 ΨΥΧΟΛΟΓΙΑ, 2011, 18 (2) ◆ 224-239

# Ο έλεγχος μηδενικών υποθέσεων: διαδικασία, παρανοήσεις και μερικές προτάσεις για καλύτερες πρακτικές

ΠΕΤΡΟΣ ΡΟΥΣΣΟΣ[1]

*ΠΕΡΙΛΗΨΗ* Το άρθρο ξεκινάει με μια παρουσίαση της διαδικασίας του ελέγχου μηδενικών υποθέσεων (Null Hypothesis Significance Testing), των αδυναμιών της και της κριτικής που έχει διατυπωθεί τις τελευταίες δεκαετίες στη διεθνή βιβλιογραφία. Στη συνέχεια παρουσιάζονται περιγραφικά και συζητώνται κριτικά οι τρόποι με τους οποίους οι συγγραφείς των ερευνητικών εργασιών που έχουν δημοσιευτεί στο περιοδικό ΨΥΧΟΛΟΓΙΑ αναφέρονται στη διαδικασία του ελέγχου υποθέσεων και ερμηνεύουν τα αποτελέσματα που προκύπτουν από αυτήν. Συγκεκριμένα, εξετάστηκαν 445 άρθρα που δημοσιεύτηκαν στο διάστημα από το 1992 ως το 2010. Αφού προβάλλονται ορισμένες από τις απόψεις που έχουν διατυπωθεί σχετικά με την ανάγκη μίας τροποποίησης των πρακτικών που ακολουθούμε στη συζήτηση των ευρημάτων από τη στατιστική επεξεργασία των ερευνητικών μας δεδομένων, διατυπώνονται μερικές προτάσεις για τη βελτίωση των πρακτικών που ακολουθούμε για την παρουσίαση των στατιστικών μας ευρημάτων ώστε να είναι αυτά σαφή και να διευκολύνεται η κατανόησή τους από τους αναγνώστες.

*Λέξεις-κλειδιά*: Έλεγχος μηδενικής υπόθεσης, Στατιστική συλλογιστική.

## Εισαγωγή

Το παρόν άρθρο έχει τέσσερις κύριους στόχους: Καταρχάς, να παρουσιάσει εν συντομία τη διαδικασία και τις αδυναμίες του ελέγχου μηδενικών υποθέσεων (Null Hypothesis Significance Testing). Δεύτερον, να εκθέσει με τη βοήθεια περιγραφικών μεθόδων και να συζητήσει κριτικά τους τρόπους με τους οποίους οι Έλληνες συγγραφείς ερευνητικών εργασιών αναφέρονται στη διαδικασία του ελέγχου μηδενικών υποθέσεων και ερμηνεύουν τα αποτελέσματα που προκύπτουν από αυτήν. Για το σκοπό αυτόν μελετήθηκαν όλα τα άρθρα που δημοσιεύτηκαν στο περιοδικό *ΨΥΧΟΛΟΓΙΑ* κατά το χρονικό διάστημα από το 1992 ως το 2010. Τρίτον, να προβάλει μερικές μόνο –καθώς η σχετική διεθνής βιβλιογραφία είναι τεράστια– από τις απόψεις που έχουν διατυπωθεί σχετικά με την ανάγκη μιας τροποποίησης των πρακτικών που ακολουθούμε στη συζήτηση των ευρημάτων από τη στατιστική επεξεργασία των ερευνητικών μας δεδομένων, «ανοί-

1. Διεύθυνση: Πρόγραμμα Ψυχολογίας, Τμήμα ΦΠΨ, Φιλοσοφική Σχολή, Εθνικό & Καποδιστριακό Πανεπιστήμιο Αθηνών, Πανεπιστημιούπολη, Ιλίσια, 15784, Αθήνα Τηλ.: 210 7277385, e-mail: roussosp@psych.uoa.gr

γοντας» ταυτόχρονα και στην ελληνική βιβλιογραφία τη συζήτηση αυτή. Τέλος, να διατυπώσει ορισμένες *προτάσεις* (και όχι οδηγίες) για τη βελτίωση των πρακτικών που ακολουθούμε για την παρουσίαση των στατιστικών μας ευρημάτων στο περιοδικό ώστε να είναι αυτά σαφή και να διευκολύνεται η κατανόησή τους από τους αναγνώστες. Ένας απώτερος και όχι άμεσος στόχος τον οποίο φιλοδοξεί να υπηρετήσει το παρόν άρθρο είναι αυτός της διαφοροποίησης και προσθήκης νέων διδακτικών στόχων κατά την εκπαίδευση των νέων επιστημόνων στη στατιστική, οι οποίοι θα συμβάλουν στην ανάπτυξη των ικανοτήτων τους για στατιστική συλλογιστική.

Να σημειωθεί ότι τα στοιχεία που παρουσιάζονται στην παρούσα εργασία βασίζονται στην τελική μορφή των δημοσιευμένων άρθρων και όχι στα υποβληθέντα άρθρα στο περιοδικό. Καταβλήθηκε επίσης κάθε δυνατή προσπάθεια ώστε να μη γίνεται γνωστή η ταυτότητα του συγγραφέα ή των συγγραφέων των άρθρων, αποσπάσματα των οποίων επιλέχτηκε να παρουσιαστούν εδώ.

### Ο έλεγχος μηδενικών υποθέσεων

Η έννοια της στατιστικής σημαντικότητας αφορά στο αν ένα ερευνητικό εύρημα θα αποδοθεί σε τυχαίους παράγοντες ή στη συστηματική επίδραση της ανεξάρτητης μεταβλητής (για πρακτικούς λόγους μόνο εστιάζουμε στην εφαρμογή της διαδικασίας ελέγχου μηδενικών υποθέσεων στα ερευνητικά δεδομένα που έχουν προέλθει από ένα απλό πείραμα. Η διαδικασία, όπως περιγράφεται παρακάτω, είναι ελαφρώς διαφορετική όταν γίνεται έλεγχος συνάφειας ή έχουν εφαρμοστεί άλλα ερευνητικά σχέδια). Η απάντηση στο δίλημμα αυτό προκύπτει μέσα από τη διαδικασία του ελέγχου μηδενικών υποθέσεων, η οποία εν συντομία προβλέπει:

α) Τη διατύπωση μιας ερευνητικής υπόθεσης, σύμφωνα με την οποία η ανεξάρτητη μεταβλητή επιδρά στην εξαρτημένη. Για παράδειγμα, αν υποτεθεί ότι η μεταβλητή του φύλου επιδρά στον προσανατολισμό στον χώρο, τότε ο ερευνητής διατυπώνει την ερευνητική υπόθεση (ή «εναλλακτική» για να μιλάμε με όρους του ελέγχου μηδενικών υποθέσεων): $\mu_α \neq \mu_γ$ (στην περίπτωση αμφίπλευρου ελέγχου) ή $\mu_α > \mu_γ$ ή $\mu_α < \mu_γ$ (στην περίπτωση μονόπλευρου ελέγχου), όπου $\mu_α$ και $\mu_γ$ είναι οι μέσες τιμές των κατανομών του πληθυσμού των ανδρών και των γυναικών αντίστοιχα.

β) Την πραγματοποίηση του πειράματος με την επιλογή δύο (τυχαίων κατά προτίμηση) δειγμάτων, τα μέλη των οποίων υποβάλλονται σε δοκιμασίες και λαμβάνονται μετρήσεις (εξαρτημένη μεταβλητή).

γ) Τον υπολογισμό των μέσων τιμών $M_α$ και $M_γ$ των δύο δειγμάτων. Γενικά, για τις τιμές αυτές ισχύει ότι $M_α \neq M_γ$, επομένως ο ερευνητής καλείται να προσδιορίσει αν η διαφορά που παρατηρήθηκε οφείλεται μόνο σε τυχαία σφάλματα μέτρησης ή είναι αποτέλεσμα των αντίστοιχων διαφορών που υπάρχουν και στους πληθυσμούς από τους οποίους προέρχονται τα δείγματα. Με άλλα λόγια, ο ερευνητής πρέπει να δείξει αν το φαινόμενο που παρατηρήθηκε είναι «πραγματικό».

δ) Την επιλογή του κατάλληλου στατιστικού κριτηρίου και την εφαρμογή του στα ερευνητικά δεδομένα ώστε να υπολογιστεί η πιθανότητα (η τιμή p) να παρατηρηθεί μια διαφορά μεταξύ των μέσων τιμών τόσο μεγάλη όσο αυτή που παρατηρήθηκε αν ισχύει η υπόθεση ότι $\mu_α = \mu_γ$ (ή $\mu_α \leq \mu_γ$ ή $\mu_α \geq \mu_γ$, εφόσον έχουμε κατευθυντική μηδενική υπόθεση). Αυτή η υπόθεση, δηλαδή ότι οι μέσοι όροι των πληθυσμών από τους οποίους προέρχονται τα δείγματα είναι ίσοι, ονομάζεται «μηδενική».

ε) Τη σύγκριση της τιμής p με το επίπεδο στατιστικής σημαντικότητας (την τιμή α, η οποία συνήθως ορίζεται στο 0,05) προκειμένου να καταλήξει σε ένα από τα ακόλουθα δύο συμπεράσματα: είτε να απορρίψει τη μηδενική υπόθεση (εφόσον η τιμή $p < α$) είτε να αποτύχει να απορρίψει τη μηδενική υπόθεση (εφόσον η τιμή $p > α$).

Τα περισσότερα στοιχεία της λογικής του ελέγχου υποθέσεων παρουσιάστηκαν το 1925 από τον Ronald Fisher, Άγγλο στατιστικό και εξελικτικό βιολόγο, στο κλασικό βιβλίο του *Statistical*

*Methods for Research Workers*[1]. Τρία χρόνια αργότερα, η διαδικασία όπως τη γνωρίζουμε και την εφαρμόζουμε σήμερα ολοκληρώθηκε από τους Jerzy Neyman, Αμερικανο-πολωνό μαθηματικό και στατιστικό, και Egon Pearson, Βρετανό στατιστικό (1928). Οι ειδικότητες των τριών επιστημόνων αναφέρονται επί τούτω, καθώς η διαδικασία αυτή κάνει την εμφάνισή της στην ψυχολογία αρκετά αργότερα (τη δεκαετία του 1950) και αφού έχει προηγηθεί η *πιθανοκρατική επανάσταση στην επιστήμη* (Krüger, Daston & Heidelberger, 1987. Porter, 1986). Έκτοτε, κυριάρχησε στις επιστήμες της συμπεριφοράς και, μάλιστα, σύμφωνα με τον Gigerenzer, ο οποίος είναι ένας από τους σημαντικότερους επικριτές της, διδάσκεται και εφαρμόζεται τυπολατρικά.

Το εντυπωσιακό με τη διαδικασία αυτή είναι πως και οι τρεις «πατέρες» της θα την είχαν απορρίψει ως ανακόλουθη προς τις ιδέες τους (Gigerenger, Swijtink, Porter, Daston, Beatty, & Krüger, 1989), σχεδόν κανένας από τους εξέχοντες στατιστικούς της εποχής εκείνης δεν την αποδέχτηκε (Gigerenzer, 1998), και αρκετοί από τους επιφανέστερους ψυχολόγους της εποχής, όπως οι Skinner, Bartlett, Simon, Luce και Stevens, έγραψαν εναντίον της (Gigerenzer & Murray, 1987).

Οι διαφορές μεταξύ του Fisher και των Neyman και Pearson ήταν τόσες και τέτοιες που ο Gigerenzer αποκαλεί τη διαδικασία που προέκυψε «υβρίδιο» (1993). Συγκεκριμένα, ο Fisher υποστήριξε ότι το επίπεδο σημαντικότητας αντιμετωπίζεται ως σύμβαση: «Συνηθίζεται και είναι βολικό για τους πειραματιστές να παίρνουν το 0,05 ως ένα τυπικό επίπεδο σημαντικότητας, με την έννοια ότι είναι προετοιμασμένοι να αγνοούν όλα τα αποτελέσματα που αποτυγχάνουν να φτάσουν αυτό το κριτήριο» (1935, σελ. 33). Ωστόσο, ο Fisher άλλαξε γνώμη αργότερα (1956) υποστηρίζοντας ότι οι ερευνητές οφείλουν να δημοσιεύουν την ακριβή τιμή p (για παράδειγμα, p = 0,02 και όχι $p < 0,05$). Η άποψη αυτή, η οποία σημειωτέον δεν ενσωματώθηκε στη διαδικασία ελέγχου υποθέσεων, προέκυψε ως αντίθεση στις θέσεις των Neyman και Pearson, οι οποίοι πρότειναν ότι το επίπεδο σημαντικότητας πρέπει να ορίζεται από τους ερευνητές πριν το πείραμα. Για αυτούς, η έννοια του επιπέδου σημαντικότητας είναι η εξής: αν η μηδενική υπόθεση είναι αληθής και το πείραμα επαναληφθεί πολλές φορές, τότε ο ερευνητής θα απορρίψει εσφαλμένα τη μηδενική υπόθεση στις 5% από τις φορές που θα πραγματοποιήσει το πείραμα. Η απόρριψη της μηδενικής υπόθεσης ενώ είναι αληθής ονομάστηκε σφάλμα Τύπου Ι στη θεωρία τους και η πιθανότητά της alpha ($\alpha$). Έτσι, αν επιλέξουμε το επίπεδο σημαντικότητας 0,05, η πιθανότητα σφάλματος Τύπου Ι είναι 5%. Η βασική διαφορά είναι η εξής (Gigerenzer, 1993): για τον Fisher το επίπεδο σημαντικότητας είναι μια ιδιότητα των δεδομένων, ενώ για τους Neyman και Pearson το $\alpha$ είναι μια ιδιότητα του στατιστικού κριτηρίου (τεστ). Το επίπεδο σημαντικότητας και το α δεν είναι το ίδιο πράγμα.

Όσο χαμηλότερα τοποθετούμε το $\alpha$, ας πούμε στο 0,001, τόσο χαμηλότερη είναι η πιθανότητα να κάνουμε σφάλμα Τύπου Ι. Ωστόσο, όσο χαμηλότερη είναι η πιθανότητα να υποπέσουμε σε σφάλμα Τύπου Ι, τόσο χαμηλότερη είναι η ισχύς του στατιστικού κριτηρίου που χρησιμοποιούμε. Ισχύς ενός στατιστικού κριτηρίου είναι η πιθανότητα να απορριφθεί (ορθά) μία εσφαλμένη μηδενική υπόθεση (Greene, 2000). Η στατιστική ισχύς είναι «η πιθανότητα να πάρουμε ένα στατιστικά σημαντικό αποτέλεσμα όταν είναι ορθό το αναμενόμενο μέγεθος επίδρασης και εφόσον η έρευνα πραγματοποιήθηκε με σωστό τρόπο» (Bausell & Li, 2002, σελ.14). Ο Cohen (1988) υποστηρίζει πως «πρόκειται για την πιθανότητα να οδηγηθούμε στο συμπέρασμα ότι το φαινόμενο υφίσταται» (σελ. 4). Έτσι, μια μέτρια ισχύς γύρω στο 0,5 δείχνει ότι υπάρχει πιθανότητα 50% να βρούμε ένα στατιστικά σημαντικό αποτέλεσμα. Το συμπλήρωμα της ισχύος (1 - ισχύς) ή $\beta$ είναι η πιθανότητα του σφάλματος Τύπου ΙΙ. Ο

1. Το βιβλίο στην πρώτη του έκδοση του 1925 (ακολούθησαν άλλες 13 εκδόσεις) είναι διαθέσιμο σε ηλεκτρονική μορφή στη διεύθυνση: http://psychclassics.yorku.ca/Fisher/Methods/

Cohen (1988) έδωσε έμφαση στη διαδικασία στάθμισης αυτών των πιθανοτήτων πριν την εφαρμογή του στατιστικού κριτηρίου. Για παράδειγμα, όταν το *α* τοποθετηθεί στο 0,001, ελαχιστοποιείται η πιθανότητα σφάλματος Τύπου Ι, αλλά ο ερευνητής μπορεί να μειώσει την ισχύ του κριτηρίου στο 0,10, αυξάνοντας επομένως την πιθανότητα σφάλματος Τύπου ΙΙ στο 0,90 (1 - 0,10).

Ο Kirk (1996) συνοψίζει σε τρία σημεία την κριτική που έχει ασκηθεί στη διαδικασία του ελέγχου μηδενικών υποθέσεων: Καταρχάς, η διαδικασία δεν αποκαλύπτει στον ερευνητή αυτό που θέλει να μάθει. Ο έλεγχος της μηδενικής υπόθεσης και η επιστημονική έρευνα στοχεύουν στην απάντηση διαφορετικών ερωτημάτων. Στην επιστημονική έρευνα αυτό που θέλουμε να μάθουμε είναι την πιθανότητα να είναι αληθής η μηδενική υπόθεση με βάση τα δεδομένα που έχουμε συλλέξει [$p\ (H_0 \mid D)$]. Αντίθετα, ο έλεγχος μηδενικών υποθέσεων μας λέει πόση είναι πιθανότητα να έχουν προκύψει τα συγκεκριμένα δεδομένα εφόσον η μηδενική υπόθεση είναι αληθής [$p\ (D \mid H_0)$]. Έτσι, όταν η πιθανότητα να έχουν προκύψει τα ερευνητικά μας δεδομένα είναι χαμηλή (για παράδειγμα, $p = 0,02$), είναι λογικό σφάλμα να καταλήξουμε στο συμπέρασμα ότι η μηδενική υπόθεση είναι πιθανότατα ψευδής (αν και αυτό θα θέλαμε να αποδείξουμε στην πραγματικότητα). Αυτό, σύμφωνα με τον Cohen (1994), είναι το «σφάλμα της αντίστροφης πιθανότητας» (inverse probability error) και, σύμφωνα με τους Falk και Greenbaum (1995), η «πλάνη της πιθανολογικής απόδειξης μέσω της αντίφασης» (illusion of probabilistic proof by contradiction). Δυο πλάνες που σχετίζονται με τη συγκεκριμένη συλλογιστική είναι οι διαδεδομένες (αλλά εσφαλμένες) πεποιθήσεις πολλών ότι η τιμή p δηλώνει την πιθανότητα να είναι αληθής η μηδενική υπόθεση και το συμπλήρωμα του p, δηλαδή 1 - p, δηλώνει την πιθανότητα να βρούμε ένα στατιστικά σημαντικό αποτέλεσμα αν επαναλάβουμε την έρευνα.

Το δεύτερο σημαντικό πρόβλημα του ελέγχου μηδενικών υποθέσεων είναι ότι πρόκειται για μία διαδικασία μικρής σημασίας: όπως αναφέρουν πολλοί (π.χ., Cohen, 1994. Thompson, 1992. Tukey, 1991. Berkson, 1938), η διαφορά μεταξύ δύο επιδράσεων δεν είναι ποτέ μηδέν: θα υπάρχει οπωσδήποτε μία μικρή διαφορά της τάξης κάποιων δεκαδικών ψηφίων. Εφόσον η μηδενική υπόθεση είναι πάντοτε ψευδής, η απόφαση του ερευνητή να την απορρίψει σημαίνει απλώς ότι ο ερευνητικός σχεδιασμός είχε την απαιτούμενη ισχύ να εντοπίσει μία πραγματική κατάσταση, η οποία μπορεί και να είναι μία μεγάλη επίδραση. Μάλιστα, ο Cohen (1994) υποστήριξε ότι η τυπολατρική εμμονή στη διαδικασία ελέγχου μηδενικών υποθέσεων έχει οδηγήσει τους ερευνητές στο να εστιάζουν στον έλεγχο του σφάλματος Τύπου Ι (το οποίο δεν μπορεί να συμβεί καθώς ούτως ή άλλως όλες οι μηδενικές υποθέσεις είναι ψευδείς) και την ίδια στιγμή να επιτρέπουν στην πιθανότητα να συμβεί σφάλμα Τύπου ΙΙ (το οποίο μπορεί να συμβεί) να υπερβαίνει τα αποδεκτά όρια, συχνά μάλιστα να φτάνει μεταξύ του 0,50 και του 0,80.

Τέλος, μια κριτική που γίνεται στη διαδικασία συνοψίζεται στο ότι υιοθετώντας ένα σταθερό επίπεδο στατιστικής σημαντικότητας μετατρέπουμε το συνεχές της αβεβαιότητας σε μία διχοτομική απόφαση (απορρίπτω-δεν απορρίπτω). Η χρήση αυτής της στρατηγικής λήψης αποφάσεων μπορεί να οδηγήσει σε περιπτώσεις όπου δύο ερευνητές, ενώ μελετούν το ίδιο ερευνητικό ερώτημα και συλλέγουν δεδομένα που αποκαλύπτουν το ίδιο μέγεθος επίδρασης, καταλήγουν σε διαφορετικά συμπεράσματα από την έρευνά τους. Για παράδειγμα, ο ένας ερευνητής μετά τη στατιστική επεξεργασία των δεδομένων του μπορεί να καταλήξει ότι δεν πρέπει να απορρίψει τη μηδενική υπόθεση καθώς η τιμή p που υπολογίστηκε ήταν 0,055. Αντίστοιχα, ένας άλλος ερευνητής, ο οποίος απλώς χρησιμοποίησε ένα λίγο μεγαλύτερο δείγμα, μπορεί να απορρίψει τη μηδενική υπόθεση. Πρόκειται για την περίπτωση στην οποία αναφέρεται το σχόλιο των Rosnow και Rosenthal: «Σίγουρα, ο Θεός αγαπάει το 0,06 σχεδόν όσο και το 0,05» (1989, σελ. 1277).

Δεκάδες άρθρα έχουν αναφερθεί σε αυτές τις αδυναμίες του ελέγχου υποθέσεων (δύο σχετικά πρόσφατες ανασκοπήσεις είναι αυτές των MacCallum, 2003 και Nickerson, 2000). Βέβαια, έχουν υπάρξει και αρκετές απόπειρες αναμόρ-

φωσης της διαδικασίας, όπως αυτή του Harris (1997), ο οποίος πρότεινε τρεις εναλλακτικές ως αποτέλεσμα του ελέγχου υποθέσεων (να απορρίψουμε τη μηδενική υπόθεση, να αποτύχουμε να την απορρίψουμε, ή να αποφασίσουμε ότι δεν έχουμε αρκετές πληροφορίες για μία απόφαση). Ωστόσο, ο Cohen (1994) είναι απόλυτος: «...μην ψάχνετε για μια μαγική εναλλακτική στον έλεγχο μηδενικών υποθέσεων, μία άλλη αντικειμενική μηχανική διαδικασία για να τον αντικαταστήσετε. Δεν υπάρχει» (σελ. 1001).

### Σφάλματα κατά τον έλεγχο μηδενικών υποθέσεων: η εμπειρία από το περιοδικό *ΨΥΧΟΛΟΓΙΑ*

Οι αδυναμίες της διαδικασίας ελέγχου μηδενικών υποθέσεων στις οποίες αναφερθήκαμε παραπάνω έχουν ως αποτέλεσμα οι ερευνητές να αντιμετωπίζουν δυσκολίες και να υποπίπτουν σε σημαντικά σφάλματα τόσο κατά την παρουσίαση όσο και κατά την ερμηνεία των ευρημάτων τους. Η παρούσα εργασία έθεσε ως σκοπό τη μελέτη των σφαλμάτων που γίνονται από τους Έλληνες ερευνητές στο χώρο της ψυχολογίας με απώτερο στόχο της τη διατύπωση προτάσεων για τη βελτίωση των ακολουθούμενων πρακτικών.

Για το σκοπό αυτόν επιλέχτηκε το περιοδικό *ΨΥΧΟΛΟΓΙΑ*, το οποίο δημοσιεύει πρωτότυπα άρθρα εμπειρικής έρευνας και πρόκειται για ένα από τα ελάχιστα και τα πρώτα επιστημονικά περιοδικά της επιστήμης της ψυχολογίας στη χώρα μας και μία από τις ισχυρότερες πηγές διαμόρφωσης πρακτικών εφόσον χρησιμοποιείται για εκπαιδευτικούς σκοπούς από πολλούς συναδέλφους. Η απόφαση για τη δημοσίευση των άρθρων που υποβάλλονται λαμβάνεται με βάση την επιστημονική τους ποιότητα μέσα από σύστημα κριτών, ενώ οι συγγραφείς είναι υποχρεωμένοι να ακολουθούν τους κανόνες και τις υποδείξεις του εγχειριδίου της APA για τη μορφοποίηση των άρθρων τους.

### Μέθοδος

Τα ερευνητικά άρθρα που δημοσιεύτηκαν στα 60 τεύχη του περιοδικού *ΨΥΧΟΛΟΓΙΑ* αποτέ-

**Πίνακας 1**
**Κατανομή συχνότητας των άρθρων του περιοδικού ως προς τη γλώσσα και το περιεχόμενό τους**

| | **f** | **rf** |
|---|---|---|
| Ελληνική γλώσσα | 328 | 73,7 |
| Αγγλική γλώσσα | 117 | 26,3 |
| **ΣΥΝΟΛΟ** | **445** | **100,0** |
| | | |
| Ανασκόπηση βιβλιογραφίας | 143 | 32,1 |
| Ποσοτική έρευνα | 261 | 58,7 |
| Ποιοτική έρευνα | 23 | 5,2 |
| Ψυχομετρικό άρθρο | 11 | 2,5 |
| Παρουσίαση στατιστικής τεχνικής | 2 | 0,4 |
| Μελέτη περίπτωσης | 3 | 0,7 |
| Πιλοτική έρευνα – διαχρονική | 1 | 0,2 |
| Ψυχοφυσική | 1 | 0,2 |
| **ΣΥΝΟΛΟ** | **445** | **100,0** |

λεσαν το υλικό πάνω στο οποίο βασίστηκε η παρούσα μελέτη. Συγκεκριμένα, από τις αρχές του 1992 ως το τέλος του 2010 κυκλοφόρησαν 17 τόμοι (το 1993 και το 1994 δεν κυκλοφόρησε το περιοδικό) ή 60 τεύχη και δύο τεύχη ευρετήρια (ένα το 2000 και ένα το 2003), στα οποία δημοσιεύτηκαν συνολικά 445 άρθρα. Όπως φαίνεται στον Πίνακα 1, το 74% περίπου αυτών των άρθρων δημοσιεύτηκαν στην ελληνική γλώσσα, ενώ τα υπόλοιπα στην αγγλική. Ένας αξιοσημείωτος αριθμός (143 άρθρα ή το 32,1%) ήταν ανασκοπήσεις βιβλιογραφίας και τα υπόλοιπα ήταν κυρίως ποσοτικές και ποιοτικές έρευνες (58,7% και 5,2% αντίστοιχα). Ένας μικρός αριθμός άρθρων αναφερόταν σε ψυχομετρικά εργαλεία και ακόμη λιγότερα σε στατιστικές τεχνικές και μελέτες περίπτωσης.

Στον Πίνακα 2 παρουσιάζονται οι συχνότητες των στατιστικών κριτηρίων και τεχνικών που χρησιμοποιήθηκαν στα άρθρα αυτά. Θα πρέπει να σημειωθεί στο σημείο αυτό ότι εκτός από τα 143 άρθρα που αναφέρονταν σε ανασκόπηση βιβλιογραφίας, υπήρχαν και πέντε άρθρα στα οποία χρησιμοποιήθηκαν μόνο περιγραφικές στατιστικές τεχνικές. Έτσι, οι 579 απόλυτες συχνότητες του Πίνακα 2 προέρχονται από τα υπόλοιπα 297 άρθρα.

Το στατιστικό κριτήριο της Ανάλυσης Διακύμανσης (F-test από τον Fisher) στην απλή ή την παραγοντική του μορφή ήταν αυτό με την υψηλότερη συχνότητα (χρησιμοποιήθηκε σε 124 άρθρα,

**Πίνακας 2**
**Κατανομή συχνοτήτων των στατιστικών κριτηρίων στα άρθρα του περιοδικού**

| | f | rf |
|---|---|---|
| Ανάλυση Διακύμανσης | 124 | 21,4 |
| Συντελεστής συσχέτισης Pearson | 73 | 12,6 |
| Κριτήριο t | 64 | 11,1 |
| Διερευνητική Ανάλυση Παραγόντων | 58 | 10,0 |
| Πολλαπλή παλινδρόμηση | 40 | 6,9 |
| $\chi^2$ | 43 | 7,4 |
| Πολυμεταβλητή Ανάλυση Διακύμανσης | 30 | 5,2 |
| Επιβεβαιωτική Ανάλυση Παραγόντων | 18 | 3,1 |
| Ανάλυση Διαδρομών | 14 | 2,4 |
| Wilcoxon | 12 | 2,1 |
| 32 άλλες τεχνικές με f < 10* | 103 | 17,8 |
| ΣΥΝΟΛΟ | 579 | 100,0 |

* Οι τεχνικές αυτές ήταν: Ανάλυση Παραγόντων με μοντέλα Rasch, Ανάλυση Συνδιακύμανσης (ANCOVA), Πολυμεταβλητή Ανάλυση Συνδιακύμανσης (MANCOVA), Συντελεστής συσχέτισης Spearman, Mann-Whitney, Απλή παλινδρόμηση, Λογιστική Ανάλυση Παλινδρόμησης, Ανάλυση ομοιοτήτων με βάση τον Δυναμοπίνακα ισοτιμίας, Πολυεπίπεδη ανάλυση συμμεταβλητής δομής, Πολυδιάστατη γεωμετρική βαθμονόμηση ομοιοτήτων, Έλεγχος δομικής ισοτιμίας, Ανάλυση Συστάδων (Cluster Analysis), Ιεραρχική Ανάλυση Συστάδων, Σύγκριση διαφορών με τυπικές τιμές, Κριτήριο t με ένα δείγμα, Kruskal-Wallis, Friedman, McNemar, Εκθετική Συνάρτηση, Κριτήριο Student-Newman-Keuls, Μοντέλα Δομικών Εξισώσεων, Ανάλυση Αντιστοιχιών, Λογαριθμογραμμικά Μοντέλα, HLM, Διωνυμικά τεστ, Ανάλυση προβλέψεων, Πολυδιάστατη κλιμακοποίηση, Μήτρα πολλαπλών χαρακτηριστικών και πολλαπλών μεθόδων, Διαφορική λειτουργία της ερώτησης, Sign test, Ανάλυση Διακριτής Συνάρτησης (DFA), Αυτόματη Ιεραρχική Ταξινόμηση.

σχεδόν τις διπλάσιες φορές από το αμέσως επόμενο, τον συντελεστή συσχέτισης Pearson r). Τη μισή συχνότητα (χρησιμοποιήθηκε σε 64 άρθρα) είχε το κριτήριο Student's t. Επειδή το κριτήριο t και η Ανάλυση Διακύμανσης αποτέλεσαν τα δύο κύρια εργαλεία για τον έλεγχο υποθέσεων στην επιστήμη της ψυχολογίας, η μελέτη των άρθρων και οι παρατηρήσεις που ακολουθούν στη συνέχεια εστίασαν στη χρήση αυτών κυρίως των κριτηρίων και σε μικρότερο βαθμό του συντελεστή συσχέτισης Pearson r, του $\chi^2$, και της παλινδρόμησης.

## Αποτελέσματα

Μια πρώτη παρατήρηση, η οποία ωστόσο δεν σχετίζεται με τον έλεγχο υποθέσεων, είναι ότι υπάρχει μια σύγχυση σε μερικούς συγγραφείς σχετικά με τους ελληνικούς όρους που αποδίδουν τα ονόματα των διαφόρων στατιστικών κριτηρίων και τεχνικών. Μερικά τυχαία παραδείγματα: «Μέθοδος της πολλαπλής διακύμανσης» ή «πολλαπλή ανάλυση διακύμανσης» για να αποδοθεί ο όρος MANOVA, «Ανάλυση συνδιασποράς» για να αποδοθεί ο όρος MANCOVA, «Ανάλυση διαμεσολάβησης» για να αποδοθεί ο όρος ANCOVA (;), «ανάλυση του δικτύου των σχέσεων» για να αποδοθεί ο όρος Path Analysis, κ.ά. Πολύ περισσότεροι είναι εκείνοι οι συγγραφείς που αφήνουν τους όρους αμετάφραστους στην αγγλική εκδοχή τους. Βέβαια, πολλοί όροι έχουν προστεθεί στο λεξιλόγιό μας πολύ πρόσφατα και δεν υπήρχε μία έγκυρη μετάφραση των περισσότερων από αυτούς[2], τουλάχιστον μέχρι το 2009, οπότε και κυκλοφόρησε από την Εταιρεία των Ελλήνων Στατιστικών ένα έγκυρο λεξικό στατιστικής ορολογίας (Ελληνικό Στατιστικό Ινστιτούτο, 2009).

Σπανιότερα εμφανίζεται σύγχυση μεταξύ του στατιστικού κριτηρίου που χρησιμοποιήθηκε για την επεξεργασία των δεδομένων και αυτού που αναφέρεται από τους συγγραφείς. Ένα παράδειγμα: «... πραγματοποιήσαμε Ανάλυση Διακύμανσης (t-test)». Θα μπορούσε κανείς να υποθέσει ότι πρόκειται για σφάλμα αβλεψίας. Ωστόσο, λίγες αράδες παρακάτω αναφέρεται: «Σύμφωνα με τα αποτελέσματα της ανάλυσης διακύμανσης, διαπιστώθηκαν στατιστικά σημαντικές διαφορές μεταξύ των ομάδων..., $t(1, 11) = 21{,}12$, $p < 0{,}001$, και ... $t(1, 11) = 5{,}87$, $p < 0{,}036$».

Μια τρίτη ενδιαφέρουσα παρατήρηση είναι ότι σε πέντε μόλις άρθρα αναφέρεται έλεγχος των προϋποθέσεων για την πραγματοποίηση παραμετρικών στατιστικών κριτηρίων (π.χ., έλεγχος κανονικότητας των τιμών των εξαρτημένων μεταβλητών ή ομοιογένειας των διακυμάνσεων). Ένα ακόμη πιο εντυπωσιακό εύρημα είναι ότι μόνο ένα (1) άρθρο παρουσιάζει διαστήματα εμπιστοσύνης των μέσων τιμών.

Όπως δείχνουν τα στοιχεία που περιλάβαμε στον Πίνακα 3, είναι πολύ συχνά τα σφάλματα

**Πίνακας 3**
**Κατανομή συχνοτήτων των σφαλμάτων κατά την αναφορά των τιμών p**

| Σφάλμα | f |
|---|---|
| $p < 0{,}000$ | 5 |
| $p < 0{,}0001$ | 13 |
| $p = 0{,}000$ | 28 |
| $p = 0{,}0001$ | 3 |
| $p = 0{,}0000$ | 2 |
| $p > 0{,}0001$ | 1 |
| $p < 0{,}00005$ | 1 |
| $p < 0{,}0005$ | 2 |
| $p > 0{,}000$ | 1 |
| **ΣΥΝΟΛΟ** | **56** |

2. Οι όροι αυτοί θα πρέπει να γίνει μία προσπάθεια να προστεθούν στο Γλωσσάρι που δημοσιεύει η Επιτροπή Ορολογίας της ΕΛΨΕ στο περιοδικό – η τελευταία δημοσίευση ήταν στον τόμο 10, σε ένα τεύχος «Παράρτημα και Γλωσσάρι», το οποίο κυκλοφόρησε τον Δεκέμβριο του 2003, και στο οποίο υπάρχουν ελάχιστοι στατιστικοί όροι.

που αφορούν στην παρουσίαση των τιμών p. Για ορισμένα από αυτά είναι βέβαιο ότι πρόκειται για λάθη αβλεψίας, άλλα όμως επαναλαμβάνονται τόσο συχνά στο ίδιο άρθρο που δεν μπορεί να θεωρηθούν τυχαία.

Όλα αυτές οι μορφές παρουσίασης της τιμής p είναι λάθος, διότι α) η πιθανότητα δεν μπορεί να πάρει την τιμή 0 (ακόμη και σε εκείνες τις περιπτώσεις που ένα λογισμικό δίνει την τιμή 0,000, αυτή είναι στρογγυλοποιημένη και υπάρχει μετά το τρίτο δεκαδικό κάποιο ψηφίο διαφορετικό του 0), β) η APA συστήνει «να μην χρησιμοποιούμε καμία τιμή μικρότερη του 0,001» (APA, 2009, σελ. 139).

Εντοπίστηκαν επίσης αρκετά εσφαλμένα συμπεράσματα στα οποία καταλήγουν συγγραφείς αφού έχουν αποφασίσει να απορρίψουν ή έχουν αποτύχει να απορρίψουν τη μηδενική υπόθεση με βάση την τιμή p:

1. Μια πρώτη πλάνη σχετικά με την τιμή p είναι ότι αυτή αποτελεί αριθμητικό δείκτη του μεγέθους μίας επίδρασης. Έτσι, όσο μικρότερη είναι μία τιμή p τόσο μεγαλύτερο είναι το μέγεθος της επίδρασης. Η πλάνη αυτή ονομάζεται «πλάνη του μεγέθους» (Kline, 2004). Ένα παράδειγμα του συγκεκριμένου σφάλματος σε άρθρο του περιοδικού:

   > «Να σημειωθεί εδώ ότι τα αποτελέσματα πρέπει να ερμηνευθούν με προσοχή, λόγω της άνισης κατανομής των υποκειμένων ανά στάθμη της μεταβλητής..., αν και η τιμή του p (=.000) εξασφαλίζει πολύ χαμηλό επίπεδο στατιστικής σημαντικότητας, και άρα ενισχύει κατά πολύ την εγκυρότητα των αποτελεσμάτων».

   Παρόμοια έλλειψη κατανόησης δηλώνουν και τα ακόλουθα αποσπάσματα: «Το κριτήριο ... βρέθηκε στατιστικώς πολύ σημαντικό ($p < 0,001$), γεγονός ...», «...η διαφορά ανάμεσα στους μέσους όρους... προσεγγίζει το όριο στατιστικής σημαντικότητας ($p = 0,056$).», και «(οριακή) επίδραση του φύλου: $F(1, 215) = 3,75$, $p < 0,06$».

   Σύμφωνα με τον Cohen (1994), οι μικρότερες τιμές p δείχνουν χαμηλότερες δεσμευμένες πιθανότητες των δεδομένων, αν δεχτούμε ότι η μηδενική υπόθεση περιγράφει με ακρίβεια τον πληθυσμό και τίποτα περισσότερο. Αυτό συμβαίνει γιατί τα στατιστικά τεστ και οι τιμές p που υπολογίζονται από αυτά λαμβάνουν υπόψη το μέγεθος του δείγματος και το μέγεθος της επίδρασης, επομένως μία επίδραση ασήμαντου μεγέθους χρειάζεται απλώς ένα αρκετά μεγάλο δείγμα ώστε να αποδειχτεί στατιστικά σημαντική. Αν το μέγεθος του δείγματος είναι μεγάλο, τότε οι χαμηλές τιμές p απλώς επιβεβαιώνουν ένα μεγάλο δείγμα, το οποίο είναι μία ταυτολογία (Thompson, 1992). Βεβαίως, αποτελέσματα τα οποία είναι πράγματι μεγάλου μεγέθους μπορεί να έχουν και αυτά χαμηλές τιμές p – ωστόσο, αυτό είναι κάτι που δεν μπορεί να το εκτιμήσει κανείς από τις τιμές p μονάχα.
2. Ένα δεύτερο σφάλμα συνδέει την τιμή p με την πιθανότητα να είναι εσφαλμένη η απόφαση του ερευνητή να απορρίψει τη μηδενική υπόθεση (εφόσον το κάνει). Έτσι, αν $p < 0,05$, η πιθανότητα να οδηγηθούμε σε σφάλμα Τύπου I εφόσον αποφασίσουμε να απορρίψουμε τη μηδενική υπόθεση είναι μικρότερη από 5%. Η πλάνη αυτή είναι μία άλλη μορφή του σφάλματος της αντίστροφης πιθανότητας, την οποία ο Pollard (1993) περιέγραψε ως σύγχυση της δεσμευμένης αρχικής πιθανότητας του σφάλματος Τύπου I [$\alpha = p(\text{απόρριψη } H_0 \mid H_0)$] με τη δεσμευμένη εκ των υστέρων πιθανότητα του σφάλματος Τύπου I με δεδομένη την απόρριψη της μηδενικής υπόθεσης [$p(H_0 \mid \text{απόρριψη } H_0)$].

   Ένα παράδειγμα του συγκεκριμένου σφάλματος είναι το ακόλουθο:

   > «Το στατιστικό κριτήριο Student's t για εξαρτημένα δείγματα ... ήταν ... και η πιθανότητα στατιστικού σφάλματος ήταν μικρότερη του 0,001»

   Η απόφαση του ερευνητή να απορρίψει τη μηδενική υπόθεση είναι είτε ορθή είτε εσφαλμένη, επομένως δεν υπάρχει κάποια πιθανότητα σχετική με αυτήν. Μόνο αν υπάρξει ένας επαρκής αριθμός επαναλήψεων της έρευνας θα μπορούσε να εκτιμήσει αν η συγκεκριμένη

απόφαση να απορρίψει τη μηδενική υπόθεση ήταν ορθή.

3. Ένα ακόμη σφάλμα είναι αυτό που γίνεται από ορισμένους συγγραφείς, οι οποίοι αντιμετωπίζουν την απόρριψη της μηδενικής υπόθεσης ως επιβεβαίωση της ποιότητας του πειραματικού τους σχεδιασμού. Ο φτωχός ερευνητικός σχεδιασμός μπορεί να δημιουργήσει τεχνητές επιδράσεις, οι οποίες με τη σειρά τους μπορεί να οδηγήσουν στην εσφαλμένη απόρριψη της μηδενικής υπόθεσης. Επίσης, το σφάλμα δειγματοληψίας μπορεί να οδηγήσει σε σφάλμα Τύπου Ι ακόμη και σε μία καλά ελεγμένη έρευνα.

   Ένα παράδειγμα αυτού του σφάλματος θα μπορούσε να θεωρηθεί η περίπτωση μίας έρευνας μεγάλης κλίμακας, κατά την οποία συγκρίθηκαν οι μέσοι όροι πέντε ομάδων που αποτελούνταν από εκατοντάδες συμμετέχοντες η καθεμία. Ο συγγραφέας κατά την παρουσίαση του αποτελέσματός του αδιαφορεί για το μέγεθος του δείγματος, το μέγεθος της επίδρασης (δεν αναφέρονται πουθενά μεγέθη επίδρασης) και την κατεύθυνση της επίδρασης στις έρευνες αυτές. Αντίθετα, στη συζήτηση του άρθρου αναφέρεται σε προγενέστερες έρευνες, όπου η επίδραση της ίδιας ανεξάρτητης μεταβλητής δεν ήταν στατιστικά σημαντική και την αποδίδει στα μικρά δείγματα που χρησιμοποίησαν αυτοί οι ερευνητές. Και στη συζήτηση του άρθρου δεν γίνεται από τον συγγραφέα καμιά αναφορά στο μέγεθος της επίδρασης.
4. Επίσης, ένα πολύ συχνό σφάλμα είναι αυτό όπου συγχέεται η τιμή p με αυτήν του α (του επιπέδου στατιστικής σημαντικότητας). Στα χυολογούμε μερικά παραδείγματα: «...διαπιστώνεται ότι υπάρχει θετική συνάφεια μεταξύ των δύο μεταβλητών σε επίπεδο στατιστικής σημαντικότητας $p < 0,01$...», «Οι διαφορές των ομάδων... αποδεικνύονται στατιστικώς σημαντικές σε διαφορετικό επίπεδο ($p < 0,036$)».
5. Τέλος, ένα μάλλον σπάνιο φαινόμενο είναι αυτό ορισμένων συγγραφέων οι οποίοι, ενώ προφανώς αμφισβητούν τον ορισμό του επιπέδου στατιστικής σημαντικότητας στο 0,05, εντούτοις δεν παρουσιάζουν την ακριβή τιμή p που υπολόγισαν αλλά θεωρούν στατιστικά σημαντικές τιμές αρκετά μεγαλύτερες από το 0,05 χωρίς να μπουν στον κόπο να δηλώσουν μέχρι ποιο επίπεδο απορρίπτουν τη μηδενική υπόθεση (π.χ., ο συγγραφέας παρουσιάζει συχνά διαφορές μεταξύ δύο μέσων όρων ως στατιστικά σημαντικές και παραθέτει αποτελέσματα όπως «...($t = 1,75$, $p < 0,082$)...» ή «...($t = 1,72$, $p < 0,088$)...»).

Ο Yates (1951) υποστηρίζει ότι η χρήση του ελέγχου μηδενικών υποθέσεων «...έχει επηρεάσει τους ερευνητές έτσι ώστε να δίνουν αδικαιολόγητα μεγάλη προσοχή στο αποτέλεσμα της στατιστικής σημαντικότητας των κριτηρίων που εφαρμόζουν στα δεδομένα τους και ελάχιστη προσοχή στην εκτίμηση του μεγέθους των επιδράσεων που μελετούν... Η έμφαση στα κριτήρια στατιστικής σημαντικότητας και η εξέταση των αποτελεσμάτων κάθε πειράματος ατομικά είχαν δυστυχώς ως συνέπεια να θεωρούν οι ερευνητές την εκτέλεση ενός στατιστικού κριτηρίου σε ένα πείραμα ως τον απόλυτο στόχο» (σελ. 32-33). Αυτό είναι εντυπωσιακά εμφανές στη μεγάλη πλειονότητα των ερευνητικών άρθρων του περιοδικού: Από τις 179 φορές που χρησιμοποιήθηκαν συνολικά η ανάλυση διακύμανσης και το κριτήριο t, μόλις σε 22 περιπτώσεις (12,3%) έγινε υπολογισμός και παρουσίαση του μεγέθους της επίδρασης.

Στα παραπάνω θα μπορούσαμε να συνυπολογίσουμε την πραγματοποίηση αναλύσεων πολλαπλής παλινδρόμησης χωρίς καμιά αναφορά στο αν πληρούνταν οι ελάχιστες προϋποθέσεις για την εφαρμογή τους, τη χρήση του κριτηρίου $\chi^2$ για τη σύγκριση ποσοστών ή χωρίς τον έλεγχο των αναμενόμενων συχνοτήτων, τη χρήση του απαραμετρικού τεστ Wilcoxon για να επιβεβαιώσει ότι η διαφορά των μέσων όρων ήταν στατιστικά σημαντική, και αρκετά ακόμη που αφορούν σε άλλες τεχνικές (σφάλματα σημαντικά, τα οποία όμως ξεφεύγουν από το κύριο αντικείμενο του παρόντος άρθρου). Τα σφάλματα αυτά, όπως αναφέρθηκε και παραπάνω, προέρχονται από τα άρθρα του περιοδικού στη δημοσιευμέ-

νη μορφή τους. Είναι προφανές ότι θα βρίσκαμε πολύ περισσότερα αν ανατρέχαμε στη μορφή των άρθρων προτού περάσουν τη διαδικασία της κρίσης και της διόρθωσης. Το γεγονός αυτό δεν θα πρέπει να μας προκαλεί έκπληξη καθώς αρκετές εμπειρικές έρευνες έχουν δείξει ότι ακόμη και έμπειροι ερευνητές συχνά κάνουν σφάλματα στην ερμηνεία της διαδικασίας ελέγχου μηδενικών υποθέσεων (π.χ., Lecoutre, Poitevineau, & Lecoutre, 2003). Ωστόσο, θα βρίσκαμε πολύ λιγότερα αν οι κριτές είχαν στη διάθεσή τους ορισμένες συστηματοποιημένες κατευθύνσεις, και, βεβαίως, αν ο διορθωτής των τελικών κειμένων που υποβάλλονται για δημοσίευση ήταν γνώστης της ερευνητικής μεθοδολογίας και στατιστικής.

## Συζήτηση

Υπάρχει σημαντική τεκμηρίωση στη διεθνή βιβλιογραφία ότι οι παρανοήσεις και τα σφάλματα που εντοπίσαμε και παρουσιάσαμε στην προηγούμενη ενότητα είναι συνηθισμένα ακόμη και μεταξύ έμπειρων ερευνητών και ακαδημαϊκών δασκάλων. Για παράδειγμα, ο Cohen (1994) αναφέρει πολλούς γνωστούς συγγραφείς (μεταξύ αυτών και τον εαυτό του!), οι οποίοι έχουν υποπέσει στα βιβλία τους σε μία ή περισσότερες από τις πλάνες που αναφέρθηκαν παραπάνω. Ο Oakes (1986) σε έρευνά του με 70 ακαδημαϊκούς ψυχολόγους διαπίστωσε ότι μόλις το 11% έκανε ορθή ερμηνεία της τιμής «$p < 0,01$». Στην παράδοξη και αντιδιαισθητική λογική της διαδικασίας του ελέγχου μηδενικών υποθέσεων (Pollard, 1993) θα πρέπει να συνυπολογίσουμε και τη δυσκολία των ανθρώπων στη συλλογιστική με υποθετικούς συλλογισμούς (Anderson, 1998). Ένα επιπλέον πρόβλημα με τον έλεγχο μηδενικών υποθέσεων είναι ότι δεν αφήνει περιθώρια αξιολόγησης και κριτικής των αποτελεσμάτων μιας έρευνας καθώς αυτοματοποιεί τις υποκείμενες διεργασίες της συλλογιστικής και της λήψης αποφάσεων (Kline, 2004). Με τον τρόπο που διδάσκεται και εφαρμόζεται στις περισσότερες περιπτώσεις, οι ερευνητές εμπλέκονται σε έναν απλοϊκό και δίχοτομικό τρόπο σκέψης: βασισμένοι στην τιμή p καλούνται απλώς να απορρίψουν ή να δεχτούν τη μηδενική υπόθεση.

Το 1996 η Αμερικανική Ψυχολογική Εταιρία (APA) συγκρότησε μια ομάδα εργασίας με αντικείμενο τη στατιστική επαγωγή (ονομάστηκε Task Force on Statistical Inference – TFSI) και αποστολή να διευκρινίσει ορισμένα από τα επίμαχα ζητήματα που αναφέραμε παραπάνω. Η ομάδα αυτή δημοσίευσε το 1999 το πόρισμά της στο περιοδικό American Psychologist (Wilkinson & Task Force on Statistical Inference, 1999), ενώ οι περισσότερες από τις προτάσεις της ενσωματώθηκαν το 2001 στην 5η έκδοση του Εγχειριδίου της APA (Publication Manual, APA, 2001). Αν και στις οδηγίες συγγραφής του περιοδικού *ΨΥΧΟΛΟΓΙΑ* αναφέρεται σαφώς ότι οι συγγραφείς πρέπει να ακολουθούν τους κανόνες του Εγχειριδίου της APA, όπως δείξαμε παραπάνω, κάτι τέτοιο δεν γίνεται συστηματικά, τουλάχιστον μέχρι σήμερα.

Με αποκλειστικό σκοπό αφενός τη βελτίωση των πρακτικών μας και αφετέρου τη βαθύτερη κατανόηση των ευρημάτων μας, στις επόμενες σελίδες επιχειρείται μια σύνοψη ορισμένων προτάσεων και οδηγιών προς τους συγγραφείς ερευνητικών αναφορών και εργασιών. Για τη διαμόρφωση αυτών των οδηγιών χρησιμοποιήθηκαν αντίστοιχες απόπειρες από τη διεθνή βιβλιογραφία (π.χ. Wilkinson & TFSI, 1999. Loftus, 1996. Cohen, 1990. Bailar & Mosteller, 1988). Οι προτάσεις αυτές ούτε είναι ούτε φιλοδοξούν να αποτελέσουν έναν πλήρη οδηγό για την αναφορά των ερευνητικών ευρημάτων. Οι αναγνώστες θα πρέπει να συνεχίσουν να ανατρέχουν στο Εγχειρίδιο της APA για διεξοδικές κατευθύνσεις και πληροφορίες.

### Ισχύς και μέγεθος του δείγματος

Η πληροφορία σχετικά με το μέγεθος του δείγματος δεν είναι αρκετή. Συνιστάται να γίνεται αναφορά στις διαδικασίες εκείνες που ακολουθήθηκαν για την επιλογή του συγκεκριμένου δείγματος, αλλά και να παρέχεται πληροφόρηση για τις υποθέσεις των ερευνητών σχετικά με τα μεγέθη επίδρασης, τη δειγματοληψία και τη μέτρηση, αλλά και για τους υπολογισμούς της στατι-

στικής ισχύος. Ακριβώς επειδή οι υπολογισμοί της στατιστικής ισχύος έχουν νόημα όταν πραγματοποιούνται πριν τη συλλογή των ερευνητικών δεδομένων, είναι χρήσιμο να παρουσιάζεται ο τρόπος με τον οποίο προέκυψαν οι εκτιμήσεις των μεγεθών της επίδρασης από προγενέστερες έρευνες και δημοσιεύσεις.

Κάτι τέτοιο είναι πολύ εύκολο σήμερα που είναι διαθέσιμα αρκετά λογισμικά (κάποια μάλιστα, όπως το G*Power[3], διατίθενται δωρεάν μέσω του Διαδικτύου) τα οποία επιτρέπουν τον υπολογισμό της στατιστικής ισχύος για διάφορους ερευνητικούς σχεδιασμούς και κατανομές δεδομένων. Βεβαίως, ο υπολογισμός της ισχύος δεν είναι δυνατός σε ορισμένους στατιστικούς ελέγχους και συγκεκριμένα όταν δεν είναι δυνατόν να προσδιοριστούν επακριβώς οι βαθμοί ελευθερίας (για παράδειγμα, στα πολυεπίπεδα μοντέλα).

Όταν το μέγεθος του δείγματος σε έναν πίνακα, σχήμα ή πρόταση είναι διαφορετικό από αυτό που αναφέρθηκε στη Μέθοδο, θα πρέπει να εξηγείται η διαφορά και πώς προέκυψε αυτή η «απώλεια».

### Έλεγχος των ερευνητικών δεδομένων

Πριν τη στατιστική επεξεργασία και την παρουσίαση των δεδομένων καλό είναι αυτά να εξετάζονται με προσοχή για ελλείπουσες και ακραίες τιμές καθώς και για σφάλματα κατά την ψηφιοποίησή τους. Στην περίπτωση ύπαρξης ελλειπουσών τιμών είναι χρήσιμο να αναφέρεται το ποσοστό τους στο σύνολο των μετρήσεων μαζί με οποιεσδήποτε ερμηνείες για την ύπαρξή τους.

Προσοχή! Ο έλεγχος των δεδομένων δεν γίνεται με σκοπό να απορρίψουμε δεδομένα ή –χειρότερα– να αλλάξουμε τιμές ώστε να ταιριάζουν με τις υποθέσεις μας. Τόσο η επιλογή της διαγραφής κατά ζεύγη (pairwise) όσο και αυτής κατά λίστα (listwise) είναι από τις χειρότερες επιλογές που μπορεί να κάνει κανείς και καλό είναι να αποφεύγονται. Για εναλλακτικές μεθόδους χειρισμού των ελλειπουσών τιμών εξαιρετικά χρήσιμα είναι τα βιβλία των Tabachnick και Fidell (2007) ή των Little και Rubin (1987).

### Επιλογή του κατάλληλου στατιστικού κριτηρίου

Σήμερα οι ερευνητές έχουν στη διάθεσή τους ένα μεγάλο αριθμό ποσοτικών μεθόδων επεξεργασίας των δεδομένων τους, ο οποίος μάλιστα αυξάνει συνεχώς. Η επιλογή του στατιστικού κριτηρίου δεν είναι σωστό να γίνεται με κριτήριο τον εντυπωσιασμό των αναγνωστών ή τον περιορισμό της αρνητικής κριτικής. Θα πρέπει να γίνεται αναφορά στους λόγους για τους οποίους επιλέχτηκε το συγκεκριμένο στατιστικό κριτήριο καθώς επίσης και στις πιθανές αδυναμίες του ερευνητικού σχεδιασμού. Με αυτό τον τρόπο οι αναγνώστες θα είναι σε θέση να σχηματίσουν μία καθαρή εικόνα σχετικά με την αξιοπιστία των δεδομένων καθώς και για τις πιθανές απειλές κατά της εγκυρότητας των ευρημάτων και των ερμηνειών τους. Αν ένα απλούστερο στατιστικό τεστ ταιριάζει για τον έλεγχο των ερευνητικών δεδομένων, είναι προτιμότερο αντί ενός περισσότερο περίπλοκου.

### Έλεγχος των προϋποθέσεων για τη χρήση ενός στατιστικού κριτηρίου

Μεγάλη προσοχή θα πρέπει να δίνεται στον έλεγχο των ερευνητικών δεδομένων για να διαπιστωθεί αν αυτά πληρούν τις προϋποθέσεις για τη χρήση κάποιου στατιστικού κριτηρίου. Καλό είναι να εξετάζουμε με προσοχή τα στατιστικά υπόλοιπα (residuals). Είναι προτιμότερος ο διαγραμματικός έλεγχος των υπολοίπων (π.χ., διαγράμματα μίσχου-φύλλων/φυλλογραφήματα, σημειογράμματα, θηκογράμματα, ιστογράμματα, κ.ο.κ.) αντί των στατιστικών δεικτών (π.χ., συμμετρία και κύρτωση) ή των τεστ ελέγχου της κατανομής (π.χ., έλεγχος Kolmogorov-Smirnov), που συχνά είναι εξαιρετικά ευαίσθητα.

3. Διαθέσιμο στη διεύθυνση: http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/

**Παρουσίαση των δεδομένων**

Δεν θα πρέπει να παραλείπεται η αναφορά πληροφοριών όπως ο μέσος όρος, η τυπική απόκλιση και το πλήθος των περιπτώσεων που πήραν μέρος στην ανάλυση (N). Από αυτές τις τρεις πληροφορίες μπορεί ο αναγνώστης να υπολογίσει την τιμή p (στην περίπτωση που έχουμε σχεδιασμό ανεξάρτητων δειγμάτων), όχι όμως και το αντίθετο...

**Διαστήματα εμπιστοσύνης**

Μια αλλαγή προς την οποία θα πρέπει να κινηθούμε είναι η εισαγωγή κατάλληλων δεικτών του σφάλματος μέτρησης: τα διαστήματα εμπιστοσύνης –που απουσιάζουν εμφαντικά από τα άρθρα του περιοδικού– δεν θα πρέπει να παραλείπονται. Τα διαστήματα εμπιστοσύνης του μέσου όρου άλλωστε μας δίνουν πληροφορίες όχι μόνο για το επίπεδο του μέσου όρου αλλά και για τη μεταβλητότητά του (variability). Ενώ ο έλεγχος υποθέσεων καταδεικνύει μόνο την ύπαρξη ή όχι μίας επίδρασης, τα διαστήματα εμπιστοσύνης σε συνδυασμό με τους δείκτες του μεγέθους της επίδρασης δείχνουν πόσο μεγάλη είναι μία επίδραση και με πόση ακρίβεια μπορεί κανείς να την εκτιμήσει.

Όπου μάλιστα είναι δυνατό, είναι εξαιρετικά χρήσιμο να κατασκευάζονται ραβδογράμματα σφάλματος (error bar charts), τα οποία αποτελούν γραφικές αναπαραστάσεις των εκτιμήσεων διαστήματος (Wilkinson & TFSI, 1999). Η χρήση των ραβδογραμμάτων σφάλματος με τη μορφή διαστημάτων εμπιστοσύνης 95% γύρω από τις μέσες τιμές των συνθηκών του πειράματος είναι ιδανικός τρόπος παρουσίασης των δεδομένων καθώς συνδυάζονται τα αποτελέσματα των δύο αυτών τεχνικών ανάλυσης των δεδομένων και απεικονίζεται η ιδιαίτερη σημασία της καθεμιάς από τις δύο (Loftus, 2002).

**Έλεγχος των ερευνητικών υποθέσεων**

Θα πρέπει πάντοτε να διευκρινίζεται αν πραγματοποιήθηκε αμφίπλευρος ή μονόπλευρος έλεγχος της μηδενικής υπόθεσης.

Η ατυχής και εσφαλμένη έκφραση «αποδοχή της μηδενικής υπόθεσης» δεν θα πρέπει να χρησιμοποιείται ποτέ.

Είναι ορθότερο να αναφέρεται η ακριβής τιμή p (π.χ. $p = 0,024$) αντί για εκφράσεις, όπως $p < 0,05$ ή «στατιστικά μη σημαντική» ή $p > 0,05$. Η πληροφορία αυτή είναι χρήσιμη σε εκείνους τους αναγνώστες που θα ήθελαν να συγκρίνουν αποτελέσματα από διαφορετικές συναφείς έρευνες ή να εκτιμήσουν το αποτέλεσμα της συγκεκριμένης έρευνας σε σχέση με ένα επίπεδο στατιστικής σημαντικότητας της δικής τους επιλογής.

Όταν πραγματοποιούνται έλεγχοι πολλαπλών συγκρίσεων, επειδή η πιθανότητα λανθασμένης απόρριψης της μηδενικής υπόθεσης σε έναν τουλάχιστον από αυτούς είναι υψηλή (πρόκειται για το σφάλμα Τύπου Ι), θα πρέπει να προσαρμόζεται κατάλληλα το επίπεδο στατιστικής σημαντικότητας (για παράδειγμα, με μία από τις μεθόδους των Dunn-Bonferroni, Holm-Bonferroni, Dunn-Šidák, κ.ά.).

Κάποια εκτίμηση μεγέθους επίδρασης θα πρέπει να συνοδεύει πάντοτε την τιμή p καθώς και τα βασικά ευρήματα. Το μέγεθος της επίδρασης επιτρέπει στους αναγνώστες να αξιολογήσουν τη σταθερότητα των ευρημάτων σε διαφορετικά δείγματα, σχεδιασμούς και αναλύσεις, ενώ διευκολύνει και τις αναλύσεις ισχύος ή τις μεταναλύσεις που μπορεί να πραγματοποιηθούν στο μέλλον. Το Εγχειρίδιο της APA ήδη από την έκδοση του 2001 αναφέρει σχετικά: «Προκειμένου ο αναγνώστης να κατανοήσει πλήρως τη σημασία των ευρημάτων σας, είναι σχεδόν πάντα απαραίτητο να περιλάβετε στο τμήμα των αποτελεσμάτων κάποιο δείκτη μεγέθους της επίδρασης ή της συνάφειας. Μπορείτε να υπολογίσετε το μέγεθος της επίδρασης ή της συνάφειας με έναν αριθμό από συνήθεις δείκτες, οι οποίοι περιλαμβάνουν (αλλά δεν περιορίζονται) το $r^2$, το $\eta^2$, το $\omega^2$, το $R^2$, το $\varphi^2$, το V του Cramer, το W του Kendall, το d και το κ του Cohen, το λ και το γ των Goodman και Kruskal, τα μέτρα κλινικής σημαντικότητας των Jacobson και Traux (1991), και του Kendall (1999), καθώς και τα πολυμεταβλητά Θ του Roy και V των Pillai και Bartlett» (σελ. 25-26).

Το βιβλίο των Rosenthal, Rosnow και Rubin

(2000) θα φανεί χρήσιμο στον αναγνώστη που επιθυμεί να μάθει περισσότερα για τα μεγέθη της επίδρασης.

### Ερμηνεία των ευρημάτων της διαδικασίας ελέγχου υποθέσεων

Ιδιαίτερη προσοχή θα πρέπει να δίνεται κατά τη μη τεχνική χρήση των τεχνικών όρων «στατιστικά σημαντικός/ή/ό», «τυχαίος», «κανονικό», «επίδραση», «συσχέτιση» κ.λπ. Πολλοί από αυτούς τους στατιστικούς όρους έχουν διαφορετική σημασία όταν χρησιμοποιούνται στον καθημερινό μας λόγο.

Οι ερευνητές που πραγματοποιούν πειράματα βασισμένα σε μη τυχαίους ερευνητικούς σχεδιασμούς (το συχνότερο) θα πρέπει να είναι πολύ προσεκτικοί με την εξαγωγή συμπερασμάτων που αφορούν σε αιτιώδεις σχέσεις, καθώς η ύπαρξη συμμεταβλητών στους σχεδιασμούς αυτούς είναι πολύ πιθανό ότι μπορεί να οδηγήσει σε διαφορετικές ερμηνείες για τα ευρήματα.

### Χρήση εναλλακτικών μεθόδων ανάλυσης

Οι συγγραφείς θα πρέπει να ενθαρρύνονται να χρησιμοποιούν εναλλακτικές μεθόδους ανάλυσης, όπως είναι οι τυχαιοποιημένοι έλεγχοι (βλ. Edgington, 1995), οι οποίοι, ενώ χρησιμοποιούν την έννοια της στατιστικής σημαντικότητας, δεν περιλαμβάνονται στους γνωστούς ελέγχους και δεν έχουν ευρεία διάδοση. Επίσης, τις τελευταίες δύο δεκαετίες έχουν αναπτυχθεί ιδιαίτερα ορισμένες τεχνικές που βασίζονται σε μαθηματικά μοντέλα, όπως τα πολυεπίπεδα μοντέλα ή τα λογαριθμογραμμικά μοντέλα, τα οποία επιτρέπουν στον ερευνητή να ξεφύγει από τη μηχανιστική εφαρμογή του ελέγχου μηδενικών υποθέσεων και να γίνει πολύ πιο δημιουργικός κατασκευάζοντας και ελέγχοντας επιστημονικά μοντέλα (βλ. Rodgers, 2010).

## Επιλεγόμενα

Ίσως οι προτάσεις που διατυπώθηκαν παραπάνω να ακούγονται περισσότερο από ό,τι πρέπει κατευθυντήριες. Ωστόσο, ως επιστημονική κοινότητα έχουμε αποδεχτεί προ πολλού λεπτομερείς οδηγίες σχετικά με το πώς θα παρουσιάζουμε λιγότερο σημαντικές πληροφορίες (όπως οι παραπομπές και οι βιβλιογραφικές μας πηγές, για την παρουσίαση των οποίων στην τελευταία έκδοση του Εγχειριδίου της APA έχουν αφιερωθεί δύο κεφάλαια και 55 σελίδες). Επομένως, ίσως αξίζει να εκχωρήσουμε λίγο περισσότερη ελευθερία για χάρη της κατανόησης.

Οι Fidler et al. (2004) στον τίτλο ενός άρθρου τους υποστήριξαν ότι «οι διευθυντές σύνταξης μπορούν να κατευθύνουν τους ερευνητές στη χρήση των διαστημάτων εμπιστοσύνης, αλλά δεν μπορούν να τους κάνουν να σκέφτονται». Οι ερευνητές εξέτασαν 594 άρθρα του American Journal of Public Health, τα οποία δημοσιεύτηκαν μεταξύ 1982 και 2000, για να διαπιστώσουν ότι, ενώ οι συγγραφείς των άρθρων αυτών ακολουθούσαν τις οδηγίες της συντακτικής επιτροπής για χρήση των διαστημάτων εμπιστοσύνης (μέσα στην εικοσαετία η χρήση των διαστημάτων εμπιστοσύνης είχε αυξηθεί από 10% σε 54%), η χρήση ήταν εντελώς επιφανειακή καθώς στη συζήτηση των ευρημάτων ελάχιστοι αναφέρονταν σε αυτά.

Είναι προφανές ότι εκτός από κατευθύνσεις για «καλύτερες πρακτικές» χρειαζόμαστε κατάλληλη εκπαίδευση (Gigerenzer, 1998). Οι πανεπιστημιακοί δάσκαλοι που διδάσκουν στατιστική έχουν να αντιμετωπίσουν πολλές προκλήσεις: πολλές από τις στατιστικές έννοιες είναι πράγματι περίπλοκες, δυσνόητες και αντιδιαισθητικές. Είναι πολύ δύσκολο να κινητοποιήσει κανείς τους φοιτητές να μελετήσουν στατιστική. Πολλοί από αυτούς τους φοιτητές αντιμετωπίζουν δυσκολίες με τα μαθηματικά που συναντούν παντού (στους αλγεβρικούς τύπους, στους δεκαδικούς αριθμούς, στα κλάσματα κ.α.) ή κάνουν το λάθος να θεωρούν ότι –όπως ακριβώς συμβαίνει στα μαθηματικά– η έμφαση είναι στους αριθμούς, στους υπολογισμούς, στους τύπους και στη μία σωστή απάντηση. Αισθάνονται πολύ άβολα με την «αναστάτωση» που χαρακτηρίζει τα ερευνητικά δεδομένα, με τις διαφορετικές πιθανές ερμηνείες ανάλογα με τις διαφορετικές υποθέσεις, καθώς και

με την εκτεταμένη χρήση των δεξιοτήτων συγγραφής και επικοινωνίας που απαιτείται. Εστιάζοντας στην αντιμετώπιση αυτών των προκλήσεων, ένας δάσκαλος πολύ συχνά αποτυγχάνει και να αγγίξει ακόμη τον πιο σημαντικό στόχο της διδασκαλίας της στατιστικής: αυτόν της ανάπτυξης των ικανοτήτων στατιστικής σκέψης και συλλογιστικής (Ben-Zvi & Garfield, 2004).

Χρειάζεται, επίσης, η ανάπτυξη τεχνικών για βελτιωμένες στατιστικές πρακτικές, οι οποίες ωστόσο θα πρέπει να βασιστούν σε εμπειρικές μελέτες των τρόπων με τους οποίους οι ίδιοι οι ερευνητές κατανοούν τις έννοιες του ελέγχου υποθέσεων, των εκτιμήσεων σημείου και των διαστημάτων εμπιστοσύνης, της στατιστικής σημαντικότητας, του μεγέθους της επίδρασης, κ.ά.

Σύμφωνα με τον Gigerenzer (1993), η καθιέρωση της διαδικασίας του ελέγχου μηδενικών υποθέσεων ως της κυρίαρχης προσέγγισης για την εξέταση των ερευνητικών ερωτημάτων οδήγησε στο να θεωρείται η στατιστική σημαντικότητα κριτήριο της καλής έρευνας και το αντίθετό της «ένα αρνητικό, άχρηστο, δυσάρεστο αποτέλεσμα» (σελ. 319). Το παρόν άρθρο επιχείρησε να ανοίξει τη συζήτηση σχετικά με τις πρακτικές μας για την ερμηνεία της στατιστικής σημαντικότητας. Ωστόσο, θα πρέπει πολύ σύντομα να ξεκινήσουμε και τη συζήτηση για την πρακτική, την ψυχολογική σημαντικότητα των ευρημάτων μας. Με άλλα λόγια, ένα στατιστικά σημαντικό αποτέλεσμα δεν σημαίνει απαραίτητα ότι το εύρημα είναι άξιο λόγου, αλλά και το αντίθετο: το να διδάξουμε σε έναν αναλφάβητο με Alzheimer να διαβάζει 10 νέες λέξεις μπορεί να μην είναι στατιστικά σημαντική διαφορά από την προηγούμενη κατάστασή του, αλλά ίσως να είναι εξαιρετικό επίτευγμα για άτομα αυτού του πληθυσμού. Συνεπώς, η στατιστική σημαντικότητα θα πρέπει να αντιμετωπίζεται ως «σχετική» μάλλον και όχι ως «απόλυτη» έννοια.

## Βιβλιογραφία


American Psychological Association (2009). *The Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: APA.

American Psychological Association (2001). *The Publication Manual of the American Psychological Association* (5th ed.). Washington, DC: APA.

Anderson, J. L. (1998). Embracing uncertainty: The interface of Bayesian statistics and cognitive psychology. *Conservation Ecology, 2* (1), 2. Available from the Internet. URL: http://www.consecol.org/vol2/iss1/art2/

Bailar, J. C. & Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals. *Annals of Internal Medicine, 108*, 266-273.

Bausell, R. B. & Li, Y. F. (2002). *Power analysis for experimental research: A practical guide for the biological, medical and social sciences.* Cambridge, UK: Cambridge University Press.

Ben-Zvi, D. & Garfield, J. (2004). Statistical literacy, reasoning, and thinking: Goals, definitions, and challenges. In D. Ben-Zvi and J. Garfield (Eds.), *The Challenge of Developing Statistical Literacy, Reasoning and Thinking* (pp. 3-15). Kluwer Academic Publishers.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33*, 526-542.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist, 49*, 997-1003.

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Marcel Dekker.

Ελληνικό Στατιστικό Ινστιτούτο (2009). *Λεξικό Στατιστικής Ορολογίας Αγγλο-Ελληνικό & Ελληνο-Αγγλικό.* Αθήνα: ΕΣΙ.

Falk, R. & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5*, 75-98.

Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science, 15*, 119-126.

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Oliver & Boyd.

Fisher, R. A. (1935). *The design of experiments*. Edinburgh: Oliver & Boyd.

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver & Boyd.

Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences*, *21*, 199-200.

Gigerenzer, G. & Murray, D. J. (1987). *Cognition as Intuitive Statistics*. Hillsdale, NJ: Erlbaum.

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). *The empire of chance. How probability changed science and every day life*. Cambridge: Cambridge University Press.

Greene, W. H. (2000). *Econometric Analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. L. Harlow, S. A. Mulaik & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145-174). Mahwah, NJ: Erlbaum.

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746-759.

Kline, R. B. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington, DC: American Psychological Association.

Krüger, L., Daston, L., & Heidelberger, M. (Eds.). (1987). *The probabilistic revolution: Vol. 1. Ideas in history.* Cambridge, MA: MIT Press.

Lecoutre, M.P., Poitevineau, J., & Lecoutre, B. (2003). Even statisticians are not immune to misinterpretations of Null Hypothesis Significance Tests. *International Journal of Psychology, 38* (1), 37-45.

Little, R. J. A. & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161-171.

Loftus, G. R. (2002). Analysis, interpretation, and visual presentation of data. *Stevens' Handbook of Experimental Psychology*, (3rd ed.), Vol 4. (pp. 339-390). New York: John Wiley and Sons.

MacCallum, R. C. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38*(1), 113-139.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika, 29A*, Part I: 175-240; Part II: 263-294.

Nickerson, R. S. (2000). Null hypothesis statistical testing: A review of an old and continuing controversy. *Psychological Methods, 5*, 241-301.

Oakes, M. (1986). *Statistical inference.* New York: Wiley.

Pollard, P. (1993). How significant is "significance"? In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 448-490). Hillsdale, NJ: Erlbaum.

Porter, T. M. (1986). *The rise of statistical thinking 1820-1900*. Princeton, NJ: Princeton University Press.

Rodgers, J. L. (2010). The epistemology of mathematical and statistical modeling: A quiet methodological revolution. *American Psychologist, 65*(1), 1-12.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach*. Cambridge University Press.

Rosnow, R. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276-1284.

Tabachnick, B. G. & Fidell, L. S. (2007). *Using Multivariate Statistics* (5th ed.). Boston: Allyn and Bacon.

Thompson, B. (1992). Two and one-half decades of leadership in measurement and evaluation. *Journal of Counseling and Development, 70,* 434-438.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*, 100-116.

Wilkinson, L. & Task Force on Statistical Inference, APA Board of Scientific Affairs. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604.

Yates, F. (1951). The influence of "statistical methods for research workers" on the development of the science of statistics. *Journal of the American Statistical Association, 46*, 19-34.

# Null Hypothesis Significance Testing: procedure, misconceptions and some suggestions for good practices

PETROS ROUSSOS[1]

*ABSTRACT*

The rationale of Null Hypothesis Significance Testing (NHST) is described, and the consequences of its hybridism are discussed. The paper presents with descriptive methods and critically discusses the ways in which the authors of the research papers published in "PSYCHOLOGY: The Journal of the HPS" refer to NHST and interpret its outcomes. We examined the 445 articles published between 1992 and 2010, we noted misuses of NHST and searched for any use of confidence intervals or error bars or use of these to support interpretation. Part of the paper focuses on the statistical-reform debate and provides detailed guidance about good statistical practices in the analysis of our research data and the interpretation of our findings. The proposed guide does not fall into the trap of mandating the use of particular procedures; it rather aims to support readers' understanding of the research results.

*Keywords:* Null hypothesis significance testing, Statistical reasoning

1. *Address*: Department of Psychology, School of Philosophy, University of Athens, University Campus, Ilisia, GR-15784, Athens, Tel.: 210 7277385, e-mail: roussosp@psych.uoa.gr