσδιορίζεται από το χώρο του δείγματος (δηλαδή το σύνολο όλων των πιθανών αποτελεσμάτων του πειράματος) και έτσι η κατανομή του κάτω από τη μηδενική υπόθεση είναι ομοιόμορφη στο διάστημα [0,1]. Για παράδειγμα, μια κλινική δοκιμή φάσης III εκτελείται για να καθορίσει αν οι τιμές της ολικής χοληστερόλης διαφέρουν μεταξύ των ατόμων της ομάδας που υποβλήθηκαν σε θεραπεία με το φάρμακο Α συγκρινόμενα με τα άτομα της ομάδας που υποβλήθηκαν σε θεραπεία με χρήση του φαρμάκου Β. Χάρη ευκολίας υποθέτουμε ότι οι αρχικές τιμές της χοληστερόλης και στις δύο ομάδες ήταν ίσες. Μετά από 12 μήνες θεραπείας παρατηρήθηκε σημαντική μείωση της τιμής της χοληστερόλης στην ομάδα Α κατά 27±10 mg/dL και στην ομάδα Β μείωση κατά 25±10 mg/dL. Αν 100 ασθενείς ήταν κατανεμημένοι σε κάθε ομάδα και λαμβάνοντας υπ' όψιν τις προϋποθέσεις για την κατάλληλη δοκιμασία σημαντικότητας, το p του ελέγχου αυτής της υπόθεσης είναι ίσο με 0,15. Σε αυτή την περίπτωση η μηδενική υπόθεση δεν απορρίπτεται, έναντι της εναλλακτικής, που ήταν ότι «στον πληθυσμό οι ολικές μειώσεις δεν ήταν ίσες». Το p που παρουσιάσαμε πιο πάνω αφορά στην πιθανότητα να παρατηρηθεί διαφορά 2 mg/dL ή και ακόμη μεγαλύτερη, μεταξύ των δύο ομάδων θεραπείας, υπό την υπόθεση ότι υπάρχει σχεδόν η ίδια μείωση στην τιμή της χοληστερόλης και στις δύο πειραματικές ομάδες (δηλαδή η μηδενική υπόθεση). Το p της τάξης του 0,15 σημαίνει ότι η παρατηρούμενη διαφορά μπορεί να αποδοθεί στην τύχη κατά 15%. Σύμφωνα με την προσέγγιση του Fisher, η μηδενική υπόθεση ποτέ δεν επαληθεύεται, αλλά είναι πιθανόν να διαψευστεί. Επιπλέον, ο Fisher προτείνει ως όριο της στατιστικής σημαντικότητας (δηλαδή το α) το 0,05. Αν το p είναι μικρότερο από α, τότε υπάρχουν αρκετές ενδείξεις που οδηγούν στην απόρριψη της μηδενικής υπόθεσης.[8]

Παρά την αξιοσημείωτη κριτική που δέχθηκε και εξακολουθεί να δέχεται αυτή η προσέγγιση, όλοι συμφωνούν ότι το επίπεδο σημαντικότητας πρέπει να αποφασίζεται πριν αναλυθούν τα δεδομένα και να συγκρίνεται με το p, αφού γίνει ο έλεγχος. Επιπλέον, παρά το γεγονός ότι τα p χρησιμοποιούνται ευρέως, υπάρχουν πολλές παρερμηνείες. Στο παρακάτω κείμενο γίνεται μια προσπάθεια να διευκρινιστεί τι πραγματικά είναι και τι όχι το p.

3.1. Τι είναι το p και τι δεν είναι…

Το p δεν αποτελεί την πιθανότητα να επαληθευτεί η μηδενική υπόθεση και αυτό γιατί οι υποθέσεις δεν έχουν πιθανότητες στην κλασική στατιστική. Επιπρόσθετα, το p δεν είναι η πιθανότητα να απορριφθεί λανθασμένα η μηδενική υπόθεση. Η εσφαλμένη απόφαση κατά την οποία απορρίπτεται η μηδενική υπόθεση ενώ είναι αληθής αποτελεί το σφάλμα τύπου I. Αυτό το σφάλμα είναι μια εκδοχή της καλούμενης «πλάνης του εισαγγελέα» ("prosecutor's fallacy"). Το ποσοστό του σφάλματος τύπου I είναι στενά συνυφασμένο με το p, αφού απορρίπτουμε τη μηδενική υπόθεση όταν το p είναι μικρότερο από κάποιο προκαθορισμένο όριο *α*. Το p δεν δηλώνει το μέγεθος ή τη σημασία του παρατηρούμενου αποτελέσματος. Έτσι, ένα πολύ μικρό p, π.χ. 0,000… (συνήθως παρουσιάζεται ως <0,001), δεν σημαίνει απαραίτητα μια πολύ ισχυρή συσχέτιση (συγκρινόμενο με το μέγεθος του αποτελέσματος που αποτελεί μέτρο της σχέσης μεταξύ δύο μεταβλητών, όπως ο σχετικός λόγος συμπληρωματικών πιθανοτήτων, ο σχετικός κίνδυνος, ο συντελεστής συσχέτισης, το d του Cohen κ.λπ.[5,6]). Επιπρόσθετα, το p επηρεάζεται από το μέγεθος του δείγματος. Για παράδειγμα, στην εικόνα 1 απεικονίζεται η εντυπωσιακή μείωση του p σε συνάρτηση με το μέγεθος του δείγματος, διατηρώντας τα παρατηρούμενα μεγέθη σταθερά. Είναι προφανές ότι αν το αρχικό δείγμα διπλασιαστεί (δηλαδή n=200 για κάθε είδος θεραπείας), τα αποτελέσματα της έρευνας αποκτούν στατιστική σημαντικότητα.

Ένα θέμα που επηρεάζει τη λήψη των ιατρικών αποφάσεων είναι οι πολλαπλές συγκρίσεις που συμβαίνουν, όταν μια οικογένεια στατιστικών συμπερασμάτων μελετάται ταυτόχρονα. Για παράδειγμα, κάνοντας μόνο έναν έλεγχο υπόθεσης σε επίπεδο στατιστικής σημαντικότητας 5%, υπάρχει μόνο 5% πιθανότητα να προκύψει αποτέλεσμα



Εικόνα 1. Θεωρητικό παράδειγμα του p σε σχέση με το μέγεθος του δείγματος για τις ίδιες παρατηρήσεις.

τουλάχιστον τόσο ακραίο όσο αυτό που παρατηρείται όταν επαληθεύεται η μηδενική υπόθεση. Κάνοντας 100 ελέγχους, όμως, με όλες τις μηδενικές υποθέσεις να επαληθεύονται, είναι πιθανότερο ότι θα απορριφθεί τουλάχιστον μία μηδενική υπόθεση. Αυτά τα σφάλματα ονομάζονται «θετικά σφάλματα» και για τον περιορισμό τους έχουν αναπτυχθεί πολλές μαθηματικές τεχνικές. Οι περισσότερες από αυτές τροποποιούν το επίπεδο σημαντικότητας *α*, προκειμένου να υπολογιστεί η επίδραση του ρυθμού του σφάλματος τύπου I και να καταστεί η σύγκριση του p πιο ακριβής.

Για όλους τους παραπάνω λόγους, πολλά επιστημονικά περιοδικά προτείνουν στους συγγραφείς να παρουσιάζουν διαστήματα εμπιστοσύνης αντί για το p, καθώς και μέτρα αποτίμησης του μεγέθους της σχέσης.[9–11] Το (1-α)% διάστημα εμπιστοσύνης είναι το διάστημα εκείνο, το οποίο περιέχει την τιμή του υπολογιζόμενου δειγματικού μέτρου (π.χ. μέση τιμή, διαφορά δύο μέσων τιμών) ή στατιστικού κριτηρίου (π.χ. t-test, F-test) στον άγνωστο πληθυσμό αναφοράς με βεβαιότητα, δηλαδή πιθανότητα (1-α)%. Ο υπολογισμός των άκρων του διαστήματος, των λεγόμενων ορίων αξιοπιστίας, γίνεται με προσθαφαίρεση στο υπολογιζόμενο μέτρο του δειγματοληπτικού σφάλματος πολλαπλασιασμένο επί μια σταθερά της κατανομής του μέτρου (π.χ. z, t), και η οποία αλλάζει ανάλογα με το επίπεδο σημαντικότητας. Για παράδειγμα, αν σε ένα δείγμα έχει υπολογιστεί η μέση τιμή $\bar{x}$ ενός χαρακτηριστικού, τότε το 95% διάστημα εμπιστοσύνης της $\bar{x}$ θα είναι ($\bar{x}$–1,96×SE,$\bar{x}$+1,96×SE), όπου 1,96 η τιμή του κριτηρίου z για α=5% και $SE=\frac{s}{\sqrt{n}}$ το δειγματοληπτικό σφάλμα. Το διάστημα εμπιστοσύνης χρησιμοποιείται για την καλύτερη αξιολόγηση της εγκυρότητας ενός στατιστικού μέτρου ή κριτηρίου στον πληθυσμό αναφοράς. Από πολλούς πλέον ερευνητές έχει υιοθετηθεί η χρήση των διαστημάτων εμπιστοσύνης ως συμπληρωματικών της στατιστικής σημαντικότητας p των ελέγχων, ενώ από ορισμένους άλλους τα διαστήματα εμπιστοσύνης έχουν αντικαταστήσει τη στατιστική σημαντικότητα, αφήνοντας στον αναγνώστη τη δυνατότητα να αποφασίσει για το μέγεθος της σημασίας των ευρημάτων.

Τέλος, το p δεν είναι η πιθανότητα ότι το πείραμα δεν θα αποφέρει το ίδιο αποτέλεσμα μετά από κ-επαναλήψεις. Γι' αυτόν το λόγο, ο Killeen[12] πρότεινε το p_{rep} ως εναλλακτικό μέτρο αντί του p, το οποίο υπολογίζει την πιθανότητα επανάληψης ενός αποτελέσματος. Μια προσέγγιση του p_{rep} είναι η ακόλουθη:

$$p_{rep} = \left[1+\left(\frac{p}{1-p}\right)^{\frac{2}{3}}\right]^{-1}$$

Όσο μικρότερο είναι το p τόσο μεγαλύτερο είναι το p_{rep}. Η Association for Psychological Science (APS) προτεί-

νει στους συγγραφείς των επιστημονικών περιοδικών να παρουσιάζουν το p_{rep} αντί του p. Παρόλα αυτά όμως, έχει γίνει αξιοσημείωτη κριτική και γι' αυτή την επιλογή. Για παράδειγμα, ένα μειονέκτημα του p_{rep} είναι ότι δεν λαμβάνει υπ' όψιν τις εκ των προτέρων πιθανότητες[13] και δεν παρέχει πρόσθετες πληροφορίες για τη σημαντικότητα του αποτελέσματος ενός δοθέντος πειράματος.

Πρόσφατα, προτάθηκε να εφαρμοστούν περισσότερο «λεπτομερείς» στατιστικές μέθοδοι, για να ερμηνευτούν «σημαντικές» σχέσεις που λαμβάνουν υπ' όψιν τις εκ των προτέρων πιθανότητες για την απόρριψη μιας υπόθεσης, όπως ο παράγοντας Β του Bayes.[14,15] Μια σημαντική παράμετρος είναι ότι ο παράγοντας Β του Bayes απαιτεί προϋπάρχουσες γνώσεις για να μετατραπεί σε συμπέρασμα. Η απλούστερη μορφή του παράγοντα του Bayes είναι ο λόγος πιθανοφανειών (δηλαδή ο λόγος Λ της μέγιστης πιθανότητας ενός αποτελέσματος κάτω από δύο διαφορετικές υποθέσεις, τη μηδενική και την εναλλακτική). Το ελάχιστο του παράγοντα του Bayes είναι αντικειμενικό και μπορεί να χρησιμοποιηθεί αντί του p ως μέτρο της αποδεικτικής ισχύος. Ωστόσο, οι ερευνητές δεν εμφανίζονται πολύ ενθουσιασμένοι με την ιδέα να καταλάβουν και να υιοθετήσουν τις στατιστικές μεθοδολογίες του Bayes, αντιλαμβανόμενοι μια υποκειμενική προσέγγιση στην ανάλυση των στοιχείων. Παρά την κριτική, για πολλούς επιστήμονες η χρήση του παράγοντα Β του Bayes είναι μια εναλλακτική μέθοδος αντί των κλασικών ελέγχων υποθέσεων που αναφέρθηκαν παραπάνω. Ειδικότερα, ο Ioannidis υπολόγισε τον παράγοντα Β του Bayes σε 272 μελέτες παρατήρησης και 50 μετα-αναλύσεις αναφορικά με γενετικά εξαρτώμενες νόσους για τις οποίες οι στατιστικά σημαντικές σχέσεις απαιτήθηκαν με όριο $p < 0,005$. Με βάση τον παράγοντα Β, τα ήδη παρατηρηθέντα στατιστικά σημαντικά αποτελέσματα δεν είχαν ισχυρή υποστήριξη για το 54–77% των επιδημιολογικών ερευνών και για το 44–70% των 50 μετα-αναλύσεων.[14]

Επιγραμματικά, αντίθετα με το p, ο παράγοντας του Bayes έχει τέτοια ερμηνεία που επιτρέπει τη χρήση του τόσο στη συμπερασματολογία όσο και στη λήψη αποφάσεων, καθώς αποσαφηνίζει τη διάκριση μεταξύ της πειραματικής απόδειξης και των τελικών συμπερασμάτων, ενώ παρέχει ένα πλαίσιο στο οποίο συνδέονται οι παλαιότερες με τις πρόσφατες ενδείξεις.

4. ΕΠΙΛΟΓΟΣ

Σε αυτό το άρθρο έγινε μια προσπάθεια ερμηνείας της σημασίας του p, μια πολύ βασική πιθανότητα στις περισσότερες βιοϊατρικές έρευνες για τη λήψη αποφάσεων. Οι πρόσφατες κατευθυντήριες γραμμές για την εξαγωγή αποτελεσμάτων των κλινικών πειραμάτων ή των μελετών παρατηρήσεων προτείνουν να παρουσιάζονται τα διαστήματα εμπιστοσύνης αντί ή μαζί με τα p, και να παρέχονται τα μεγέθη των επιδράσεων των σχέσεων που διερευνώνται.

Συμπερασματικά, θα μπορούσε να λεχθεί ότι, παρά τα μειονεκτήματά του, το p εξακολουθεί να έχει σημαντική αξία όταν χρησιμοποιείται και ερμηνεύεται σωστά.

ABSTRACT

Statistical "errors" in biomedical research: The value of the p-value

D.B. PANAGIOTAKOS, A. CHAIMANI, M. SITARA

Office of Biostatistics and Epidemiology, Department of Nutrition Science-Dietetics, Harokopeio University of Athens, Kallithea, Greece

Archives of Hellenic Medicine 2010, 27(1):113–118

Statistical analysis of biomedical data has played an important role in the development of health sciences during recent years. The role of statistical science in medical decision making is multidimensional. Calculation of sample size, and assessment of measurement error or decision errors are some of the issues that arise when statistical science is used in the support of evidence based medicine. In this article, common misinterpretations observed in medical articles regarding the arithmetic mean and the correct interpretation of the p-value are discussed.

Key words: Data analysis, Mean, Medicine, p-value, Statistics, Variance

Βιβλιογραφία

1. ΤΡΙΧΟΠΟΥΛΟΣ Δ, ΤΖΩΝΟΥ Α, ΚΑΤΣΟΥΓΙΑΝΝΗ Κ. *Βιοστατιστική*. Εκδόσεις Παρισιάνος, Αθήνα, 2000
2. ELSTEIN AS. On the origins and development of evidence-based medicine and medical decision making. *Inflamm Res* 2004, 53:S184–S189
3. ΣΠΑΡΟΣ Λ, ΓΑΛΑΝΗΣ Π. *Δοκίμια επιδημιολογίας*. Εκδόσεις

Παρισιάνος, Αθήνα, 2006
4. ΣΤΑΥΡΙΝΟΣ Β, ΠΑΝΑΓΙΩΤΑΚΟΣ ΔΒ. *Βιοστατιστική*. Εκδόσεις Gutenberg, Αθήνα, 2007
5. SACKETT DL, STRAUS SE, RICHARDSON S, ROSENBERG W, HAYNES B. *Επί ενδείξεων βασιζόμενη Ιατρική* (ελληνική μετάφραση). Εκδόσεις Πασχαλίδης, Αθήνα, 2002
6. GOODMAN SN. Toward evidence-based medical statistics. 1: The p fallacy. *Ann Intern Med* 1999, 130:995–1004
7. SCHERVISH MJ. Ps: What they are and what they are not. *Am Stat* 1996, 50:203–206
8. FISHER RA. *Statistical methods for research workers.* Oliver & Boyd Publ, London, UK, 1950
9. NAKAGAWA S, CUTHILL IC. Effect size, confidence interval and statistical significance: A practical guide for biologists. *Biol Rev Camb Philos Soc* 2007, 82:591–605
10. ROTHMAN KJ. Random error and the role of statistics. In: *Epidemiology: An introduction.* Oxford University Press, New York, 2002:113–129
11. INTERNATIONAL COMMITTEE OF MEDICAL JOURNAL EDITORS. Uniform requirements for manuscripts submitted to biomedical journals. *Ann Intern Med* 1988, 108:258–265
12. KILLEEN PR. An alternative to null-hypothesis significance tests. *Psychol Sci* 2005, 16:345–353
13. McDONALD RR. Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen. *Psychol Sci* 2005, 16:1006–1008
14. IOANNIDIS JP. Effect of formal statistical significance on the credibility of observational associations. *Am J Epidemiol* 2008, 168:374–383
15. ΓΑΛΑΝΗΣ Π, ΣΠΑΡΟΣ Λ. Ανάλυση δεδομένων: Μη μπαγιεσιανή προσέγγιση. *Αρχ Ελλ Ιατρ* 2005, 22:377–391

Corresponding author:

D. Panagiotakos, Harokopeio University of Athens, 70 E. Venizelou Ave., GR-176 71 Kallithea, Greece
e-mail: dbpanag@hua.gr