Η σημαντικότητα των ευρημάτων

Τὰ ὄργανα ἅτινα ὁ ἄνθρωπος κατεσκεύασε διὰ νὰ εὑρύνῃ καὶ βαθμολογήσῃ τὸ πλαίσιον τῶν ἰδιοτήτων αὐτοῦ, ἐπέτρεψαν εἰς τοῦτον νὰ ἀντικαταστήσῃ εἰς πλεῖστα σημεῖα τὰς ποιοτικὰς ἐντυπώσεις διὰ ποσοτικῶν μετρήσεων. Ἔκτοτε ἠδυνήθη νὰ προσδιορίσῃ μετ' ἀκριβείας τὰς σταθερὰς σχέσεις καὶ νὰ ἀνακαλύψῃ τοὺς νόμους.

Κωσταντίνος Αθανασιάδης *Η Στατιστική και αι Επιστήμαι Παρατηρήσεως* σελ. 30.

Η στατιστική σημαντικότητα

Το γράμμα που συμβολίζει το επίπεδο στατιστικής σημαντικότητας στη

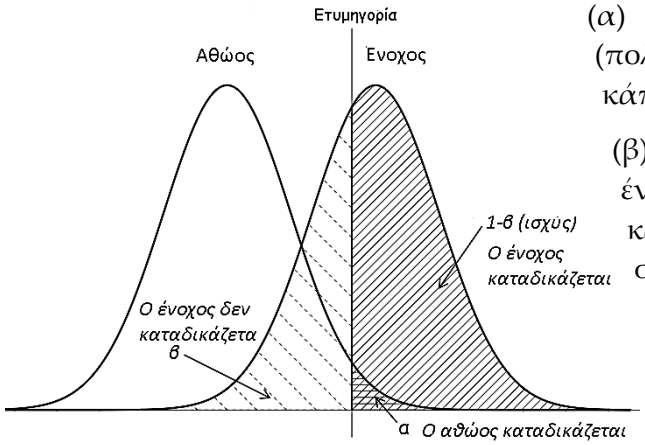
Είδαμε μέχρι τώρα ότι ως ερευνητές, συλλέγουμε στατιστικά δεδομένα ώστε να ελέγξουμε στατιστικές υποθέσεις ή να κατασκευάσουμε μοντέλα της πραγματικότητας. Είδαμε επίσης πώς ο έλεγχος των υποθέσεων και η κατασκευή των μοντέλων απαιτεί αναγωγή των αποτελεσμάτων σε θεμελιωμένες και ευρέως χρησιμοποιούμενες θεωρητικές κατανομές πιθανοτήτων (π.χ. z, t, F, x^2 κλπ.). Οι κατανομές αυτές χρησιμοποιούνται στον έλεγχο υποθέσεων με τη βασική αρχή ότι μπορεί στην έρευνά μας να έχουμε μόνο ένα μικρό δείγμα παρατηρήσεων (εξάλλου η αδυναμία να ερευνήσουμε όλον τον πληθυσμό είναι η βάση της Στατιστικής), αλλά θεωρητικά θα μπορούσαμε (αν μπορούσαμε) να πάρουμε πάρα πολλά τέτοια δείγματα και να φτιάξουμε με αυτά τις λεγόμενες δειγματοληπτικές κατανομές. Αν σύμφωνα με τις δειγματοληπτικές κατανομές το εύρημά μας είναι συνηθισμένο τότε διατηρούμε την αμφιβολία μας σε σχέση με αυτό το οποίο ελέγχουμε (π.χ. μια συσχέτιση μεταξύ μεταβλητών). Αν όμως το εύρημά μας είναι σπάνιο (π.χ. μια μεγάλη διαφορά στους μέσους όρους δύο ομάδων μαθητών) τότε δύο πράγματα μπορεί να συμβαίνουν: είτε δεν υπάρχει διαφορά μεταξύ των δύο αυτών ομάδων και απλώς εμείς «έτυχε» να επιλέξουμε στο δείγμα μας τις ακραίες περιπτώσεις είτε υπάρχει διαφορά μεταξύ των δύο αυτών ομάδων και εμείς την βρήκαμε. Τι από τα δύο συμβαίνει; Η απάντηση στο πιο πάνω ερώτημα εμπεριέχει ένα επίπεδο σφάλματος.

Για να δώσουμε ένα παράδειγμα για την έννοια της στατιστικής σημαντικότητας θα φέρουμε ένα νοητικό παράδειγμα από τον δικαστικό χώρο. Σε μια δίκη ο κατηγορούμενος είναι είτε αθώος είτε ένοχος. Στο Σχήμα 25 η αριστερή

κατανομή είναι αυτή της αθωότητας ενώ η δεξιά είναι αυτή της ενοχής. Σε μία από τις δύο κατανομές ανήκει ο κατηγορούμενος, ανάλογα με το αν διέπραξε ή δεν διέπραξε το αδίκημα. Εμείς δεν είμαστε σίγουροι αν ο κατηγορούμενος είναι αθώος ή ένοχος. Δεν τον είδε κανείς να διαπράττει το αδίκημα. Κατ' αρχήν, όμως, πιστεύουμε στην αθωότητά του[1] ή, για να το πούμε στατιστικά, πιστεύουμε στη «μηδενική υπόθεση».

Κατά τη διάρκεια της δίκης έρχονται στο φως στοιχεία και αρχίζουμε να αμφιβάλλουμε για την αθωότητα του κατηγορουμένου. Μάλιστα τα στοιχεία αυτά τοποθετούνται στον «άξονα της ενοχής», ο οποίος είναι ο οριζόντιος άξονας που βλέπουμε κάτω από τις δύο κατανομές στο Σχήμα 25. Όσο πιο πολλά στοιχεία έχουμε τόσο μετακινούμαστε νοερά από τα αριστερά του οριζόντιου άξονα προς τα δεξιά..

Αν τα επιβαρυντικά στοιχεία που συσσωρεύονται περάσουν ένα νοητό κάθετο όριο που στο Σχήμα 25 το έχουμε ονομάσει «ετυμηγορία» τότε ψηφίζουμε ότι ο κατηγορούμενος δεν ανήκει στην αριστερή κατανομή (αυτή της «αθωότητας») αλλά αντίθετα ανήκει στη δεξιά κατανομή (αυτή της «ενοχής») και ψηφίζουμε «ένοχος». Αν τα επιβαρυντικά στοιχεία δεν περάσουν το κρίσιμο αυτό σημείο τότε δεν είμαστε σίγουροι ότι ο κατηγορούμενος διέπραξε το αδίκημα και βάσει του τεκμηρίου της αθωότητας ψηφίζουμε «αθώος». Εδώ υπάρχουν δύο πιθανά λάθη:



Σχήμα 1. Η ισχύς ενός στατιστικού τεστ είναι η πιθανότητα διάψευσης μιας ψευδούς μηδενικής υπόθεσης (1-β).

(α) να καταδικαστεί ένας αθώος (πολύ σοβαρή δικαστική πλάνη διότι κάποιος αθώος πάει στη φυλακή)

(β) να μην καταδικαστεί ένας ένοχος (σφάλμα εξίσου σοβαρό αν και μερικοί το θεωρούν λιγότερο σοβαρό από το προηγούμενο).

Στο Σχήμα 25 βλέπουμε την πιθανότητα β, δηλαδή να μην καταδικαστεί ένας άνθρωπος που είναι ένοχος και ανήκει στη δεξιά κατανομή. Να αποτύχουμε δηλαδή να απορρίψουμε μια ψευδή μηδενική υπόθεση. Αυτό είναι το «Σφάλμα Τύπου βήτα» στη Στατιστική.

Βλέπουμε επίσης στο Σχήμα 25 και την πιθανότητα 1-β που είναι να πετύχουμε να βρούμε ότι ένας άνθρωπος που ανήκει στη δεξιά κατανομή (ένοχος) καταδικάζεται ως ένοχος. Αυτό στη Στατιστική ονομάζεται «ισχύς» ενός τεστ. Ισχύς είναι η πιθανότητα να απορρίψουμε μια ψευδή μηδενική υπόθεση.

Τέλος, στο Σχήμα 25 βλέπουμε και την πιθανότητα α, δηλαδή την πιθανότητα να κάνουμε αυτό που στη Στατιστική ονομάζεται «Σφάλμα Τύπου άλφα»· να καταδικάσουμε έναν αθώο· να απορρίψουμε μια μηδενική υπόθεση που στην

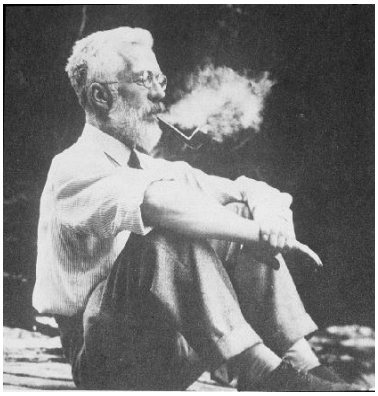
[1] Βλ Άρθρο 48 στον *Χάρτη Θεμελιωδών Δικαιωμάτων της Ευρωπαϊκής Ενωσης* (2000).

πραγματικότητα είναι αληθής. Από τι εξαρτάται η αθώωση ή η καταδίκη του κατηγορουμένου; Από το αν είναι αθώος ή ένοχος (δηλαδή από την αλήθεια της μηδενικής υπόθεσης), από την ισχύ των αποδεικτικών στοιχείων (την ισχύ του τεστ), καθώς και από το σημείο που με βάση τα αποδεικτικά στοιχεία θα δώσουμε την ετυμηγορία «ένοχος» (δηλαδή το άλφα). Το άλφα είναι το επίπεδο στατιστικής σημαντικότητας· η πιθανότητα να κάνουμε σφάλμα τύπου άλφα. Η ισχύς των αποδεικτικών στοιχείων σε μία δίκη θα μπορούσε να παραλληλιστεί με την ισχύ των στατιστικών τεστ σε μια στατιστική ανάλυση. Αν η πιθανότητα για λάθος τύπου βήτα είναι β, η ισχύς είναι η συμπληρωματική πιθανότητα του βήτα, δηλαδή 1-β.

Η έννοια της στατιστικής σημαντικότητας πρωτοεμφανίστηκε στον χώρο των κοινωνικών επιστημών μετά από το 1925 με τη δημοσίευση δηλαδή του βιβλίου του Ronald Fisher *Statistical Methods for Research Workers*. Δέκα χρόνια αργότερα, το 1935, ο ίδιος ο στατιστικός στο βιβλίο του *The Design of Experiments* (1935) παρουσίασε τα πρώτα τεστ για τον έλεγχο των στατιστικών υποθέσεων. Κατά τον έλεγχο των υποθέσεων η πιθανότητα p της τιμής του στατιστικού τεστ συγκρίνεται με μια θεωρητική τιμή, η οποία έχει πιθανότητα α. Για τον Fisher, η τιμή της πιθανότητας p είναι το μέτρο εναντίον της μηδενικής υπόθεσης και δεν υπάρχει κάποιο προαποφασισμένο όριο πέρα από το οποίο αυτή η τιμή γίνεται «σημαντική». Η πιθανότητα p=0,045, για παράδειγμα, έχει για τον Fisher όση αποδεικτική δύναμη έχει περίπου η πιθανότητα p=0,051. Για τους Jerzy Neyman και Egon Pearson, όμως, σφοδρούς επιστημονικούς αντίζηλους του Fisher, η τιμή α δηλαδή το λεγόμενο «επίπεδο στατιστικής σημαντικότητας», ορίζεται ως μια πιθανότητα-όριο του να απορρίψουμε τη μηδενική υπόθεση με την προϋπόθεση ότι στην πραγματικότητα η μηδενική υπόθεση είναι αληθής (Barros, Peypoch, & Louçã, 2008). Το α δηλαδή είναι η πιθανότητα να υποπέσουμε σε «Σφάλμα Τύπου α», όπως λέγεται και να απορρίψουμε μια στην πραγματικότητα αληθή μηδενική υπόθεση. Όμως η απόρριψη μιας μηδενικής υπόθεσης με βάση κάποιο δεδομένο α θα είχε νόημα αν και μόνο αν είχε νόημα η εναλλακτική υπόθεση. Σε ένα πείραμα για παράδειγμα. Όταν απλώς ελέγχουμε αν μια κατανομή είναι κανονική ή όχι τότε η απόρριψη της μηδενικής υπόθεσης για p=0,049 δεν έχει και πολύ νόημα.

Στο χώρο της εκπαίδευσης και της ψυχολογίας το χρησιμοποιούμενο επίπεδο στατιστικής σημαντικότητας είναι συνήθως το $\alpha = 0,05$, ενώ συχνή είναι στα άρθρα έγκυρων περιοδικών και οι πίνακες στα κελιά των οποίων και υπάρχει ένας αστερίσκος για στατιστική σημαντικότητα μικρότερη του 0,05, δύο αστερίσκοι για στατιστική σημαντικότητα μικρότερη του 0,01 και τρεις αστερίσκοι για στατιστική σημαντικότητα μικρότερη του 0,001. Παλαιότερα κυριαρχούσε η τάση να διατηρούμε ένα και μόνο επίπεδο στατιστικής σημαντικότητας σε όλη την έκταση της δημοσιευμένης εργασίας (π.χ. ό,τι συνδεόταν $p<0,05$ ήταν στατιστικώς σημαντικό). Σήμερα επικρατεί η τάση να δημοσιεύουμε την ακριβή τιμή του p για το αποτέλεσμα κάθε στατιστικού τεστ. Αυτό οδηγεί πολλούς ερευνητές να αξιολογούν μια σχέση που συνδέεται με $p<0,05$ ως «λιγότερο σημαντική» από μια σχέση που συνδέεται με $p<0,001$. Αυτό, όπως υποστηρίχθηκε πιο πάνω, δεν είναι αυτό το οποίο υποστήριξαν οι θεμελιωτές της στατιστικής ανάλυσης στον χώρο των κοινωνικών επιστημών.

Η τιμή του $\alpha = 0,05$ αν και δεν είναι αυθαίρετη είναι σίγουρα συμβατική, αφού η υιοθέτηση ενός μεγαλύτερου επίπεδου στατιστικής σημαντικότητας (π.χ. $\alpha = 0,10$), θα οδηγούσε αναπόφευκτα σε περισσότερα λάθη τύπου Ι. Από την άλλη μεριά, όμως, όσο περισσότερο μεγαλώνει το α, τόσο μειώνεται η πιθανότητα για σφάλμα τύπου ΙΙ, το οποίο αναφέρεται στην αποτυχία του να απορρίψουμε μια ψευδή μηδενική υπόθεση. Γεγονός είναι ότι η πιθανότητα β να απορρίψουμε μια ψευδή μηδενική υπόθεση *δεν* ισούται με 1-α, ενώ ένας ερευνητής που θα υιοθετούσε ως επίπεδο στατιστικής σημαντικότητας μεγαλύτερο του 0,05 (π.χ. $\alpha = 0,10$) θα είχε πολλές πιθανότητες να δει ότι η εργασία του δεν δημοσιεύεται σε κάποιο έγκυρο επιστημονικό περιοδικό. Είναι γνωστό επίσης ότι στο χώρο της ψυχολογικής και εκπαιδευτικής έρευνας έχει επικρατήσει η άποψη ότι το σφάλμα τύπου άλφα είναι σοβαρότερο από το σφάλμα τύπου βήτα. Πιο εύκολα, δηλαδή, δεχόμαστε να αποτύχουμε να απορρίψουμε μια ψευδή μηδενική υπόθεση παρά το αντίστροφο. Έτσι όπως έχουν τα πράγματα, συνηθίζουμε ως ερευνητές να καταλήγουμε σε ένα ερευνητικό συμπέρασμα για $p<\alpha$ και στο ακριβώς αντίθετο ερευνητικό συμπέρασμα για $p>\alpha$. Βλέπουμε δηλαδή πώς η τιμή του α οδηγεί πολλές φορές σε συγκρουόμενα ερευνητικά συμπεράσματα, χωρίς να έχουμε στα χέρια μας κάτι χειροπιαστό για να καταλάβουμε πώς σχετίζονται οι μεταβλητές μεταξύ τους.



Εικόνα 1. Ο sir Ronald Fisher είχε διαμάχη με τον Karl Pearson σε σχέση με τη φαινομενολογία της στατιστικής σημαντικότητας. Σήμερα στα εγχειρίδια μεθοδολογίας έρευνας υιοθετείται μια υβριδική λύση. Το λογισμικό δίνει την ακριβή πιθανότητα *p* (σύμφωνα με τον Fisher) και κατόπιν οι ερευνητές φέρονται στη μηδενική υπόθεση με βάση την κρίσιμη τιμή του άλφα (σύμφωνα με τον Pearson).

Η διάψευση στατιστικών υποθέσεων συνδέεται με διάφορα προβλήματα, όπως είναι, για παράδειγμα, η ισχύς των χρησιμοποιούμενων κριτηρίων ή το μέγεθος του δείγματος. Ισχύς ενός κριτηρίου είναι η πιθανότητα να απορρίψει μια ψευδή μηδενική υπόθεση για χάρη της εναλλακτικής. Όμοια συνδεδεμένο με τη διάψευση των μηδενικών υποθέσεων είναι και το μέγεθος του δείγματος. Για παράδειγμα αν ένας ερευνητής έχει στοιχεία για την επίδοση διακοσίων χιλιάδων μαθητών και συγκρίνει την επίδοση δώδεκα χιλιάδων από αυτούς (ας υποθέσουμε με μέσο όρο 100 και τυπική απόκλιση 15) με την επίδοση των υπολοίπων εκατόν ογδόντα οκτώ χιλιάδων μαθητών (ας υποθέσουμε με μέσο όρο 99,85 και τυπική απόκλιση 15) θα βρει ότι η διαφορά των τριών δέκατων του βαθμού συνδέεται με z τιμή ίση με 2,12 και p = 0,017. Η διαφορά δηλαδή είναι στατιστικώς σημαντική για επίπεδο στατιστικής σημαντικότητας $\alpha = 0,05$. Η μηδενική υπόθεση μπορεί να απορρίπτεται στατιστικά, αλλά ουσιαστικά η εναλλακτική της είναι ασήμαντη.

Συμπερασματικά, θα λέγαμε ότι το p σε ένα στατιστικό τεστ εκτιμάει απλώς την πιθανότητα, ώστε η τιμή που βρίσκουμε στο δείγμα μας να αποκλίνει τόσο όσο

φαίνεται ότι αποκλίνει από την τιμή που ορίζεται για τον πληθυσμό από τη μηδενική υπόθεση. Τα στατιστικά τεστ υποθέτουν ότι στον πληθυσμό ισχύει η μηδενική υπόθεση και, στο πλαίσιο αυτό, ελέγχουν τη πιθανότητα να πάρουμε τις συγκεκριμένες τιμές που φανερώνει το δείγμα. Έτσι, ακόμα και επουσιώδεις διαφορές ή ασήμαντες πραγματικές σχέσεις είναι δυνατόν να δώσουν στατιστικώς σημαντικά αποτελέσματα. Η απόρριψη ή η αποτυχία απόρριψης της μηδενικής υπόθεσης δεν μας λέει όλα όσα θα θέλαμε, σχετικά με την ένταση της σχέσης που συνδέει δύο μεταβλητές. Αυτό συμβαίνει γιατί η μηδενική υπόθεση μπορεί να διαψευστεί όχι μόνο από την ύπαρξη κάποιας «ουσιαστικής» σχέσης μεταξύ μεταβλητών αλλά και από άλλους παράγοντες.

Η πραγματική σημαντικότητα

Από μόνη της, όμως, η στατιστική σημαντικότητα δεν είναι αρκετή για να φτάσουμε σε ερευνητικά συμπεράσματα σημαντικά από πρακτική άποψη. Χρειαζόμαστε μια άλλου είδους σημαντικότητα, η οποία θα σηματοδοτεί την «ουσιαστική» πλευρά του αποτελέσματος, αυτό δηλαδή που ο Alan Kazdin (1999) από το Πανεπιπστήμιο του Yale ονομάζει «πρακτική σημαντικότητα» (“practical significance”). Οι σχετικές συζητήσεις σχετικά με την πρακτική σημαντικότητα των ευρημάτων της εκπαιδευτικής έρευνας υπήρξαν τόσο έντονες στο παρελθόν, ώστε η American Psychological Association (APA) που είδαμε στο **Σφάλμα! Το αρχείο προέλευσης της αναφοράς δεν βρέθηκε.** σύστησε το 1996 ειδική Ομάδα Εργασίας (Task Force) για τη μελέτη του όλου ζητήματος. Δεχόμενη εν μέρει τις προτάσεις της Ομάδας Εργασίας (Wilkinson, 1999), η APA έδωσε οδηγίες προς τους συγγραφείς ότι είναι σχεδόν πάντα αναγκαίο να δίνουμε πληροφορίες όχι μόνο για τη στατιστική σημαντικότητα των ευρημάτων μας αλλά και πληροφορίες σχετικά με την ένταση των παρατηρούμενων σχέσεων.

Ο όρος Effect Size, δηλαδή «μέγεθος της επίδρασης» παραπέμπει μάλλον στα πειράματα, στα οποία κατά κύριο λόγο βλέπουμε «επιδράσεις» ανεξάρτητων μεταβλητών σε εξαρτημένες μεταβλητές. Με πείραμα μελετάμε, για παράδειγμα, το «αποτέλεσμα» μιας εκπαιδευτικής παρέμβασης, μέσω της «επίδρασης» που αυτή έχει σε κάποια ή σε κάποιες εξαρτημένες μεταβλητές. Η λέξη «επίδραση» όμως στη φράση «μέγεθος της επίδρασης» μπορεί να αναφέρεται όχι μόνο στην ύπαρξη αιτιώδους σχέσης και επίδρασης μεταξύ δύο μεταβλητών Χ και Υ αλλά ακόμα και στην απλή συμμεταβολή τους. Πάντως είτε στη μία είτε στην άλλη περίπτωση μιλάμε για «σχέσεις» μεταξύ των μεταβλητών και όχι αναγκαστικά για «επιδράσεις» και αυτό πρέπει να ξεκαθαριστεί από στατιστικη άποψη.

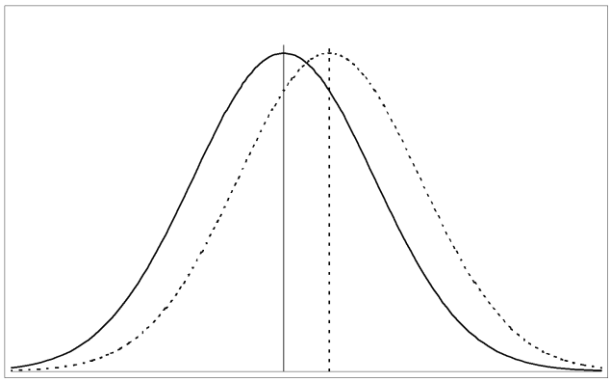
Στη ψυχολογική και εκπαιδευτική έρευνα οι σχέσεις επίδρασης και οι αλληλεπιδράσεις εμφανίζονται είτε ως διαφορές μεταξύ μέσων όρων είτε ως συντελεστές συσχέτισης και συνάφειας. Και στη μία και στην άλλη περίπτωση οι

εν λόγω σχέσεις βασίζονται στην έννοια της διακύμανσης. Είναι γνωστό, για παράδειγμα, ότι υψώνοντας τον δείκτη r του Pearson στο τετράγωνο, μπορούμε να πάρουμε ένα μέτρο του πώς «επιδρά» μία μεταβλητή στη διακύμανση των τιμών της άλλης. Είναι γνωστό επίσης ότι ο ίδιος δείκτης r —ουσιαστικά ως point biserial (ρ_{bs})— είναι ο συντελεστής βαρύτητας στη περίπτωση ενός γραμμικού μοντέλου που δημιουργούμε για να συσχετίσουμε μια κατηγορική μεταβλητή που έχει μόνο δύο κατηγορίες (π.χ. φύλο) με μια ποσοτική μεταβλητή. Αν λάβουμε δε υπόψη ότι ανεξάρτητα από τη στατιστική σημαντικότητα, τόσο οι διαφορές μεταξύ στατιστικών στοιχείων, όσο και οι σχέσεις μεταξύ μεταβλητών συνδέονται με τη διακύμανση, καταλήγουμε στο συμπέρασμα ότι η μελέτη του μεγέθους της επίδρασης συνδέεται με τη μελέτη των πηγών της διακύμανσης. Αυτοί είναι λοιπόν και οι τρεις άξονες πάνω στους οποίους μελετάμε το μέγεθος της επίδρασης: (α) διαφορές μέσων όρων, (β) οι συσχετίσεις και συνάφειες και τελικά (γ) η μελέτη της διακύμανσης.

Το μέγεθος της επίδρασης ως διαφορά μέσων όρων

Πολλές φορές στη εκπαιδευτική έρευνα χρειάζεται να αξιολογήσουμε τη διαφορά δύο μέσων όρων. Στη περίπτωση αυτή, όπως είναι γνωστό, χρησιμοποιούμε το t test, ένα παραμετρικό κριτήριο, το οποίο εξετάζει τη πιθανότητα της τιμής της «τυπικής» (standardised) διαφοράς των δύο μέσων. Για σχετικά μικρά δείγματα, η τιμή του t test συγκρίνεται με τις κρίσιμες τιμές της κατανομής του "Student", για δεδομένους βαθμούς ελευθερίας και όχι με τις τιμές της τυπικής κανονικής κατανομής (δηλαδή των τιμών z). Ο λόγος για τον οποίο γίνεται αυτό είναι ότι δεν γνωρίζουμε τη διακύμανση της δειγματοληπτικής κατανομής και πρέπει να την υπολογίσουμε από τη διακύμανση των τιμών του δείγματος. Κι ενώ, σύμφωνα με τη θεωρία, η δειγματοληπτική κατανομή της διαφοράς δύο μέσων όρων είναι η κανονική κατανομή, η δειγματοληπτική κατανομή της διακύμανσης ακολουθεί τη κατανομή χι τετράγωνο για $n-1$ βαθμούς ελευθερίας (όπου n το μέγεθος του δείγματος). Επίσης, όταν μελετάμε τη διαφορά μεταξύ μέσων όρων, συμβαίνει καμιά φορά να μην είμαστε σε θέση να αξιολογήσουμε την ουσία μιας κατά τα άλλα στατιστικώς σημαντικής διαφοράς διότι δεν ήμαστε εξοικειωμένοι με τη κλίμακα μέτρησης. Καταλαβαίνουμε, για παράδειγμα, μια στατιστικώς σημαντική διαφορά μεταξύ του 19 και του 15, όταν οι αριθμοί αυτοί αντιπροσωπεύουν μέσους όρους μαθητικής επίδοσης μαθητών λυκείου, αλλά δεν καταλαβαίνουμε, ενδεχομένως, μια διαφορά πενήντα μονάδων σε μια παγκόσμια μελέτη μαθητικής επίδοσης στα οποία η κλίμακα είναι από το 0 μέχρι το 700.

Ένας τρόπος λοιπόν να αξιολογήσουμε τη διαφορά των μέσων όρων είναι να μελετήσουμε το ποσοστό στο οποίο οι δύο σχετικές κατανομές αλληλεπικαλύπτονται. Ας υποθέσουμε, για παράδειγμα, ότι έχουμε ένα πείραμα με δύο ομάδες: την πειραματική (experimental) και την ομάδα ελέγχου (control).



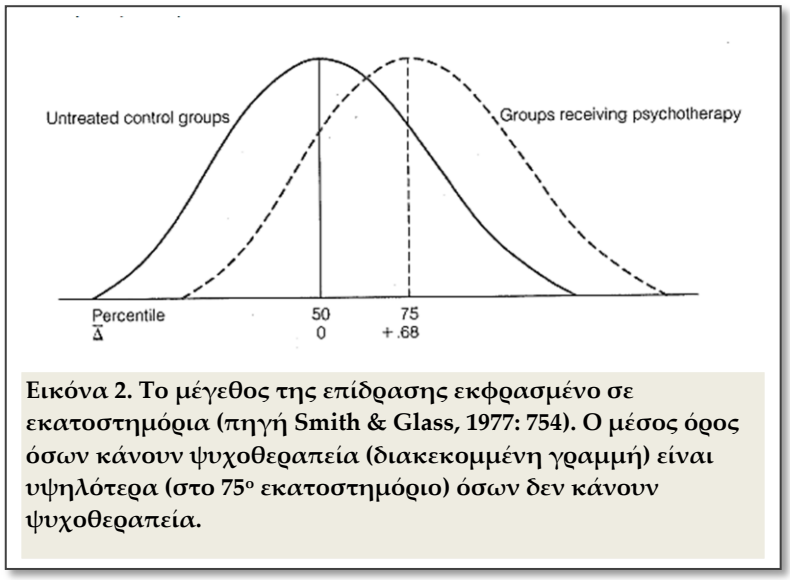
Σχήμα 2. Πειραματική Ομάδα (διακεκομμένη γραμμή), Ομάδα Ελέγχου (συνεχής γραμμή) και οι κατανομές τους.

Ας υποθέσουμε ότι οι κατανομές συχνοτήτων των δύο ομάδων είναι κανονικές, με μέσους όρους μ^C και μ^E αντιστοίχως και κοινή διακύμανση σ^2 . Λέμε τότε ότι το μέγεθος της επίδρασης, το οποίο συμβολίζουμε με δ , ισούται με τη διαφορά $\mu^E - \mu^C$ διά την τυπική απόκλιση σ .

Στο που ακολουθούν, βλέπουμε τις κατανομές συχνοτήτων για την πειραματική ομάδα (διακεκομμένη γραμμή) και την ομάδα ελέγχου (συνεχής γραμμή). Οι κάθετες γραμμές αντιστοιχούν στους μέσους όρους των κατανομών. Κι ενώ η απόσταση που χωρίζει τους μέσους όρους είναι ίδια και στα δύο σχήματα, στο Σχήμα α η διακύμανση είναι μικρότερη. Χωρίς να αναφερθούμε καθόλου ούτε στη στατιστική σημαντικότητα των διαφορών ούτε στη κλίμακα μέτρησης, μπορούμε, μέσω του ποσοστού αλληλοεπικάλυψης των δύο κατανομών να αξιολογήσουμε το μέγεθος της διαφοράς μεταξύ της πειραματικής ομάδας και της ομάδας ελέγχου. Βέβαια, χρειάζεται να έχουμε κανονικές κατανομές. Συμπερασματικά, όταν έχουμε για την Ομάδα Ελέγχου ότι $Y_j^C \sim N(\mu^C, \sigma^2)$ με $j = 1, \cdots, n^C$ και για τη Πειραματική Ομάδα ότι $Y_j^E \sim N(\mu^E, \sigma^2)$ με $j = 1, \cdots, n^E$ το μέγεθος της επίδρασης (Effect Size) είναι $\delta = \frac{\mu^E - \mu^C}{\sigma}$. Ως παράδειγμα μπορούμε να αναφέρουμε ότι ένας ερευνητής πήρε 38 παιδιά πρωτοβάθμιας εκπαίδευσης, εκ των οποίων τα μισά επιλέγηκαν τυχαία για να ακούσουν μια ιστορία από την νηπιαγωγό κατά τις πρωινές ώρες. Τα υπόλοιπα 19 παιδιά άκουσαν την ίδια ιστορία το απόγευμα από κασετόφωνο. Τα παιδιά και των δύο ομάδων κλήθηκαν να απαντήσουν αργότερα σε 20 ερωτήσεις σχετικά μ' αυτά που άκουσαν. Ο μέσος όρος των σωστών απαντήσεων ήταν για το πρωί 15,2 και για το απόγευμα 17,9. Η τυπική απόκλιση και στις δύο περιπτώσεις ήταν 3,3. Αντικαθιστώντας στον τύπο έχουμε ότι ένα μέγεθος επίδρασης ίσο με 0,8 γιατί $\delta = \frac{17,9 - 15,2}{3,3} = 0,8$.

Τρόποι παρουσίασης της διαφοράς μεταξύ δύο μέσων όρων

Το μέγεθος της επίδρασης στη διαφορά μεταξύ δύο μέσων όρων ισοδυναμεί κατ' αρχάς με κάποια τιμή z, δηλαδή με τιμή της τυπικής κανονικής κατανομής. Έτσι, στο παράδειγμα το οποίο αναφέραμε και με βάση τις περιοχές της τυπικής κανονικής κατανομής για $z = 0{,}8$, ο μέσος όρος των τιμών της απογευματινής ομάδας είναι μεγαλύτερος από το 79% των τιμών της πρωινής ομάδας. Εκτός από αυτό και αφού γνωρίζουμε ότι στο προαναφερθέν πείραμα έχουμε δύο ομάδες παιδιών των 19 ατόμων η καθεμία, μπορούμε να ισχυριστούμε π.χ. ότι το 10ο παιδί της πειραματικής ομάδας, στο οποίο αντιστοιχεί ο μέσος όρος επίδοσης του απογευματινού γκρουπ, θα έχει επίδοση ίση με εκείνη του παιδιού με την 4η υψηλότερη επίδοση στην ομάδα ελέγχου (πρωινό γκρουπ). Έτσι, έχουμε μια χειροπιαστή ένδειξη για το μέγεθος της επίδρασης.



Εικόνα 2. Το μέγεθος της επίδρασης εκφρασμένο σε εκατοστημόρια (πηγή Smith & Glass, 1977: 754). Ο μέσος όρος όσων κάνουν ψυχοθεραπεία (διακεκομμένη γραμμή) είναι υψηλότερα (στο 75ο εκατοστημόριο) όσων δεν κάνουν ψυχοθεραπεία.

Ένας άλλος τρόπος να παρουσιάσουμε το μέγεθος της διαφοράς των μέσων όρων σε μια εργασία είναι να μετατρέψουμε την τιμή του δ σε συντελεστή συσχέτισης ρ του Pearson και να εκτιμήσουμε πόσο τοις εκατό της διακύμανσης της συνεχούς μεταβλητής «οφείλεται» στη παρέμβασή μας. Πράγματι, αν λάβουμε υπόψη μας την βασική ιδέα της Ανάλυσης Διακύμανσης, αποδεικνύεται ότι $\rho^2 \approx \frac{\delta^2}{\delta^2 + df}$, όπου *df* είναι οι βαθμοί ελευθερίας του *t test*, δηλαδή $n^E + n^C - 2$. Οι Larry Hedges και Ingram Olkin (1985) προτείνουν την ίδια σχέση με ελαφρώς διαφορετικό παρονομαστή: $\rho^2 \approx \frac{\delta^2}{(\delta^2 + df)/\tilde{n}}$, όπου $\tilde{n} = n^E n^C / (n^E + n^C)$. Για το παράδειγμα που παρουσιάσαμε προηγουμένως

έχουμε με αντικατάσταση ότι $\rho^2 \approx \frac{0,8^2}{0,8^2+4}$ και τελικά $\rho^2 = 0,138$. Περίπου, δηλαδή, το 14% της διακύμανσης στην επίδοση των παιδιών οφείλεται στην επίδραση της ανεξάρτητης μεταβλητής.

Αξιολογώντας τη τιμή του μεγέθους της επίδρασης

Οι πιο γνωστές απόψεις για την αξιολόγηση του μεγέθους της επίδρασης στη περίπτωση της διαφοράς μέσων όρων είναι αυτές του Cohen (1969), ο οποίος πρότεινε τις κατ' απόλυτη τιμή 0,2, 0,5, και 0,8 τυπικές αποκλίσεις πάνω ή κάτω από τον μέσο όρο των κατανομών ως «μικρό», «μέτριο», και «μεγάλο» «αποτέλεσμα», αναφορικά με τις τυπικές διαφορές των μέσων όρων (standardized mean differences), ενώ τα ποσοστά 1, 6, και 14 τοις εκατό θεωρούνται από τον ίδιο ως «μικρό», μέτριο, και μεγάλο αποτέλεσμα. Οι απόψεις του Cohen (ό. π.) για τα μεγέθη των αποτελεσμάτων της στατιστικής ανάλυσης, έχουν επηρεάσει τον σχολιασμό των ευρημάτων στα κείμενα των έγκυρων διεθνών περιοδικών ψυχοπαιδαγωγικής έρευνας. Τα αριθμητικά μεγέθη που ο Cohen συνδέει με το μέγεθος της επίδρασης –αν και δεν είναι τελείως αυθαίρετα-— δεν έχουν ακόμα πλήρη εμπειρική στήριξη στο χώρο της ψυχοπαιδαγωγικής έρευνας. Παρόλα αυτά δεν υπάρχουν μέχρι στιγμής δημοσιευμένα εναλλακτικά αριθμητικά μεγέθη σχετικά με το μέγεθος της επίδρασης και οι περισσότεροι ερευνητές βασίζονται σε αυτά.

Ένα παράδειγμα αξιολόγησης του μεγέθους της επίδρασης από τη βιβλιογραφία.

Οι συγγραφείς Χριστίνα Τσιλφίδου και Μαρία Πλατσίδου (2011), αν και δεν το δηλώνουν, χρησιμοποιούν και αξιολογούν ως δείκτη μεγέθους της επίδρασης την τιμή 0,79 του δείκτη alpha του Cronbach, στον οποίο θα αναφερθούμε κι εμείς σε επόμενο κεφάλαιο. Γράφουν οι ερευνήτριες:

«Οι συμμετέχοντες κλήθηκαν να δηλώσουν τον βαθμό στον οποίο ίσχυε γι' αυτούς η καθεμία από τις προτάσεις χρησιμοποιώντας μια κλίμακα 5 σημείων όπου 1=Ποτέ και 5=Σχεδόν πάντα. Το ερωτηματολόγιο αυτό ελέγχθηκε ως προς την εσωτερική συνοχή του και διαπιστώθηκε ότι η αξιοπιστία του ήταν αρκετά υψηλή (Cronbach = 0,79).» (Τσιλφίδου & Πλατσίδου, 2011: 179).

Μια λύση στο πιο πάνω ζήτημα έχουν προσπαθήσει να δώσουν οι Rosenthal και Rubin (1982, στο Randolph, 2005), οι οποίοι αξιολογούν το μέγεθος της επίδρασης στη περίπτωση δύο μέσων όρων με τον δείκτη BESD (Binomial Effect Size Display). Οι δύο στατιστικολόγοι, προτείνουν τη «διχοτόμηση» της εξαρτημένης

(ποσοτικής) μεταβλητής -η ανεξάρτητη είναι εκ των πραγμάτων διχοτομημένη- έτσι ώστε να υπολογίσουν τα ποσοστά πάνω και κάτω από το σημείο της διχοτόμησης, για κάθε κατηγορία της ανεξάρτητης ξεχωριστά. Για παράδειγμα, στην έρευνα με τα παιδιά του νηπιαγωγείου που αναφέρεται στη προηγούμενη σελίδα , θα μπορούσαμε να διχοτομήσουμε την εξαρτημένη μεταβλητή με βάση, ας πούμε, τη διάμεσο όλων των τιμών και κατόπιν να ορίσουμε τις κατηγορίες «επιτυχία» και «αποτυχία», ανάλογα με το αν ένα παιδί απάντησε αντιστοίχως σε περισσότερες ή λιγότερες ερωτήσεις από τη τιμή της διαμέσου. Τα ποσοστά των τιμών πάνω και κάτω από το σημείο της «αποτυχίας-επιτυχίας» για την Πειραματική Ομάδα και την Ομάδα Ελέγχου ξεχωριστά, συνδέονται με τον συντελεστή ρ του Pearson και αποτελούν, σύμφωνα με τους Rosenthal και Rubin (1982) έναν πολύ καλό δείκτη για το μέγεθος της επίδρασης. Πράγματι, αν υποθέσουμε ότι στην έρευνα που αναφέραμε η πιθανότητα επιτυχίας είναι για τον πληθυσμό 0,5, τότε για ένα effect size ίσο με 0,8, η πιθανότητα επιτυχίας για τους μαθητές της Ομάδας Ελέγχου είναι 0,5 + ρ, ενώ για την Πειραματική Ομάδα είναι 0,5 – ρ. Έτσι, η πιθανότητα επιτυχίας για την απογευματινή ομάδα είναι 0,5 + 0,37 = 0,87 ή 87%, ενώ για την πρωινή ομάδα μόλις 13%. Βλέπουμε λοιπόν πώς ένας σχετικά μικρός συντελεστής ρ συνδέεται με μεγάλες διαφοροποιήσεις όταν έχουμε να κάνουμε με το δίπολο «επιτυχία – αποτυχία».

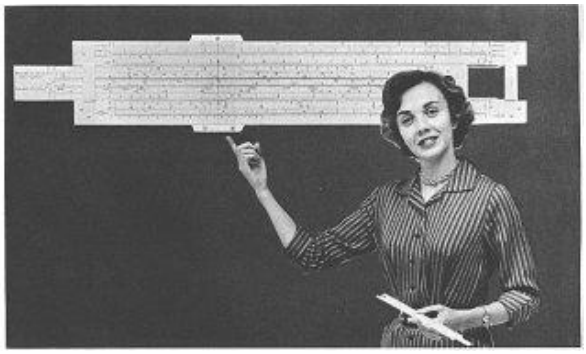
Το μέγεθος της επίδραση στον πληθυσμό

Υπάρχουν πολλοί τρόποι για να υπολογίσουμε το μέγεθος της επίδρασης δ στον πληθυσμό. Ένας από αυτούς είναι να χρησιμοποιήσουμε τη ποσότητα $g = (\bar{Y}^E - \bar{Y}^C)/s^*$ του δείγματος, όπου $\bar{Y}^E$ και $\bar{Y}^C$ είναι οι μέσοι όροι της πειραματικής ομάδας και της ομάδας ελέγχου αντιστοίχως, ενώ s^* η τυπική απόκλιση, στη οποία θα αναφερθούμε πιο κάτω. Σημειώστε ότι, όπως και οι Hedd, χρησιμοποιούμε εδώ το σύμβολο g ως εκτιμητή του δ στον πληθυσμό και όχι τα ευρέως διαδεδομένα στη βιβλιογραφία d και Δ. Αυτό γίνεται γιατί η ποσότητα $(\bar{Y}^E - \bar{Y}^C)/s^*$ δεν είναι αμερόληπτη εκτιμήτρια του δ. Θα αναφερθούμε στο ζήτημα αυτό στην επόμενη παράγραφο. Όσον αφορά τώρα την τυπική απόκλιση s^* ο Glass (1976) έχει προτείνει τη χρήση της τυπικής απόκλισης της ομάδας ελέγχου s^C στον παρονομαστή, με το σκεπτικό ότι είναι πιο ασφαλές να χρησιμοποιούμε τη διακύμανση της ομάδας ελέγχου στα πειράματα, όπου έχουμε πολλές πειραματικές ομάδες αλλά μόνο *μία* ομάδα ελέγχου. Όμως, είναι γενικά παραδεκτό στη βιβλιογραφία -ακόμα και στον υπολογισμό του *t test*- ότι χρησιμότερο είναι να χρησιμοποιούμε τη συνδυασμένη

(*pooled*) τυπική απόκλιση για τον υπολογισμό του μεγέθους της επίδρασης. Η συνδυασμένη διακύμανση

δίδεται από τον τύπο $s_{pooled} = \sqrt{\frac{(n^E - 1)(s^E)^2 + (n^C - 1)(s^C)^2}{n^E + n^C - 2}}$.

Ας αναφέρουμε ένα παράδειγμα. Οι Λεονταρή, Κυρίδης, και Γιαλαμάς (2000) μελέτησαν το επαγγελματικό άγχος 447 εκπαιδευτικών της πρωτοβάθμιας εκπαίδευσης της χώρας μας, χρησιμοποιώντας την κλίμακα Άγχους της Επαγγελματικής Ζωής (Professional Life Stress Scale, PLSS, των Fontana & Abouserie, 1993). Βρέθηκε στην εν λόγω έρευνα ότι το επαγγελματικό άγχος των 47 εκπαιδευτικών που εργάζονται σε αγροτικές περιοχές είχε μέσο όρο 20,53 και τυπική απόκλιση 9,26. Ο αντίστοιχος μέσος όρος και τυπική απόκλιση για τους 327 εκπαιδευτικούς της ομάδας ελέγχου ήταν 17,46 και 7,67. Έτσι, μια εκτίμηση για το μέγεθος της επίδρασης, με τη χρήση της συνδυασμένης τυπικής απόκλισης θα ήταν:



Εικόνα 3. Η μέτρηση. Πριν την εμφάνιση των υπολογιστών τσέπης στα μέσα της δεκαετίας του 1970 οι μαθηματικοί υπολογισμοί στη γη αλλά και στο διάστημα γίνονταν μέσω ξύλινων ή μεταλλικών κανόνων, οι οποίοι με έναν αριθμό συρόμενων ράβδων και ενός κινητού κέρσορα μετέτρεπαν ρητούς αριθμούς στη λογαριθμική κλίμακα και έδιναν αποτελέσματα αριθμητικών δύσκολων πράξεων. Από σημειωτική άποψη ο χάρακας ήταν η δια του συμβόλου δήλωση του επαγγέλματος του μηχανικού όπως το στηθοσκόπιο ήταν η δια ενός συμβόλου δήλωση του επαγγέλματος του γιατρού.

$$g = (\bar{Y}^E - \bar{Y}^C) / s_{pooled} = (20,53 - 17,46) \Big/ \sqrt{\frac{(47-1)(9,26)^2 + (327-1)(7,67)^2}{47 + 327 - 2}} = 0,389 .$$

Αυτή είναι μια εκτίμηση για το μέγεθος της επίδρασης στον πληθυσμό.

Αναφέραμε προηγουμένως, ότι η ποσότητα *g* δεν είναι αμερόληπτη εκτιμήτρια του μεγέθους της επίδρασης στον πληθυσμό. Πράγματι, σύμφωνα με τους Hedges και Olkin (1985) η εκτιμώμενη τιμή του *g* δεν είναι ίση με *δ*, αλλά περίπου $\delta + \frac{3\delta}{4N - 9}$, όπου $N = n^E + n^C$. Η εκτιμώμενη τιμή του *g*, σύμφωνα με τους ίδιους (ό.

π.) είναι $E(g) = d / J(N-2)$, όπου $J_{(N-2)} = 1 - \frac{3}{4(N-2)-1}$ (υπάρχουν σχετικοί πίνακες για την τιμή του $J_{(m)}$ για m από 2 και πάνω, στους οποίους φαίνεται ότι όσο μεγαλύτερο είναι το μέγεθος του δείγματος, τόσο η τιμή του J πλησιάζει τη μονάδα). Πρακτικά, λοιπόν, ένας καλύτερος εκτιμητής του δ είναι το $d \cong \left(1 - \frac{3}{4N-9}\right) g$. Για το παράδειγμα των Λεονταρή, Κυρίδη και Γιαλαμά (2000) που αναφέραμε προηγουμένως το μέγεθος της επίδρασης είναι $d \cong \left(1 - \frac{3}{4(434)-9}\right) \times 0{,}389 = 0{,}388$, ελάχιστα μικρότερο της τιμής g που υπολογίσαμε προηγουμένως. Για σχετικά μεγάλα δείγματα, η διαφορά μεταξύ d και g είναι σχεδόν αμελητέα (L Hedges & Olkin, 1985). Είναι, όμως, για τους συγγραφείς ορθότερο να χρησιμοποιούμε το d και όχι το g. Η θεωρητική κατανομή του g σχετίζεται με την t κατανομή, αλλά παρεκκλίνει από τις τιμές της t με μια παράμετρο ίση με $\sqrt{\tilde{n}}\delta$, όπου $\tilde{n} = \frac{n^E n^C}{n^E + n^C}$. Από την άλλη μεριά, σύμφωνα με τους Hedges και Olkin (1985) η θεωρητική κατανομή του d (θυμηθείτε ότι $d = J_{(m)} g$) είναι η κανονική κατανομή με μέσο όρο δ και εκτιμώμενη διακύμανση για μεγάλα δείγματα $\sigma^2_\infty(d) = \frac{n^E + n^C}{n^E n^C} + \frac{\delta^2}{2\left(n^E + n^C\right)}$ (). Σύμφωνα με τους ίδιους, τια επαρκώς μεγάλα δείγματα (δηλαδή τόσο το n^C, όσο και το n^E να είναι τουλάχιστον μεγαλύτερα του 10), μπορούμε να χρησιμοποιήσουμε τον τύπο της διακύμανσης $\sigma^2_\infty(d)$, έχοντας για δ^2 την τιμή του d^2 (ό. π.). Η εκτιμώμενη διακύμανση $\hat{\sigma}^2(d)$ μπορεί να χρησιμοποιηθεί αργότερα με τον γνωστό τρόπο για να βρούμε και το *διάστημα εμπιστοσύνης* του δ, δηλαδή το κατώτερο και ανώτερο όριο του μεγέθους της επίδρασης στον πληθυσμό (ό. π.). Αυτή άλλωστε ήταν και η σύσταση της Ομάδας Εργασίας της APA (Wilkinson, 1999). Για επίπεδο στατιστικής σημαντικότητας α = 0,05 το διάστημα εμπιστοσύνης του δ στον πληθυσμό είναι d συν/πλην 1,96 φορές το $\hat{\sigma}^2(d)$. Για τη δική μας περίπτωση, αν κάνουμε τις πράξεις, βρίσκουμε ότι $\hat{\sigma}^2(d) = 0{,}0442$. Το διάστημα εμπιστοσύνης του δ στον πληθυσμό είναι (0,344, 0,432).

Η ανάλυση διακύμανσης ως γραμμικό μοντέλο

Στις προηγούμενες ενότητες είδαμε το μέγεθος της επίδρασης ως τυποποιημένη διαφορά μέσων όρων. Σε αυτή την ενότητα θα δούμε το μέγεθος της επίδρασης ως επεξηγούμενη διακύμανση (variance-accounted-for). Η βασική ιδέα στη παρούσα ενότητα είναι ότι σε μια ερευνητική παρέμβαση, η ανεξάρτητη ή «επεξηγούσα» μεταβλητή θα «εξηγήσει» μέρος της συνολικής διακύμανσης της εξαρτημένης ή «επεξηγούμενης» μεταβλητής. Έτσι, ο λόγος της «επεξηγημένης» διακύμανσης προς τη συνολική διακύμανση μπορεί να θεωρηθεί ως το μέγεθος της επίδραση της ανεξάρτητης μεταβλητής στην εξαρτημένη. Σε σχέση με τις τυποποιημένες διαφορές που έχουμε δει, η επεξηγούμενη διακύμανση μας βοηθάει να αξιολογήσουμε το μέγεθος της επίδρασης σε περισσότερο σύνθετους ερευνητικούς σχεδιασμούς. Όμως, η μέθοδος αυτή έχει το μειονέκτημα ότι αναφέρεται στη συνάφεια μεταξύ μεταβλητών και όχι ακριβώς στην επίδραση της μίας στην άλλη. Η διαφορά της έννοιας της συνάφειας από την έννοια της επίδρασης και πολύ περισσότερο την έννοια της αιτιώδους σχέσης είναι πολύ ουσιαστική, αλλά θεωρούμε ότι το ζήτημα αυτό είναι ξεκάθαρο. Έτσι, ενώ στις μεθόδους που στηρίζονται στην επεξηγούμενη διακύμανση μπορούμε να μιλάμε για «μέγεθος της επίδρασης», δεν είναι πάντα δυνατό να γνωρίζουμε αν αυτή η «επίδραση» είναι και «αιτιότητα» ή είναι απλώς το αποτέλεσμα κάποιας συμμεταβολής.

Υπάρχουν πολλοί τρόποι με τους οποίους μπορούμε να εκφράσουμε τον λόγο της επεξηγούμενης διακύμανσης προς τη συνολική διακύμανση. Αυτό μένει να το αποφασίσουμε, ανάλογα με τον ερευνητικό σχεδιασμό ή τον τύπο της στατιστικής ανάλυσης που χρησιμοποιούμε. Ειδικά στον τομέα της ψυχολογικής έρευνας η μελέτη του μεγέθους της επίδρασης στην Ανάλυση Διακύμανσης (ANOVA) μπορεί να είναι πολύ πολύπλοκο ζήτημα.

Τη βασική ιδέα της απλής Ανάλυσης Διακύμανσης μπορούμε να την εκφράσουμε μέσω ενός γραμμικού μοντέλου. Για παράδειγμα, αν έχουμε μια εξαρτημένη ποσοτική μεταβλητή Y, με m περιπτώσεις, και η ανεξάρτητη κατηγορική μεταβλητή περιλαμβάνει k κατηγορίες, γράφουμε για τη περίπτωση

i της κατηγορίας j ότι $Y_{ij} = \mu + \tau_j + e_{ij}$. Στη σχέση αυτή, μ είναι ο γενικός μέσος όρος, τ_j η διαφορά του γενικού μέσου όρου από τον μέσο όρο της κατηγορίας j, ενώ e_{ij} είναι ένας όρος «σφαλμάτων», τα οποία ορίζονται ως η διαφορά της τιμής Y_{ij} από τον μέσο όρο μ_j της οικείας κατηγορίας j. Μπορούμε, δηλαδή, να γράψουμε με μορφή πινάκων ότι **y** = **Xτ** + **e**, όπου **y** είναι ένα διάνυσμα m στοιχείων (όσο είναι το μέγεθος του δείγματος), **e** ένα διάνυσμα m «σφαλμάτων», **τ** ένα διάνυσμα με k+1 στοιχεία (το πλήθος k των ομάδων της παρέμβασης συν ένα επιπλέον στοιχείο, το οποίο προορίζεται για τον γενικό μέσο όρο), και **X** ένας πίνακας (design matrix), με m γραμμές και k+1 στήλες. Ο πίνακας **X** έχει ως σκοπό να συνδέσει την εξαρτημένη συνεχή μεταβλητή y με την ανεξάρτητη μεταβλητή. Είναι γνωστό κατ' αρχάς ότι σχέσεις μεταξύ μεταβλητών και διαφορές μεταξύ μέσων όρων εκφράζονται με τη βοήθεια γραμμικών μοντέλων.

Αν υπολογίσουμε τη σχέση αυτή για όλους τους συμμετέχοντες στην παρέμβαση και για όλες τις κατηγορίες παρέμβασης, καταλήγουμε στο ότι το συνολικό άθροισμα τετραγώνων των διαφορών των τιμών από τον γενικό μέσο όρο (SS_{total}) ισούται με το άθροισμα των τετραγώνων των διαφορών των μέσων όρων της παρέμβασης από τον γενικό μέσο όρο ($SS_{treatment}$), συν το άθροισμα τετραγώνων των διαφορών των τιμών από τους μέσους όρους των οικείων ομάδων (SS_{error}). Αν διαιρέσουμε τα αθροίσματα των τετραγώνων (SS) με τους οικείους βαθμούς ελευθερίας, έχουμε τα μέσα αθροίσματα τετραγώνων (mean squares, MS). Όπως είναι γνωστό, στην Ανάλυση Διακύμανσης (ANOVA) χρησιμοποιούμε το F *test*, το οποίο ακολουθεί την F κατανομή, για κάποιους βαθμούς ελευθερίας, και το οποίο αξιολογεί τη στατιστική σημαντικότητα του λόγου $F = \frac{MS_{treatment}}{MS_{error}}$.

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΓΙΑ ΔΥΟ ΚΑΤΗΓΟΡΙΕΣ

Σε προηγούμενη ενότητα αναφερθήκαμε στο πώς μπορούμε να μετατρέψουμε την τυποποιημένη διαφορά μεταξύ δύο μέσων όρων σε συντελεστή ρ του Pearson. Είδαμε επίσης πώς ο συντελεστής αυτός μπορεί να χρησιμοποιηθεί για την αξιολόγηση του μεγέθους της επίδρασης είτε ως έχει είτε -κυρίως- υψωμένος στο τετράγωνο. Αυτά ισχύουν στη περίπτωση που η επεξηγούσα (ή ανεξάρτητη)

κατηγορική μεταβλητή έχει δύο κατηγορίες (π.χ. αγόρι – κορίτσι). Στη περίπτωση που η κατηγορική μεταβλητή έχει περισσότερες από δύο κατηγορίες, χρησιμοποιούμε τον correlation ratio. Προς το παρόν είναι ανάγκη να αναφερθούμε και πάλι στον δείκτη ρ του Pearson, για να δούμε τον υπολογισμό και την ερμηνεία του μεγέθους της επίδρασης, με τη βοήθεια γραμμικών μοντέλων. Θα αναφερθούμε στο μέγεθος της επίδρασης σε δύο περιπτώσεις: (α) η επεξηγούσα μεταβλητή να ορίζει μια γνήσια διχοτομία (π.χ. αγόρι, κορίτσι), (β) η επεξηγούσα μεταβλητή να ορίζει μια τεχνητή διχοτομία (π.χ. μέγεθος του σχολείου). Για παράδειγμα, ο γράφων μελέτησε την επίδοση τριάντα πέντε χιλιάδων μαθητών και μαθητριών από τριακόσια εβδομήντα πέντε λύκεια κατά τις πανελλαδικές εξετάσεις του έτους 2000. Από τα δεδομένα της έρευνας μπορεί να κατασκευαστεί το μοντέλο: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$. Στο μοντέλο αυτό, η εξαρτημένη μεταβλητή y είναι η κανονικοποιημένη (normalised) επίδοση των μαθητών στα μαθηματικά, με μέσο όρο μηδέν και τυπική απόκλιση ίση με 1. Η μεταβλητή χ_1 αναφέρεται στο «φύλο» των μαθητών, με τιμή «0» για τα αγόρια και «1» για τα κορίτσια. Η μεταβλητή χ_2 αναφέρεται στο «μέγεθος του σχολείου», με τιμή «0», αν το σχολείο είχε λιγότερους από 101 συμμετέχοντες στις γενικές εξετάσεις και «1» αν το σχολείο είχε περισσότερους από 100 συμμετέχοντες στις εξετάσεις. Το 40% των σχολείων ήταν «μεγάλα», είχαν δηλαδή πάνω από 100 συμμετέχοντες. Τέλος, το e είναι ένας όρος σφαλμάτων, ο οποίος ακολουθεί την τυπική κανονική κατανομή με μέσο όρο μηδέν και τυπική απόκλιση ίση με 1. Το μοντέλο υπολογίστηκε ως $y = 0,083\ (0,010) - 0,250\ (0,012)x_1 + 0,122\ (0,012)x_2 + e$.

Όπως φαίνεται τα κορίτσια είχαν χαμηλότερη επίδοση στα μαθηματικά από τα αγόρια, ενώ οι μαθητές που φοίτησαν σε «μεγάλα» λύκεια είχαν επίσης κατά μέσο όρο καλύτερες επιδόσεις από τους μαθητές των «μικρών» λυκείων. Και οι δύο συντελεστές είναι στατιστικώς σημαντικοί για α = 0,001. Η έννοια των συντελεστών των X_1 και X_2, δεν θα μας απασχολήσει εδώ.

Οι μεταβλητές X_1 και X_2 (κάθε μία ξεχωριστά), ορίζουν διχοτομίες (αγόρια και κορίτσια, μεγάλα και μικρά σχολεία). Υπάρχει, όμως, μια διαφορά μεταξύ των X_1 και X_2. Το φύλο είναι μια μεταβλητή σε επίπεδο μαθητών· μια γνήσια

διχοτομία. Από την άλλη μεριά, η διαφορά μεταξύ «μικρού» και «μεγάλου» λυκείου είναι μια τεχνητή διαφορά και στην ουσία το μέγεθος της σχολικής μονάδας ακολουθεί θεωρητικά την κανονική κατανομή. Έτσι, ενώ μπορούμε να συγκρίνουμε απευθείας τους συντελεστές για τις μεταβλητές «φύλο» και «μέγεθος σχολείου μεγαλύτερο από 100», δεν μπορούμε να γράψουμε το ίδιο για την επίδραση της μεταβλητής «μέγεθος του σχολείου» στην εξαρτημένη μεταβλητή. Το μέγεθος του σχολείου είναι μια κατασκευασμένη διχοτομία.

Επειδή η μεταβλητή Y ακολουθεί την τυπική κανονική κατανομή (μέσος όρος 0, τυπική απόκλιση 1), ο συντελεστής της μεταβλητής «φύλο» (-0,250) είναι στην ουσία ο δείκτης ρ του Pearson και μπορεί να χρησιμοποιηθεί ως έχει για τον υπολογισμό του μεγέθους της επίδρασης. Στην πραγματικότητα, ο δείκτης αυτός είναι ο point biserial ρ_{pb}. Ας ξεχάσουμε για λίγο την ύπαρξη της μεταβλητής X_2 και ας φανταστούμε την επίδοση των αγοριών και κοριτσιών σε ένα σύστημα κάθετων αξόνων. Οι τιμές Y_{ij} εκτείνονται κατακορύφως, με γενικό μέσο όρο μ_y, χωρισμένες, όμως, σε δύο ομάδες: μία για τα αγόρια (j = 0), με μέσο όρο μ_0, και μία για τα κορίτσια (j = 1), με μέσο όρο μ_1. Από τη συνολική διακύμανση σ^2, μπορούμε να αφαιρέσουμε τη διακύμανση που οφείλεται στην επίδραση της ανεξάρτητης κατηγορικής μεταβλητής X , δηλαδή τη διακύμανση $\sigma^2_{y/\chi}$. Να γράψουμε δηλαδή $\sigma^2 - \sigma^2_{y/x}$. Αν διαιρέσουμε τη διαφορά αυτή προς τη συνολική διακύμανση σ^2, παίρνουμε τον δείκτη ρ του Pearson για τον πληθυσμό, ο οποίος, όπως είναι γνωστό, υψωμένος στο τετράγωνο, μας πληροφορεί για το ποσοστό της διακύμανσης στην εξαρτημένη μεταβλητή, το οποίο «οφείλεται» στην ανεξάρτητη μεταβλητή. Το ποσοστό αυτό για την έρευνα που αναφέραμε είναι 6,25%. Άρα, περίπου το 6,25% της διακύμανσης στην επίδοση των μαθητών και μαθητριών φαίνεται να εξηγείται από τη μεταβλητή «φύλο». Σύμφωνα με όσα έχουμε πει μέχρι τώρα, ο ρ μπορεί να μετατραπεί σε τυποποιημένη διαφορά μέσων όρων, με τον τύπο $\delta = \frac{2\rho}{1-\rho^2}$. Για τα δεδομένα της έρευνας που αναφέραμε, έχουμε ότι $\delta = 2(-0,250)/1-(-0,250)^2 = -0,470$. Άρα, κατ' απόλυτη τιμή, το 0,47 είναι το μέγεθος της επίδρασης του «φύλου», εκφρασμένο ως

τυποποιημένη διαφορά μέσων όρων. Εκφρασμένο ως z τιμή, το 0,47 μας πληροφορεί ότι το «μέσο» αγόρι έχει καλύτερη επίδοση από το 68% των κοριτσιών.

Παρόμοια είναι τα πράγματα και για την επίδραση της μεταβλητής «σχολείο με περισσότερους από 100 μαθητές στις εξετάσεις». Με την ίδια λογική παίρνουμε ότι περίπου το 1,49% της διακύμανσης στη μαθητική επίδοση «οφείλεται» στη κατηγορική μεταβλητή που προαναφέραμε, ενώ η τυποποιημένη διαφορά των μέσων όρων είναι 0,25. Υπολογίζεται, έτσι, ότι ο «μέσος μαθητής» των σχολείων με πάνω από 100 συμμετέχοντες στις εξετάσεις, έγραψε καλύτερα στα μαθηματικά από το 60% περίπου των μαθητών που φοίτησαν σε «μικρά» λύκεια. Υπάρχει, όμως, ένα σημαντικό ζήτημα εδώ. Είδαμε παραπάνω ότι η μεταβλητή «σχολείο με περισσότερους από 100 μαθητές στις εξετάσεις» έρχεται να αντικαταστήσει τη συνεχή μεταβλητή «μέγεθος του σχολείου» διότι ο ερευνητής αυθαίρετα υιοθέτησε τους 100 μαθητές ως ένα τεχνητό όριο μεταξύ των «μικρών» και των «μεγάλων» σχολείων. Μάλιστα, βρέθηκε ότι το 40% των σχολείων του δείγματος είχε πάνω από 100 μαθητές στις εξετάσεις.

Επειδή κάτω από τη προαναφερθείσα διχοτομία υπάρχει κανονική κατανομή, και επειδή θέλουμε να μιλήσουμε γενικά για την επίδραση του «μεγέθους» του σχολείου και όχι για την επίδραση της διχοτομίας «μεγάλο – μικρό», χρησιμοποιούμε τη λογική της *probit analysis (*από το "probability unit"*)*, ώστε να μπορέσουμε να συγκρίνουμε απευθείας την επίδραση της μεταβλητής X_2 τον δείκτη β_2 τόσο με τον δείκτη β_1 , όσο και με άλλους πιθανούς δείκτες συνεχών επεξηγούντων μεταβλητών. Η λογική της *probit* ξεφεύγει από τον σκοπό του άρθρου, αλλά γενικά η κεντρική της ιδέα είναι ότι η τυποποιημένη διαφορά δύο ποσοστών στο δείγμα συνδέεται με την αθροιστική συχνότητα της κανονικής κατανομής. Για το σκοπό αυτό πολλαπλασιάζουμε τον συντελεστή β_2 , ο οποίος -θυμηθείτε- είναι ο r_{bs}, με το κλάσμα $\frac{\sqrt{pq}}{y}$, όπου p είναι το ποσοστό των μεγάλων σχολείων στο δείγμα, q είναι το 1-p, ενώ y είναι η τετμημένη της κανονικής κατανομής στο σημείο διαχωρισμού μικρών και μεγάλων σχολείων,

εδώ δηλαδή στο 0,40. Τώρα ο δείκτης point biserial r_{pb} ονομάζεται απλώς biserial r_{bis}.

ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΜΕ ΠΕΡΙΣΣΟΤΕΡΕΣ ΚΑΤΗΓΟΡΙΕΣ

Όταν η παρέμβασή μας έχει περισσότερες από 2 κατηγορίες, μπορούμε να γενικεύσουμε και να χρησιμοποιήσουμε αντί για ρ τον δείκτη η^2. Στην περίπτωση αυτή μπορούμε να φανταστούμε αντί της γνωστής ευθείας γραμμής, η οποία περνάει από τους μέσους όρους δύο κατηγοριών της Χ, μία τεθλασμένη γραμμή, η οποία περνάει από τους μέσους όρους των Υ, για κάθε τιμή των Χ. Το σχήμα της γραμμής αυτής δεν έχει κανένα ουσιαστικό νόημα, αφού οι τιμές της Χ είναι απλώς οι «ταμπέλες» (labels) των κατηγοριών. Μάλιστα, αν η μεταβλητή Χ έχει k κατηγορίες, η γραμμή αυτή εκφράζεται από ένα πολυώνυμο (k-1) βαθμών.

Στην περίπτωση αυτή χρησιμοποιούμε τον δείκτη η, ο οποίος εμφανίζεται συχνά στη σύγχρονη ψυχοπαιδαγωγική έρευνα υψωμένος στο τετράγωνο, δηλαδή ως η^2, και συνδέεται με την Ανάλυση Διακύμανσης (ANOVA) αφού είναι ουσιαστικά ο λόγος του αθροίσματος των τετραγώνων των διαφορών *μεταξύ* των ομάδων της παρέμβασης προς το συνολικό άθροισμα του τετραγώνου των διαφορών. Είναι δηλαδή $\eta^2 = \frac{\sum(Y_{ij} - \bar{Y})^2 - \sum(Y_{ij} - \bar{Y}_j)^2}{\sum(Y_{ij} - \bar{Y})^2}$. Όμως, από τη θεωρία της Ανάλυσης Διακύμανσης γνωρίζουμε ότι το πρώτο άθροισμα στον αριθμητή της *Εξίσωσης 2* είναι το συνολικό άθροισμα των τετραγώνων (SS_{total}), ενώ το δεύτερο άθροισμα είναι το άθροισμα τετραγώνων του σφάλματος (SS_{error}), τα οποία αφαιρούμενα δίνουν το άθροισμα τετραγώνων της παρέμβασης. Έτσι, $\eta^2 = \frac{SS_{total} - SS_{error}}{SS_{total}}$ και τελικά $\eta^2 = \frac{SS_{treatment}}{SS_{total}}$. Ο συντελεστής η^2 φανερώνει, με λίγα λόγια, τον λόγο του αθροίσματος των διαφορών της παρέμβασής προς το συνολικό άθροισμα των διαφορών. Ο συντελεστής η^2 δημοσιεύεται όπως θα δημοσιευόταν οποιοσδήποτε άλλος συντελεστής συνάφειας υψωμένος στο τετράγωνο, ενώ αναφέρεται στο ίδιο ακριβώς πράγμα: στο ποσοστό της διακύμανσης που οφείλεται στη παρέμβαση. Ένα μέτρο, δηλαδή, του μεγέθους

της επίδρασης. Άλλοι δείκτες, εναλλακτικοί προς τον η^2 είναι ο δείκτης ε^2, καθώς και ο δείκτης ω^2 (Kelley, 1935), οι οποίοι χρησιμοποιούν διαφορετικό τρόπο υπολογισμού από αυτόν του η^2. Είναι λοιπόν $\varepsilon^2 = \frac{SS_{treatment} - MS_{error}}{SS_{total}}$ και

$$\omega^2 = \frac{SS_{treatment} - (k-1)MS_{error}}{SS_{total} + MS_{error}}$$

Ακολουθεί ένα παράδειγμα από μια Ανάλυση Διακύμανσης με το SPSS και ο υπολογισμός δεικτών επίδρασης.

Πίνακας 1. Ανάλυση διακύμανσης στο SPSS

Source	d.f.	SS	MS	F
Treatments	4	351,52	87,88	9,08
Error	45	435,30	9,67	
Total	49	786,82		

$$\eta^2 = \frac{SS_{treatment}}{SS_{total}} = \frac{351,52}{786,82} = 0,447$$

$$\omega^2 = \frac{SS_{treatment} - (k-1)MS_{error}}{SS_{total} + MS_{error}} = \frac{351,52 - 4(9,67)}{786,82 + 9,67} = \frac{312,84}{796,49} = 0,393$$

ΤΕΛΙΚΕΣ ΠΑΡΑΤΗΡΗΣΕΙΣ

Θα μπορούσε κάποιος να ισχυρισθεί ότι η στατιστική σημαντικότητα των ευρημάτων της έρευνας συνδέεται με τη παιδαγωγική σημαντικότητα. Όμως το *p* δεν έχει σχέση με την ουσία. Το *p* εκτιμάει απλώς την πιθανότητα, ώστε η τιμή που βρήκαμε στο δείγμα μας να αποκλίνει τόσο όσο φαίνεται ότι αποκλίνει από την τιμή που ορίζεται για τον πληθυσμό από τη *μηδενική υπόθεση*. Είναι ανάγκη λοιπόν όλοι όσοι εμπλέκονται σε ψυχοπαιδαγωγική έρευνα να δημοσιεύουν - εκτός από το *p*- στοιχεία για την παιδαγωγική σημαντικότητα κι ένας τρόπος για να γίνει αυτό είναι η δημοσίευση στοιχείων για το «μέγεθος του αποτελέσματος» ή *effect size*. Υπάρχουν πολλοί τρόποι για τη μέτρηση και δημοσίευση του μεγέθους του αποτελέσματος, ένας εκ των οποίων είναι οι διάφοροι δείκτες. Οι υπάρχοντες δείκτες επικεντρώνονται είτε στις παρατηρούμενες *διαφορές* μεταξύ δύο ή περισσοτέρων στατιστικών στοιχείων είτε στην «αποτελεσμένη» διακύμανση (*variance-accounted-for*). Και επειδή όλη η στατιστική ανάλυση είναι στην ουσία μια ανάλυση *σχέσεων*, η μελέτη των διαφορών και της διακύμανση

συνδέεται με τη μελέτη *σχέσεων* συνάφειας και συσχέτισης. Πέντε χρήσιμοι δείκτες μεγέθους του αποτελέσματος είναι οι εξής:

Πίνακας 2. Δείκτες για το μέγεθος της επίδρασης στην παιδαγωγικη έρευνα

Διαφορές μέσων όρων	*Διακύμανση (variance-accounted-for)*
Glass's g'	Eta squared (η^2) λόγος συνάφειας
Cohen's d	Kelley's epsilon squared (ε^2)
	Hays's omega squared (ω^2)

Με τη δημοσίευση δεικτών για το μέγεθος του αποτελέσματος πετυχαίνουμε: (α) να συγκρίνουμε διαφορετικά στατιστικά στοιχεία σε ιδιαίτερα σύνθετους ερευνητικούς σχεδιασμούς, (β) να συγκρίνουμε στατιστικά στοιχεία και ευρήματα διαφορετικών ερευνών, και (γ) να παρουσιάζουμε τα ευρήματα της ψυχοπαιδαγωγικής έρευνας σε ανθρώπους που δεν είναι εξοικειωμένοι μαζί της και (δ) να συζητήσουμε για τη σημαντικότητα ενός στατιστικού στοιχείου, ανεξάρτητα από το p.

Βιβλιογραφικές αναφορές

Barros, C. P., Peypoch, N., & Louçã, F. (2008). *Should The Widest Cleft in Statistics - How and Why Fisher opposed Neyman and Pearson*. Retrieved from http://www.iseg.utl.pt/departamentos/economia/wp/wp022008deuece.pdf

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

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

Fontana, D., & Abouserie, R. (1993). Stress levels, gender and personality factors in teachers. *Educational Psychology*, *63*, 261–270.

Glass, G. (1976). Primary, secondary, and meta-analysis of research. *Educational Researche*, *5*(10), 3–8.

Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London: Academic Press.

Kazdin, A. (1999). The meaning and measurement of clinical significance. *Journal of Consulting and Clinical Psychology*, *67*(2), 332–339.

Randolph, J. J. (2005). Using the Binomial Effect Size Display (BESD) to Present the Magnitude of Effect Sizes, *10*(14). Retrieved from http://pareonline.net/pdf/v10n14.pdf

Rosenthal, R., & Rubin, D. (1982). A simple general purpose display of magnitude and experimental effect. *Journal of Educational Psychology*, *74*, 166–169.

Smith, M., & Glass, G. (1977). Meta-analysis of psychotherapy outcome studies. *American Psychologist*, *32*(9), 752–760. doi:10.1037/0003-066X.32.9.752

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, *54*(8), 594–604.

Λεονταρή, Α., Κυρίδης, Α., & Γιαλαμάς, Β. (2000). Το επαγγελματικό άγχος των εκπαιδευτικών της πρωτοβάθμιας εκπαίδευσης. *Παιδαγωγική Επιθεώρηση*, *7*(30), 139–161.

Τσιλφίδου, Χ., & Πλατσίδου, Μ. (2011). Κίνητρο επίτευξης και συμμετοχή στη Συνεχιζόμενη Εκπαίδευση των Εκπαιδευτικων Πρωτοβάθμιας Εκπαίδευσης. *Μέντορας*, *13*, 165–180.

Χάρτης Θεμελιωδών Δικαιωμάτων της Ευρωπαϊκής Ενωσης. , Pub. L. No. 364/01 (2000). Επίσημη Εφημερίδα των Ευρωπαϊκών Κοινοτήτων. Retrieved from http://www.europarl.europa.eu/charter/pdf/text_el.pdf