Psihološka obzorja / Horizons of Psychology, 14, 3, 7-40 (2005)

Znanstveni teoretsko-pregledni prispevek

Statistična pomembnost in njen pomen[#]

Christina Bachmann, Riccardo Luccio in Emilia Salvadori[*]
Oddelek za psihologijo, Katedra za metodologijo, Univerza v Firenzah

Povzetek: Glede na najbolj uporabljeno metodo v psihologiji, in sicer testiranje ničelne hipoteze (TNH), je namen tega prispevka izpostaviti napačna razumevanja in napake, ki izhajajo iz »neprimerne« uporabe hibridnega pristopa, nastalega z združevanjem Fisherjevega (1925) »pristopa *p*-vrednosti« (PVA) ter Pearsonovega in Neymanovega (1933) pristopa »določene alfa verjetnosti« (FAA). V prispevku želimo prispevati h kritični razpravi o uporabi statističnih testov in interpretaciji rezultatov s strani raziskovalcev ter tudi o sami logiki TNH pristopa. Poleg tega smo, kot je nedavno predlagala delovna skupina APA *Task Force*, pregledali metode analize podatkov, ki jih moramo priključiti TNH, da bi zagotovili večjo zanesljivost rezultatov. Te metode so intervali zaupanja, mere velikosti učinka, analiza moči testa in prevzorčenje.

Ključne besede: testiranje ničelne hipoteze, statistična pomembnost, analiza moči, velikost učinka

Statistical significance and its meaning

Christina Bachmann, Riccardo Luccio and Emilia Salvadori
Dipartimento di Psicologia, Sezione di Metodologia, Universita di Firenze, Italia

Abstract: The most frequently method of data analysis in psychology is null hypothesis significance testing (NHST); this work identifies the misunderstandings and consequent biases due to the unwitting use of this hybrid approach, derived by the fusion of the *p*-value approach (PVA) of Fisher (1925) and the fixed alpha approach (FAA) of Neyman and Pearson (1933). The aim of this work is to contribute to the critical debate about the use of this test, how researchers interpret test results, and the logic of the NHST approach. In the same vein of recent recommendations of the APA Task Force, we examine different methods of data analysis, that are used with NHST including confidence intervals, effect size measures, power analysis, and resampling.

Key-words: null hypothesis testing, statistical significance, power analysis, effect size (statistical)

CC = 2240

[#] Prispevek je bil predstavljen na konferenci 16. dani Ramira Bujasa v Zagrebu, 11.-13.12.2003 (kot vabljeno predavanje avtorja R. Luccia), in je izvleček iz knjige La verifica della significativita dell'ipotesi nulla in psicologia istih treh avtorjev, izdane leta 2005 pri založbi Firenze University Press. Besedilo je iz italijanskega izvirnika prevedla Janja Tomšič. Besedilo prevoda sta strokovno pregledala Gregor Sočan in Anja Podlesek.

[*] Naslov / address: Emilia Salvadori, Dipartimento di Psicologia, via S. Niccola 93, 50125 Firenze, Italia; e-mail: lmsalv@tin.it

Testiranje ničelne hipoteze (TNH) v psihologiji

Prevladujoča paradigma pri analizi podatkov psiholoških raziskav je t.i. testiranje ničelne hipoteze (v angleščini *NHST, Null Hypothesis Significance Testing*). TNH je produkt združevanja pristopa *p-value approach* (PVA, Fisher, 1925) in pristopa *fixed alpha approach* (FAA, Neyman in Pearson, 1933), ki imata skupne poteze, a z opaznimi razlikami. Vendar pa se citirani avtorji niso strinjali niti s pristopom drugega avtorja niti z mešanico obeh pristopov.

Za razlago metode TNH študentom pri lekcijah statistike za psihologe navadno uporabljamo sledečo shemo, ki omogoča vizualizacijo procesa odločanja. Samo Bog ve, ali je v naravi resnična ničelna hipoteza H_0 (ki taji učinek obravnave ali korelacijo med spremenljivkami itd.) ali alternativna hipoteza H_1 (ki potrjuje učinek obravnave itd.). Statistik skuša na osnovi računane verjetnosti uganiti, kar ve Bog.

		Bog	
		H_0	H_1
Statistik	H_0	Pravilna (ustrezna) zavrnitev	Napačna zavrnitev Napaka II tipa β
	H_1	Napačna potrditev Napaka I tipa α	Zadetek Moč (1- β)

Opomba: Tabela govori o zavrnitvi in potrditvi obstoja pojava (učinka itd.) in ne o zavrnitvni in potrditvi ničelne hipoteze (op. ur.).

Inferenčni proces naj bi vključeval naslednje stopnje: (i) določiti vrednost verjetnosti α (običajno 0,05); (ii) privzeti, da je H_0 resnična; (iii) določiti verjetnost dejanskih podatkov, če je H_0 resnična [tj, p (podatki | H_0)]; (iv) če je $p > \alpha$, sprejeti H_0; (v) v nasprotnem primeru zavrniti H_0 in sprejeti H_1. Ta postopek imamo lahko za »fisherjanskega«, za vse matematično orodje pa imata zasluge Neyman in Pearson (1933), ki sta med drugim predlagala bistveno drugačno strategijo odločanja. Da si stvari nekoliko razjasnimo, je koristno, da si pobliže pogledamo pristopa PVA in FAA.

Pristop PVA (*p*-value approach)

Fisher je pristop PVA najbolj jasno predstavil v delu *Statistical Methods for Research Workers* (1925), kjer je trdil, da če ob poznavanju karakteristik določene populacije iz nje vzamemo vzorec in ta krši naša pričakovanja, »...lahko rečemo, da je bil vzet iz druge populacije. […Če] imamo na razpolago kritične teste za pomembnost, lahko ugotovimo, ali je drugi vzorec pomembno drugačen od prvega ali ne« (str. 44). Nekoliko kasneje se pojavi meja verjetnosti 0,05, ki »je koristna […] za presojanje, ali

je določeno odstopanje pomembno ali ne« (str. 48). Kasneje je Fisher (1955, 1956) trdil, da moramo raven pomembnosti računati po tistem, ko je bil test že izveden, in je to torej lastnost podatkov.

Pristop FAA (fixed-alpha approach)

Zgoraj predstavljena shema pa izhaja predvsem iz Neymanovega in Pearsonovega (1933) pristopa FAA. Neyman in Pearson sta gledala na teste pomembnosti kot na metodo za izbiro ene od dveh možnih hipotez. Ta teorija je dobila svojo dokončno obliko v znani lemi, ki ga v zgoščeni obliki predstavljamo v Prilogi. Neyman in Pearson sta najprej jasno definirala ničelno in alternativno hipotezo, v nasprotju z njima pa se je Fisher omejil na ničelno hipotezo (čeprav je ni imenoval tako). Če privzamemo, da je bil vzorec vzet iz določenega univerzuma, se parameter θ, za katerega menimo, da je relevanten (npr. povprečje pri normalni porazdelitvi univerzuma in vzorca), ne bo pomembno razlikoval od ustreznega parametra populacije, če je resnična H_0, ter se bo od le-tega pomembno razlikoval, če je resnična H_1. Neyman in Pearson sta torej trdila, da je treba določiti raven α za prvo hipotezo. Iz tega izhaja tudi ime pristopa. Nato nadaljujemo z ocenjevanjem parametra z metodo največjega verjetja. Izračunamo odnos med koeficienti največjega verjetja pod pogojem, da bi bila resnična H_0, in pod pogojem, da bi bila resnična H_1. Glede na to, da α ustreza neka določena kritična vrednost tega odnosa, se v primeru, da je dosežena vrednost nižja od določene, zaključi v prid ničelni hipotezi, v nasprotnem primeru pa v prid alternativni hipotezi. Za Neymana in Fisherja to ni bila končna odločitev, ampak le prvi korak, ki ga uporabimo kot izhodišče plodne poti, na kateri se dokopljemo do nadaljnjih potrditev.

Metodo TNH imamo lahko za »fisherjansko«, ker se pri sprejetju ali zavrnitvi ničelne hipoteze, za katero privzamemo, da je resnična, naslanjamo na verjetnost podatkov. Vendar pa je gotovo ne moremo imeti za »fisherjansko« glede koncepta dveh nasprotnih si hipotez, čeprav lahko v nekem smislu le-ti obravnavamo kot implicitno vsebovani v Fisherjevi razlagi. Koristna je lahko shema, ki poudarja glavne razlike med obema pristopoma. Če delno sledimo Hubertyju (1993), lahko rečemo, da (i) je pri pristopu PVA verjetnost variabilna, pri pristopu FAA pa je α vnaprej določena (fiksirana); (ii) imamo pri pristopu PVA samo H_0, pri FAA pa tudi H_1; (iii) se pri pristopu PVA določi p (podatki | H_0 resnična), pri pristopu FAA pa se določi napaka I. tipa in kritično območje C; (iv) pri PVA H_0 zavrnemo, če je p majhen, sicer jo dopustimo, pri FAA pa H_0 zavrnemo v prid H_1, če se T nahaja v kritičnem območju distribucije, sicer jo dopustimo; (v) na tej točki se za Fisherja eksperiment konča, medtem ko se za Neymana in Pearsona ponovi – če je možno na novih vzorcih. V naslednjem poglavju bomo videli, kako je bilo vse to združeno v en sam pristop.

Obravnava TNH

V štirih glavnih mednarodnih revijah eksperimentalne psihologije so med letoma 1930 in 1940 TNH uporabili samo v štirih člankih, med letoma 1940 in 1955 pa so to metodo uporabili za analizo podatkov v 80 % člankov (Sterling, 1959) in se je uveljavila kot »paradigma« (Kuhn, 1962). Kot je opazil Yates (1951, str. 33), so jo »raziskovalci pogosto imeli za cilj eksperimenta«.

Že od 30-ih let je metoda TNH sprožila veliko kritičnih razprav o njeni uporabi in logiki. Veliko matematičnih psihologov, z Duncanom Luceom na čelu, je ni sprejelo in so jo imeli za oviro znanstvenemu napredku. Tudi v sami eksperimentalni psihologiji so obstajala nasprotovanja metodi TNH: Skinner je ustanovil revijo *Journal of Experimental Analysis of Behavior* ravno zato, da bi ubežal poniževanju TNH.

Ne glede na kritike pa je bilo v praksi malo sprememb (Cohen, 1994). Izpostaviti je treba objavo posebne številke revije *Journal of Experimental Education*,Vol. 61, št. 4, in revije *Psychological Science*, Vol. 8, št. 1, ter knjige, posvečene prihodnjim perspektivam testov pomembnosti v psihologiji, s prispevki »za« in »proti« (Harlow, Mulaik in Steiger, 1997; za pregled glej Nickerson, 2000; Sullivan, 2000). Kot glavni problemi so se pokazali naslednji.

Moč testa

Prvi od problemov (Cohen, 1977, 1988) je zanemarjanje moči uporabljenega testa. Le-to bi morali, tudi zaradi določanja optimalne velikosti uporabljenih vzorcev, določiti pred zbiranjem podatkov, kar pa raziskovalci skoraj nikoli ne naredijo (k tej temi se bomo še vrnili).

Strogost “binarne” odločitve

Določitev kritičnega praga izbire med dvema alternativama je problematična. Tako je treba sprejeti H_0, ker je dobljena vrednost p enaka 0,052. Ta vrednost pa je tako blizu pomembnosti, da kljub sprejetju ničelne hipoteze pušča dvom o tem, ali ni morda učinek vseeno dejansko prisoten. Pomislimo na metaanalize, ki predpostavljajo, da se rezultati povzetih raziskav porazdeljujejo vzdolž določenega kontinuuma.

“Poljubnost” α

Vrednost α je običajno določena na 0,05 (5 %), včasih pa na 0,01 (1 %). Za to odločitev ni nobene teoretične motivacije. Kot pravita Rosnow in Rosenthal (1989, str. 1277): »Bog ima 0,06 ravno toliko rad kot 0,05«. Primerno je imeti nizko raven α v raziskavah, ki imajo lahko resne posledice na zdravstveni in ekonomski ravni. Bolj kot je raziskava pomembna, bolj bi morali biti pozorni na moč testa in na ponovljivost rezultatov, torej na probleme, neodvisne od ravni α.

Precenitev *p*

Običajno verjamemo, da nižja kot je vrednost *p*, manj je H_0 sprejemljiva. Vendar pa *p* računamo glede na predpostavko, da je ničelna hipoteza resnična. Ni dovolj, da razberemo samo, ali je *p* večji ali manjši od α, saj obstoj razlike ne prispeva nobene informacije o ničelni hipotezi. V osnovi te napake je ideja, da *p* meri stopnjo, v kateri je rezultat dosežen po slučaju (Carver, 1993; Daniel, 1998): *p* 0,03 naj bi pomenil, da obstaja 3-odstotna verjetnost, da so rezultati posledica slučaja, medtem ko naj bi bila 97-odstotna verjetnost, da je to resničen učinek.

Statistična pomembnost nasproti praktični pomembnosti

Težimo tudi k temu, da vrednost pomembnosti interpretiramo kot kazalec velikosti učinka. Za vrednosti *p*, ki so precej manjše od 0,05, se uporabljajo izrazi kot »zelo pomembni rezultati«. Statistična pomembnost ne ustreza velikosti učinka. Pusti nam domnevati, da obstaja določen učinek, vendar nam o njegovih značilnostih ne pove ničesar. Vrednost *p* ni odvisna le od upoštevanega učinka (θ), ampak tudi od velikosti vzorca. Poleg tega sam odnos med θ in *p* ni linearen – majhnemu povečanju učinka lahko ustreza mnogo večji upad vrednosti *p* in obratno.

Izraz »pomembnost« tako samovoljno razumemo kot »pomembnost (relevantnost)« (primerjaj Schafer, 1993; Meehl, 1997). Predvsem nekateri ugledni člani združenja *American Educational Research Association* (prim. Thompson, 1996, 1997; Cahan, 2000) so predlagali, naj izraz »pomembnost« vedno spremlja pridevnik »statistična«, da bi izločili učinek nejasnosti, ki ga spremlja (za drugačno mnenje glej Robinson in Levin, 1997; Levin in Robinson, 1999). Thompson (2003) razlikuje tri tipe pomembnosti: statistično, praktično in klinično. Statistična pomembnost ne pove, ali so rezultati praktično pomembni. Obstajajo določeni redki in nenavadni dogodki, ki niso relevantni, in nekateri pogosti in verjetni dogodki, ki pa so zelo pomembni.

Precenitev rezultatov

Ničelna hipoteza H_0 vsebuje predpostavko, da so učinki ničelni, torej da vzorčna ocena θ povsem ustreza »pravi« θ^*. Vendar pa je težko, da bi učinek, četudi zelo majhen, imeli za neobstoječega, in prav tako je samovoljno misliti, da je podobnost med ocenjenim in pravim θ lahko popolna.

Sprejetje ničelne hipoteze: težka in neprijetna odločitev

Raziskovalec praviloma, če je njegov *p* (podatki | H_0) večji od vnaprej določene vrednosti α, meni, da bo njegovo delo verjetno končalo v košu za star papir, in zaradi lastne cenzure del s takimi »negativnimi« izidi ne bo objavil (Hubbard in Armstrong, 1997). Wilson, Smoke in Martin (1973) so pri pregledu člankov, objavljenih v letih 1969 in 1970 v revijah *American Journal of Sociology, American Sociological*

Review in *Social Forces*, ugotovili, da je bila ničelna hipoteza zavrnjena v 80,3 % primerov, v katerih je bila uporabljena metoda TNH. Greenwald (1975) je pri pregledu letnika 1972 revije *Journal of Personality and Social Psychology* ugotovil, da se je tu enako zgodilo v 87,9 % člankov. V člankih o marketingu ta odstotek naraste na 92,2 % (Hubbard in Armstrong, 1992, 1994), po Lindsayu (1994) pa na 84,2 %. Sterling, Rosenbaum in Weinkam (1995) so pokazali, da je z metodo TNH zavrnjena ničelna hipoteza v 95,6 % objavljenih psihološśkih delih, v medicinskih revijah pa je ta odstotek 85,4 %. Hubbard in Armstrong (1992) sta pregledala moč v maloštevilnih raziskavah, v katerih je bila ničelna hipoteza sprejeta. Ugotovila sta, da so bile povprečne verjetnosti, da ima lahko obravnava velik ($p = 0{,}01$) ali samo srednje velik učinek ($p = 0{,}11$), zelo nizke, kar je nedvomno informativno.

Pri politiki objavljanja revij obstaja celo odpor zoper objave del, kjer je raven p samo 0,05, tako da se daje prednost ravni p 0,01 ali manjšši, kot je iskreno priznal Melton (1962), ko se je po 12 letih poslovil od vodenja revije *Journal of Experimental Psychology.*

Posebnost (omejitve) ničelne hipoteze v naravi

Hipoteza H_0 v splošno priznani praksi testiranja hipotez, vsebuje premiso, da so predvidevani učinki nični, torej da je razlika med povprečij populacij nič, da je korelacija nič, da je odstotek moških in žensk v kateri koli populaciji 50 %, vendar je to arbitrarno določeno (Bakan, 1966; Loftus, 1996). Chow (1996, 1998), Rindskopf (1997) ter Levin in Robinson (1999) dvomijo, da je problem resnično pomemben. Kot opozarja Tukey (1991, str. 100), noben raziskovalec ne privzame, da je razlika med dvemi povprečji ali korelacija med dvema spremenljivkama točno 0. Mulaik, Raju in Harshman (1997) menijo, da problem prej predstavljajo odkloni od ničelne hipoteze zaradi velikosti obravnavanega vzorca.

Vpliv velikosti vzorca

Velikost vzorca vpliva na pomembnost dobljenih rezultatov: »Pri velikem številu podatkov vrednosti p težijo k majhnosti« (Berkson, 1938, str. 526; glej Nunnally, 1960; Kerlinger, 1979; Hays, 1963; Cohen, 1994). Vrednost r 0,6 ni statistično pomembna pri 10 osebah, pri 500 osebah pa bi, za p manjši od 0,05, zadostoval r 0,088 (0,77 % pojasnjene variance) (Daniel, 1998). Pri 10000 osebah bi zadostovala vrednost r, ki bi bila enaka 0,0196! Razmerje med dvema χ^2, ki ju dobimo na vzorcih velikosti N_1 in N_2, je (prim. Long, 1983, str. 75) enako:

$$\chi_2^2 = \chi_1^2 \frac{N_2 - 1}{N_1 - 1} \tag{1}$$

Tako bi npr., če je N_1 enak 100 in N_2 enak 1000, z eno stopnjo svobode χ^2 prvega vzorca 1,64, kateremu bi ustrezala vrednost p 0,20, v drugem vzorcu postal

16,55, z vrednostjo *p* manjšo od 0,001! Podobno lahko razmišljamo tudi pri Studentovem *t* in ANOVA-i.

Bonferronijev problem

Določiti raven α na 0,05 je enako kot reči, da obstaja verjetnost 0,95, da bomo prišli do »nepomembnega« zaključka. Če pa preverjamo dve neodvisni ničelni hipotezi in če je raven α za obe enaka, je verjetnost, da ni nobena od dveh pomembna, enaka $0,95 \times 0,95 = 0,90$. In če je neodvisnih ničelnih hipotez 15, je verjetnost, da ni nobena od teh pomembna, enaka $0,95^{15} = 0,46$, z verjetnostjo 0,54 (večjo od slučaja), da bomo morali zavrniti vsaj eno. Z 20 ničelnimi hipotezami bi bila verjetnost celo 1 (pravzaprav 0,64; op. ur.). To je znan Bonferronijev problem (1936; prim. Bland, 2000; za drugačno mnenje glej Perneger, 1998). Bonferroni je predlagal, da bi določili popravljen α, $\alpha_c = \alpha/k$, kjer je *k* število neodvisnih ničelnih hipotez, ki jih preverjamo. Tako bi bil s petnajstimi ničelnimi hipotezami $\alpha_c = 0,05/15 = 0,003$. Vendar pa se ta popravek skoraj nikoli ne uporablja.

Ponovljivost rezultatov

Falsifikacijo (zavrnitev) ničelne hipoteze pogosto interpretiramo kot dokaz ponovljivosti dobljenih rezultatov (Carver, 1978, govori o želji po ponovljivosti). Vendar bi morala biti moč statističnega testa zadovoljiva (vsaj 0,80; prim. Cohen, 1988), da bi upravičeno mislili, da bi v enakih pogojih ponovitev zajemala iz istih dveh med seboj različnih univerzumov, iz katerih smo dobili dva obravnavana vzorca.

Intenziteta in smer učinka

Pri TNH pomembnost ne more potrditi drugega kot obstoj in smer neke razlike. Če je povprečje vzorca, izpostavljenega obravnavi A, večje od povprečja vzorca, izpostavljenega obravnavi B, in če je ničelna hipoteza falsifikabilna, očitno lahko potrdimo, da A proizvaja naraščanje učinka v primerjavi z B, vendar pa ne moremo povedati, kolikšen je ta prirastek.

Logika testa pomembnosti in formalna logika

Logika, na kateri temelji zavrnitev ničelne hipoteze in posledično sprejetje alternativne hipoteze, se opira na pravila silogističnega razmišljanja (Cohen, 1994), še posebej na *modus tollens*, po katerem negacija sledečega implicira negacijo predhodnega. Na videz bi bila logika testa hipotez naslednja:

Pri vsaki ničelni hipotezi, ki je resnična, ni učinka obravnave.
Našli smo učinek obravnave.

$\therefore$ Torej je ničelna hipoteza napačna.

Vendar je testiranje hipotez po naravi verjetnostno in to pravilo v formalni logiki ni uporabno. To je lahko razumljivo z naslednjim silogizmom:

Za vsakega italijanskega prebivalca je malo verjetno, da bo predsednik republike.
Ciampi je predsednik republike.

Torej je malo verjetno, da je Ciampi italijanski prebivalec.

Problem linearnosti modelov

Klasičen linearen model v »fisherjanski« statistiki predpostavlja, da je opazovani podatek linearna kombinacija, določena s pripadnostjo podatka določenemu univerzumu (predstavljeno s θ), vpliva *j*-te obravnave (α_j) in napake *i*-te osebe (e_j):

$$y_{i,j} = \theta + \alpha_j + e_i \quad (2)$$

Ta model vzdrži kot linearen, vendar ga nič ne zagotavlja. Jasno je, da v veliko primerih odnosi (npr. pri interakciji obravnav) niso linearni. Zato so potrebne vsaj tri dopolnilne postavke: i) da se napake porazdeljujejo normalno; ii) da napake med seboj niso korelirane; iii) da je varianca porazdelitev napak enaka za vsako eksperimentalno situacijo. Vendar tudi tu nič ne zagotavlja, da bodo stvari šle na tak način (glej Micceri, 1989).

V bran metode TNH

Ne želimo pa dati vtisa, da statistiki TNH smatrajo kot že zastarel pristop. Še vedno ima privržence. Eden izmed njih je Chow (1996, 1998, 1999), ki meni, da so kritike ponavljajoča se »litanija«, ki je štirideset let predlagala vedno znova enake in netehtne argumente. Kot opozarja Abelson (1997), je problem vezan predvsem na to, da je TNH treba uporabljati previdno, v okviru omejitev in predvsem tveganja pri njeni nekritični uporabi. Po Sullivanu (2000) lahko glavne argumente v bran TNH razvrstimo v naslednje kategorije.

Odločitve na nominalni ali ordinalni ravni

TNH je koristna, kadar želimo delati na ravni neurejenih kategorij (nominalne lestvice) in asimetričnih odnosov (ordinalne lestvice). V teh primerih ni pomembna velikost učinka, ampak samo njegova smer. Tu je falsifikacija ničelne hipoteze zadostna.

Pomanjkanje alternativnih analiz podatkov

Raziskovalci niso zadovoljni s predlaganimi alternativnimi analizami. Zaradi zahtev raziskave (npr. glede števila oseb) ni vedno možno dobiti zadovoljive moči in pogosto se zato s tem ni vredno ukvarjati. Večji del psiholoških raziskav se ne ukvarja z razvojem modelov (napovedjo funkcije, na primer, v odvisnosti od časa) kot je to v drugih znanostih. Prav tako se ne ukvarja npr. s študijami kovariančnih struktur, kot v LISREL-u.

Odgovornost raziskovalcev za napake in napačna razumevanja

TNH je bila primarno zasnovana kot logičen in dobro utemeljen način statistične analize. Raziskovalci so krivi, če je metoda TNH slabo uporabljena in je povzročila veliko napak.

Pomanjkanje alternativnih oblik analize

Čeprav je to šibka metoda, ne smemo pozabiti, da je pogosto edina analiza, ki nam je resnično na razpolago. Temu lahko le pritrdimo.

Priporočila Ameriškega psihološkega združenja (APA)

Po sledeh teh razprav in predvsem zaradi obširne razprave, ki jo je povzročila objava Cohenovega članka *The earth is round (p < .05)* v letih 1994 in 1996, je na burni razpravi na konvenciji Ameriškega psihološkega združenja (APA) (Abelson, 1997, meni, da je bilo videti, kot bi »skupina radikalnih aktivistov« vzela za talce »10 statistikov in 6 direktorjev revij, pod geslom 'Podpri popolno opustitev testov (pomembnosti)' in 'razveljavi ničlo!'«) Znanstveni svet APA zaupal komisiji strokovnjakov, imenovani *Task Force on Statistical Inference* (TFSI), nalogo, naj preuči situacijo in pokaže na možne alternative. TFSI je identificirala predvsem dva povezana, a različna predmeta študija: (i) vlogo TNH v psihološkem raziskovanju; (ii) spremembe, ki so se zgodile skozi čas kot posledica obravnave podatkov v psihologiji. TFSI je predlagala vrsto akcij za izboljšanje stanja, in sicer najprej revizijo APA priročnika (APA, 1994). Ta opažanja so združena v reviziji priročnika (APA, 2001).

Glede na TFSI (Wilkinson in TFSI, 1999) je potrebno pri objavi rezultatov (i) natančno poročati o vseh problemih, ki so se pojavili med zbiranjem podatkov (npr. o manjkajočih podatkih ali osamelcih), s priporočili za preprečitev njihovega ponovnega pojavljanja; (ii) zbrati *preproste in parsimonične analize*, ne pa sofisticiranih metod, ki k znanju dodajo malo; (iii) odpovedati se dihotomni izbiri med sprejetjem in zavrnitvijo H_0 ter podati le vrednost p, še bolje pa je podati *intervale zaupanja* in indekse *velikosti učinka*; (iv) pri prikazovanju rezultatov uporabljati *slike* z grafično predstavitvijo intervalov zaupanja.

Problem velikosti učinka

Kot dopolnilo TNH je bilo predlaganih več indeksov, t. i. mer velikosti učinka (Maxwell in Delaney, 1990), ki jih lahko razdelimo v dve kategoriji (Nix in Barnette, 1998): (i) moč odnosa med spremenljivkami (mere povezanosti); (ii) velikost učinka (v ožjem pomenu izraza).

Mere povezanosti

Mere povezanosti navadno kažejo odstotke pojasnjene variance (Maxwell in Delaney, 1990), oziroma, koliko variance odvisne spremenljivke je povezane z variiranjem neodvisne spremenljivke.

ANOVA

Pri ANOVI ponavadi uporabljamo štiri tipe mer povezanosti: (i) eta kvadrat (η^2); (ii) delni eta kvadrat (η_p^2); (iii) omega kvadrat (ω^2); (iv) intraklasna korelacija (ρ_i). Indeksa ω^2 in ρ_i sta oceni stopnje povezanosti, računane na populaciji, medtem ko sta η^2 in η_p^2 oceni stopnje povezanosti, računane na vzorcu.

Eta kvadrat (η^2), imenovan korelacijsko razmerje (*correlation ratio*), predstavlja odstotek totalne variance, ki ga lahko pripišemo učinku. Dobimo ga z odnosom med odklonom zaradi učinka (SS_{eff}) in totalnim odklonom (SS_t). Računanje η^2 je naslednje:

$$\eta^2 = \frac{SS_{eff}}{SS_t} \tag{3}$$

Eden od problemov, povezanih z uporabo η^2, izvira iz dejstva, da je vrednost tega indeksa pri enem od učinkov odvisna od števila in velikosti ostalih preučevanih učinkov. Tako bi se, če bi dvema neodvisnima spremenljivkama dodali še tretjo, vrednost učinka, pripisana interakciji med prvima dvema, zmanjšala, medtem ko bi varianca, pripisana tej isti interakciji, ostala nespremenjena.

Parcialni eta kvadrat (η_p^2) se od η^2 razlikuje v tem, da v imenovalcu enačba ne uporablja totalne variance (SS_t), ampak vsoto med varianco zaradi učinka (SS_{eff}) in varianco zaradi napake (SS_{err}). V enačbi:

$$\eta_p^2 = \frac{SS_{eff}}{(SS_{eff} + SS_{err})} \tag{4}$$

Indeks *omega kvadrat* (ω^2) ocenjuje količino variance, pojasnjene z neodvisno spremenljivko v populaciji. Temelji na parametrih populacije, ki pa so ponavadi nepoznani

in jih je treba oceniti na osnovi vzorčnih podatkov. Ocenjen ω^2 indeks ($\hat{\omega}^2$) je izražen z enačbo:

$$\hat{\omega}^2 = \frac{(k-1)(F-1)}{(k-1)(F-1)+kn} \tag{5}$$

kjer *k* pomeni število skupin, *n* je število oseb v skupini in *F* vrednost testa. Teh enačb ne moremo uporabiti pri ponovljenih merjenjih.

Indeks *intraklasne korelacije* (ρ_I) ocenjuje povezanost med neodvisno in odvisno spremenljivko v populaciji za model *random effects* (model naključnih učinkov, op. prev.):

$$\rho_I = \frac{MS_{\text{eff}} - MS_{\text{err}}}{MS_{\text{eff}} + gl_{\text{eff}}\, MS_{\text{err}}} \tag{6}$$

Analogno koeficientu determinacije je kvadrat indeksa ρ_I indeks deleža variance odvisne spremenljivke, ki ga pojasni neodvisna spremenljivka.

Povezanost in korelacija

Koeficiente korelacije lahko pojmujemo kot mere velikosti učinka. Bravais-Pearsonov koeficient *r* (kot tudi točkovno biserialni koeficient r_{pb} in Φ) lahko računamo na osnovi vrednosti Studentovega *t* testa in χ^2:

$$r = \sqrt{\frac{\chi^2_{(1)}}{N}} \tag{7}$$

$$r = \sqrt{\frac{t^2}{t^2 + df}} \tag{8}$$

Koeficienta *r* in r_{pb} lahko dobimo tudi iz mer velikosti učinka v ožjem pomenu izraza, ki bosta podrobno predstavljeni v naslednji točki: Cohenovega *d* in Hedgesovega *g*. Za celovitost razlage pa prikazujemo tudi enačbe, ki omogočajo izračun *r* iz teh indeksov. Za pretvorbo *r* v *d* se uporablja naslednja enačba:

$$r = \frac{d}{\sqrt{d^2 + 4}} \tag{9}$$

Če je vzorec majhen ali so numerusi vzorcev zelo različni, je bolj natančna naslednja enačba (Aaron, Kromrey in Ferron, 1998):

$$r = \frac{d}{d^2 + \sqrt{\frac{N^2 - 2N}{n_1 n_2}}} \tag{10}$$

Za izračun r iz Hedgesovega g se uporablja naslednja enačba:

$$r = \frac{g}{\sqrt{g^2 + 4 \cdot \frac{gl_w}{N}}} \tag{11}$$

Povezanost in kontingenčne tabele

Najbolj uporabljena mera povezanosti v kontingenčnih tabelah je Pearsonov koeficient kontingence C, ki variira med 0 (odsotnost učinka) in 1 (zgornja meja). Enačba je:

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}} \tag{12}$$

Kot smo videli pri prejšnjih enačbah, lahko tudi koeficient korelacije uporabimo za določitev velikosti učinka.

Na koncu prikazujemo še Cramerjev ϕ' indeks, ki ga uporabljamo za χ^2, dobljen iz kontingenčnih tabel, in ki ima to prednost, da ga lahko uporabljamo s tabelami katere koli dimenzije:

$$\phi' = \sqrt{\frac{\chi^2}{N \cdot gl_{min}}} \tag{13}$$

kjer je $gl_{min} = k - 1$, k pa je tista vrednost števila stolpcev ali števila vrstic, ki je manjša.

Povezanost in regresija

Pogosto uporabljena mera povezanosti v regresijski analizi je koeficient determinacije r^2 (pogosto označen z R^2). Ta daje kvantifikacijo deleža variance odvisne

spremenljivke, določenega z njenim odnosom z neodvisno spremenljivko.

$$r^2 = \frac{SS_{\text{eff}}}{SS_{\text{t}}} \tag{14}$$

Koeficient r^2 lahko zavzema vrednosti, ki variirajo med 0 (odsotnost učinka – vsa pojasnjena varianca je posledica napake) in 1 (popoln učinek ali popolna predikcija – vsa pojasnjena varianca je posledica neodvisne spremenljivke).

Mere velikosti učinka

Te mere predvidevajo analizo razlike med povprečji, ki bi, glede na ničelno hipotezo, morala biti ničelne velikosti. Poglejmo nekatere.

Dva neodvisna vzorca

Mere velikosti učinka, ki se uporabljajo pri neodvisnih skupinah, temeljijo na standardiziranih razlikah med povprečji. Najbolj uporabljeni indeksi so: (i) Cohenov *d*; (ii) Hedgesov *g*; (iii) Glassov Δ.

Indeks d je odnos med razliko povprečij ($M_1 - M_2$) in standardno deviacijo (σ) ene od dveh skupin (Cohen, 1988):

$$d = \frac{M_1 - M_2}{\sigma} \tag{15}$$

Ta oblika enačbe naj bi se uporabljala samo za homogene variance. Ker homogenost ni vedno zagotovljena, se v praksi (Rosnow in Rosenthal, 1996) namesto σ uporablja skupna standardna deviacija (*pooled*) (σ_p, Cohen, 1988), ki jo dobimo s kvadratnim korenom povprečja varianc.

Izračun *d* temelji na vzorčnih podatkih. Pri vzorcih z manjšim numerusom Hedges in Olkin (1985) predlagata uporabo popravljene enačbe:

$$d_c = d\left[1 - \frac{3}{4(n_1 + n_2) - 9}\right] \tag{16}$$

Indeks *d* lahko izračunamo tudi na podlagi vrednosti Studentovega *t* za dva neodvisna vzorca (Rosenthal in Rosnow, 1991):

$$d = \frac{t\,(n_1 + n_2)}{\sqrt{gl \times n_1 \times n_2}} \tag{17}$$

Indeks *d* lahko izračunamo iz *F* testa za univariatno analizo z dvema skupinama ali iz koeficienta korelacije *r* (Friedman, 1968):

$$d = \frac{2\sqrt{F_{(1,?)}}}{\sqrt{gl_{\text{err}}}} \tag{18}$$

Iz enačbe (9) lahko dobimo *d* iz katerega koli koeficienta korelacije.

Hedgesov (1981) indeks *g* dobimo z enačbo:

$$g = \frac{M_1 - M_2}{\sqrt{MS_{\text{w}}}} \tag{19}$$

ki ima v imenovalcu srednji kvadrat znotraj skupin (*within*) iz ANOVA-e za dve skupini. Hedgesov *g* lahko izračunamo iz *t* za neodvisne vzorce (Rosenthal in Rosnow, 1991):

$$g = \frac{t\sqrt{n_1 + n_2}}{\sqrt{n_1 n_2}} \tag{20}$$

Glassov (1976) Δ indeks je odnos med razliko povprečij eksperimentalne in kontrolne skupine ($M_1 - M_2$) in standardno deviacijo kontrolne skupine (le-ta je privzeta kot najbolj podobna populacijski):

$$D = \frac{M_1 - M_2}{\sigma_{\text{c}}} \tag{21}$$

Dva odvisna vzorca

Za izračun indeksa *d* pri ponovljenih merjenjih na istem vzorcu bi lahko uporabili: (i) vrednost *t* za ponovljena merjenja za pridobitev Cohenovega indeksa *d* (Rosenthal, 1991 – vendar bo vrednost indeksa *d*, dobljenega iz *t* za ponovljena merjenja, ki vključuje korelacijo, vedno večja od vrednosti *d*, dobljene iz *t* za neodvisne vzorce); (ii) standardni deviaciji obeh originalnih povprečij (prim. Dunlap, Cortina, Vaslow in Burke, 1996 – vendar tukaj napaka ne teži k izničenju, ampak k seštevanju). Enak problem bi se pojavil, če bi se odločili za oceno indeksa *d* z uporabo testa *F*, izhajajoč iz ANOVA-e za ponovljena merjenja za dve skupini. Alternativo predstavlja Cohenov (1988) indeks *f*:

$$f = \frac{m_y}{s_y} \tag{22}$$

kjer je m_y povprečje razlik med povprečji in je s_y dobljen z enačbo:

$$s_y = \sqrt{s_A^2 + s_B^2 - 2\, r\, s_A\, s_B} \tag{23}$$

Drugi indeksi velikosti učinka

Obstajajo tudi drugi indeksi velikosti učinka. Cohenov (1988) indeks *q* uporabljamo, da bi dobili velikost učinka pri razlikah med koeficienti korelacije. Dobimo ga z naslednjo enačbo:

$$q = z_1 - z_2 \tag{24}$$

kjer so *r* spremenjeni v točke *z* s sledečo enačbo:

$$z = \frac{1}{2}\log_e \frac{1+r}{1-r} \tag{25}$$

Indeks *h* (Cohen, 1988) uporabljamo za velikost učinka pri razlikah med proporci. Izmerimo neko dihotomno odvisno spremenljivko na dveh neodvisnih vzorcih in dobimo za vsakega proporc uspeha (P_1 in P_2). Uporabili bomo nelinearno transformacijo *P*, kot npr. arcus sinus:

$$f = 2\arcsin\sqrt{P} \tag{26}$$

Indeks *h* torej dobimo z enačbo:

$$h = f_1 - f_2 \tag{27}$$

Indeks *w* uporabljamo za velikost učinka za χ^2 test pri preverjanju prileganja porazdelitve in pri analizi kontingenčnih tabel. V prvem primeru imamo frekvence kategorij (empirične) in jih želimo primerjati s frekvencami iste skupine kategorij, za katere poznamo porazdelitev na osnovi H_0 (pravokotna ali normalna porazdelitev itd). V drugem primeru so frekvence računane hkrati na več ravneh več spremenljivk, za katere H_0 predvideva, da so med seboj neodvisne. Za vsako celico imamo dva proporca, ki ju predvidevata ničelna in alternativna hipoteza. Indeks *w* meri diskrepanco med proporcema v celicah:

$$w = \sqrt{\sum_{i=1}^{m} \frac{(P_{1i} - P_{0i})^2}{P_{0i}}} \tag{28}$$

kjer je P_{0i} proporc v i-ti celici, ki ga predvideva ničelna hipoteza, P_{1i} proporc v i-ti celici, ki ga predvideva alternativna hipoteza in m število celic.

Indeks f (Cohen, 1988) je mera velikosti učinka pri ANOVA-i in ANCOVA-i. Lahko ga pojmujemo kot razširitev indeksa d na primere, v katerih imamo več kot dve skupini. Lahko ga izračunamo z naslednjo enačbo:

$$f = \frac{\sqrt{\frac{\sum_{i=1}^{k} (m_i - m)^2}{k}}}{s} \tag{29}$$

kjer so m_i povprečja k skupin, m pa je skupno povprečje.

Indeks f^2 (Cohen, 1988) uporabljamo za velikost učinka v multipli regresiji in v drugih multivariatnih metodah. Glede na to, da te temeljijo na F porazdelitvi, je v resnici indeks f^2 vsebinsko enak že obravnavanemu indeksu f. Za izračun indeksa f^2 se uporablja naslednja enačba:

$$f^2 = \frac{\mathrm{PV}_{\mathrm{eff}}}{\mathrm{PV}_{\mathrm{err}}} \tag{30}$$

kjer je PV_{eff} delež variance, ki je posledica obravnave, PV_{err} pa rezidualna varianca.

Rosenthalova in Rubinova BESD

Rosenthal in Rubin (1982) sta predlagala metodo za interpretacijo velikosti učinka, imenovano *Binomial Effect Size Display* (BESD), z namenom, da bi postala očitna sprememba v deležu uspeha (izboljšanje, ozdravitev, izboljšanje samovrednotenja), ki je posledica obravnave. Metoda predpostavlja možnost dihotomizacije odvisne spremenljivke, kar pojasnjuje njen uspeh v raziskavah, ki ocenjujejo učinkovitost terapevtskih obravnav ali psihopedagoških metod, pa tudi njen uspeh v metaanalizah. Za izračun ustreznih vrednosti prirastka uspeha izhajamo iz točkovno biserialnega koeficienta r in koeficienta r^2. Delež uspeha v kontrolni skupini (ali v primeru samo enega vzorca delež uspeha pred ali v odsotnosti obravnave) in delež uspeha v eksperimentalni skupini (ali v primeru enega vzorca delež uspeha po

obravnavi) sta dana po naslednjih enačbah:

$$0{,}50 - \frac{r}{2} \tag{31}$$

$$0{,}50 + \frac{r}{2} \tag{32}$$

Tabela 1 prikazuje prirastek uspehov na osnovi vrednosti r in r^2.

Pojasnimo metodo z enim primerom. Predstavljajmo si, da delamo raziskavo za preverjanje učinka kirurškega posega na preživetje pacientov, ki imajo resno kardiopatijo. Točkovno biserialni koeficient korelacije, izračunan na zbranih podatkih, znaša $r_{pb} = 0{,}33$. Če v enačbah 31 in 32 zamenjamo r s to vrednostjo, dobimo naslednji vrednosti:

$$0{,}50 - \frac{0{,}33}{2} = 0{,}50 - 0{,}165 = 0{,}335$$

$$0{,}50 + \frac{0{,}33}{2} = 0{,}50 + 0{,}165 = 0{,}665$$

To pomeni, da delež uspeha, torej verjetnost preživetja pacientov, naraste s 33,5 % na 66,5 % za tiste, ki so izpostavljeni posegu.

Metodo podpira možnost pretvorbe najbolj uporabljenih statističnih testov v r, še bolj pa jo podpira enostavnost spremembe mer velikosti učinka v BESD v metaanalizah (prim. Lyons in Woods, 1991). Poleg uporabe pri dihotomnih odvisnih spremenljivkah (uspeh *vs.* neuspeh) lahko BESD uporabljamo tudi pri odvisnih

Tabela 1. *Deleži prirastka uspehov, izhajajoč iz nekaterih vrednosti r in* r^2.

r	r^2	Prirastek deleža uspeha		Razlika v deležu uspeha
		od	do	
0,10	0,01	0,45	0,55	0,10
0,20	0,04	0,40	0,60	0,20
0,30	0,09	0,35	0,65	0,30
0,40	0,16	0,30	0,70	0,40
0,50	0,25	0,25	0,75	0,50
0,60	0,36	0,20	0,80	0,60
0,70	0,49	0,15	0,85	0,70
0,80	0,64	0,10	0,90	0,80
0,90	0,81	0,05	0,95	0,90
1,00	1,00	0,00	1,00	1,00

spremenljivkah, ki so na metrični lestvici, če je varianca v skupinah s skoraj enakim numerusom približno enaka (glej Rosenthal, Rosnow in Rubin, 2000; za previdnost pri uporabi metode glej Thompson and Schumacker, 1997).

Interpretacija velikosti učinka

Ne obstaja splošno sprejeto pravilo za interpretacijo velikosti učinka. V odvisnosti od problemov raziskave se lahko zgodi, da je velik učinek irelevanten, majhen učinek pa pomemben (Rosenthal, 1993). Literatura ponuja določene referenčne vrednosti, ki pa jih moramo uporabljati previdno.

Referenčne vrednosti za posamezne indekse

Tabela 2 povzema referenčne vrednosti za veliko indeksov velikosti učinka (prim. Cohen, 1988, 1992a, 1992b; Cooper in Findley, 1982; Haase, Waechter in Solomon, 1982; Sedlmeier in Gigerenzer, 1989). Cohen (1988) interpretira velikost učinka na osnovi prekrivanja porazdelitev dosežkov dveh vzorcev (eksperimentalnega in kontrolnega). Transformacijo *d* lahko interpretiramo s standardno normalno porazdelitvijo.

Narejenih je bilo nekaj poskusov, da bi dobili kritične točke ali razpone za interpretacijo vrednosti indeksov. Vendar pa se je dobro izogniti rigidni uporabi le-teh. Velikost učinka je treba najprej interpretirati v odnosu do rezultatov predhodnih raziskav (Wilkinson in TFSI, 1999; Welkowitz, Ewen in Cohen, 1982; Thompson, 2002). Kot pravi Thompson (2001): če je tu rigidnost enaka kot pri uporabi 0,05 pri metodi TNH, naredimo enako napako na drugi lestvici.

Tabela 2. *Pomen velikosti učinka pri najpogosteje uporabljenih indeksih.*

		Velikost učinka		
Tip analize	Test	Majhen	Srednji	Velik
t test za neodvisne vzorce	d	0,20	0,50	0,80
t test za odvisne vzorce	f	0,10	0,25	0,40
ANOVA	f	0,10	0,25	0,40
	ω^2	0,01	0,06	0,15
ANCOVA	f	0,10	0,25	0,40
Korelacija	r	0,10	0,30	0,50
Multipla regresija in multivariatne metode	f^2	0,02	0,15	0,35
Hi kvadrat	w	0,10	0,30	0,50
Razlike med koeficienti korelacije	q	0,10	0,30	0,50
Razlike med proporci	h	0,20	0,50	0,80

Intervali zaupanja

Eden od predlaganih načinov za integracijo informacij metode TNH so *intervali zaupanja*. Širino intervala določa verjetnost, imenovana *koeficient zaupanja*, da interval vsebuje parameter. To pomeni, da bolj kot se poveča velikost vzorca *n*, bolj se zmanjša meja napake, in se torej poveča natančnost ocene. Bolj kot se zmanjša koeficient zaupanja, bolj se pri nespremenjenem *n* interval zoži. S povečanjem koeficienta zaupanja se poveča tudi interval, ki pa postane malo informativen.

Intervale zaupanja lahko izračunamo za vsak parameter: povprečje, varianco, proporc, mediano itd. Za povprečje: ne glede na porazdelitev izvorne populacije, po centralnem limitnem izreku pri *n*, večjem ali enakem 30, teži vzorčna porazdelitev povprečij k normalni obliki. Če je $\alpha = 0{,}05$, je kritična vrednost *z* (pri dvostranskem testiranju) ±1,96. Če σ ne poznamo, jo ocenimo na osnovi standardne deviacije vzorcev (*s*) in vzorčna porazdelitev povprečij dobi obliko Studentove *t* porazdelitve z *n*-1 stopnjami svobode. Porazdelitev za varianco je χ^2 porazdelitev: $\chi^2_{n-1} = (n-1)\hat{s}^2 / \sigma^2$

Če moramo oceniti proporc π, s katerim je določena lastnost prisotna v populaciji, je porazdelitev naključne spremenljivke $\hat{p}$ normalna standardizirana porazdelitev, s povprečjem $\mu_p = np$ in standardno napako proporcev $\sigma_p = \sqrt{pq/n}$, če je *n* dovolj velik. Če je *n* vzorca med 1 in 10, je dobro uporabiti metodo grafičnega postopka, ki sta jo predlagala Clipper in Pearson (1934). Do sedaj predstavljeni postopki se nanašajo na vzorec, ki smo ga vlekli iz neskončne populacije, medtem ko moramo za končno populacijo ocenjeno varianco popraviti, saj bi bila sicer večja od dejanske (prim. Burstein, 1975).

Tudi za velikost učinka je ustrezna uporaba intervalov zaupanja. Če interval zaupanja vključuje vrednost 0, je to enakovredno temu, da imamo rezultat, ki ni statistično pomemben. Nasprotno, če vrednost 0 ni vključena v intervalu, je to statistično pomembno. Primer Cohenovega *d* je posebno lahko rešljiv. Ta indeks je namreč normalno porazdeljen in ima standardno deviacijo enako:

$$s_d = \sqrt{\frac{n_1 + n_2}{n_1 n_2} + \frac{d^2}{2(n_1 + n_2)}} \tag{33}$$

kjer sta n_1 in n_2 numerusa eksperimentalne in kontrolne skupine (Hedges in Olkin, 1985). Lahko torej določimo desne in leve *z* vrednosti, ki ustrezajo prej določeni ravni α. Na žalost se velika večina ostalih pregledanih indeksov ne porazdeljuje normalno, kar implicira nujnost uporabe drugih metod. V teh primerih uporabimo lahko dostopno programsko opremo, kot je SPSS ali STATISTICA (glej Smithson, 2001; Steiger in Fouladi, 1997). Glede tega problema glej Fidler in Thompson (2001).

Za neparametrične teste vzemimo za primer Kendallov koeficient korelacije rangov τ (tau), na katerega za razliko od *r* in ρ ne vpliva nenormalnost porazdelitve vzorcev, zaradi česar ga raziskovalci pogosto raje uporabljajo (Long in Cliff, 2004). Če bi vmesne korake izpeljali po šolski metodi – na dolgo (prim. Siegel, 1956), bi videli, da interval dobimo z naslednjo enačbo:

$$-z_{\text{krit}}\frac{\sqrt{2(2n+5)}}{3\sqrt{n(n-1)}} \le t \le z_{\text{krit}}\frac{\sqrt{2(2n+5)}}{3\sqrt{n(n-1)}} \tag{34}$$

Tako se privzame, da je $n!$ možnih razvrstitev rangov enako verjetnih. Če ocenjujemo τ »parametrično« (Cliff in Charlin, 1991), moramo zamenjati vrednost z_{krit} z vrednostjo t_{krit}. Vendar se zdi, da te simulacije niso prinesle velikih prednosti (prim. Long in Cliff, 2004).

Moč statističnega testa

Moč, vezana direktno na napako drugega tipa, je zmožnost prepoznati učinek, kadar pride do njega. Dobimo jo z 1 – β. Pomembno se je zavedati, da določiti α implicira določiti hkrati tudi β. Ti dve verjetnosti sta med seboj povezani. H_0 in H_1 namreč ustrezata dogodkom z »nepopolno diskriminativnostjo«, torej se ustrezni porazdelitvi delno prekrivata. Npr. pri primerjavi povprečij z določitvijo α določimo področje zavrnitve v porazdelitvi H_0 (pri enostranskem testiranju je to področje desno od odločitvene osi). Vendar odločitvena os preseka tudi porazdelitev, ki ustreza H_1 in področje na levi strani osi ustreza β. Parametri porazdelitev H_0 in H_1 ter razdalja med njima (ki je enaka Cohenovemu *d*) so podani, in če hočemo najti optimalen β, moramo spremeniti velikost vzorca. Če že poznamo μ_0, μ_1, σ_0 in σ_1, torej povprečja in standardne deviacije univerzumov, ki ustrezajo H_0 in H_1, bodo področja zavrnitve ničelne in alternativne hipoteze ustrezala odgovarjajočim *z* točkam (enosmernim). Če bomo torej določili α na 0,05 in β na 0,2, bodo ustrezne točke *z* 1,645 in -0,845. Lahko bomo torej postavili naslednji sistem enačb, katerega neznanki sta *M* (kritično povprečje, nad katerim zavrnemo ničelno hipotezo in sprejmemo alternativno, ki ustreza $\alpha = 0{,}05$) in ravno *N*:

$$\begin{aligned} \frac{M-m}{\frac{s_0}{\sqrt{N}}} &= 1{,}645 \\ \frac{M-m}{\frac{s_0}{\sqrt{N}}} &= -0{,}845 \end{aligned} \tag{35}$$

Moč določajo trije viri: (i) raven α (in usmerjenost ničelne hipoteze); (ii) velikost vzorca; (iii) velikost učinka (razdalja med μ_0 in μ_1). Na splošno z njihovim povečanjem raste tudi vrednost moči.

Simon (1999) je predlagal uporabo neformalnega pravila, ki predvideva določitev α na 0,05 in β na 0,2. To pomeni, da bi morala biti vrednost moči najmanj 0,80, da bi bila sprejemljiva, medtem ko bi morala biti verjetnost, da naredimo napako drugega tipa, 20 %. Odnos *q* med vrednostmi β in α določa tudi težo, pripisano obema tipoma napak (Cohen, 1988). Če bi npr. imeli $\alpha = 0{,}001$ in $\beta = 0{,}50$, bi bil odnos *q* enak (0,50/0,001) = 500, kar bi pomenilo enako kot trditi, da je zmotna zavrnitev ničelne hipoteze 500-krat hujša kot njeno sprejetje. Če sledimo Simonovemu nasvetu, bi odnos *q* ustrezal vrednosti (0,20/0,005) = 4, kar pomeni, da je zavrnitev ničelne hipoteze zaradi napake štirikrat hujša od njenega zmotnega sprejetja.

Na moč vpliva tudi *usmerjenost ničelne hipoteze* (Cohen, 1988). V normalni porazdelitvi z $\alpha = 0{,}05$ je kritična vrednost *z* pri enosmernem testiranju 1,654 in pri dvosmernem testiranju 1,96. Z dvosmerno α bo imel test manjšo moč določanja učinka v primerjavi z načrtom raziskave z enako vrednostjo α, ampak enosmerno, pod pogojem, da gre rezultat v predvideni smeri.

Povezanost med močjo in *velikostjo vzorca* nas pripelje do problema zanesljivosti dobljenih rezultatov. Večji kot je numerus vzorca, manjša je napaka in večja je zanesljivost ali natančnost rezultatov. *Velikost učinka* najbolj pomembno določa moč. Večja kot je velikost učinka pri enaki ravni α in velikosti vzorca, večja je moč testa (Cohen, 1988). Analiza moči se nanaša še na en temeljni cilj velike večine raziskav: ponovljivost (Ottenbacher, 1996). Muller in Lavange (1992) sta odkrila pogoste napake raziskovalcev pri izbiri ustrezne analize moči glede na uporabljen načrt.

A priori analizo moči uporabljamo za določitev velikosti vzorca, ki je potrebna, da jamči določeno moč testa. *A posteriori analizo moči* uporabljamo, da bi spoznali moč testa, ki je bil uporabljen. *A priori* ali *a posteriori* analize moči se je možno lotiti z uporabo tabel (Kraemer in Thiemann, 1987; Cohen, 1988; Lipsey, 1990; Zar, 1996) ali statističnih programov. V apriornih analizah je treba poznati α, moč in velikost učinka, bodisi na osnovi poznanega iz literature bodisi na osnovi posebnih vidikov raziskave. V aposteriornih analizah je postopek enak, saj poznamo α, velikosti vzorca in učinka, ki so vsi vnaprej določeni (izbor α in *N* pri definiciji načrta raziskave, velikost učinka pa izračunamo na zbranih podatkih). V obeh primerih kontroliramo križanje vrednosti, podanih v tabelah, ki je relevantno za pridobitev moči testa. Predvsem Cohen (1988) predstavlja veliko tabel za analizo moči *a priori* in *a posteriori* z veliko statističnimi testi, za katere navaja referenčne vrednosti glede na tri različne ravni α (0,01, 0,05 in 0,10), z izpeljankami na osnovi uporabljenega testa.

Obstaja veliko programov za analizo moči: (i) *Power and Precision* (Borestein, Cohen in Rothstein, 1997), v SPSS kot *Sample Power* (ne za ponovljena merjenja in za MANOVA-o); (ii) *PASS* (NCSS Statistical Software, 1999), tudi za ponovljena merjenja, a ne za MANOVA-o; (iii) *G*Power* (2000), ki je, poleg tega, da je brezplačen, tudi preprost za uporabo; (iv) *SAS Macro* (Friendly, 1991), imenovan *mpower*, samo

za aposteriorne analize. Tudi dva zelo uporabljena programa ponujata analize moči: (i) *SPSS*, samo aposteriorno, za ANOVA-o; (ii) *Systat 10*, tako za apriorne kot za aposteriorne analize, za veliko večino statističnih testov.

Prevzorčenje

Obstajajo alternativni načini spopadanja s problemom metode TNH, ki temeljijo na tehnikah simulacije, imenovani na splošno načini ponovnega vzorčenja (*resampling*) (Simon, 1969; Simon, Atkinson in Shevokas, 1976). V resnici ti segajo nekaj desetletij nazaj, k Fisherju (1935), ki je zasnoval t. i. eksaktne teste. Še pred njim pa je v dvajsetih letih to idejo zasnoval von Mises (1928; prim. von Mises, 1964).

Randomizacijski test

Znan je Fisherjev t. i. test eksaktne verjetnosti za kontigenčne tabele, ki ga ponavadi računamo s hipergeometričnim koeficientom. Logiko, ki je v osnovi tega testa, lahko posplošimo. Imamo vzorec **x**: $\{x_1, x_2, ..., x_n\}$, sestavljen iz vrednosti, ki so na metrični ravni. Pogledamo, ali je tistih nad povprečjem več (ali manj) od tistih pod povprečjem. Ni pomemben odklon od povprečja, ampak samo frekvenca nadpovprečnih (ali podpovprečnih) vrednosti. H_0 predvideva, da je verjetnost, da bomo imeli pozitiven dogodek p(P), enaka verjetnosti negativnega dogodka p(N), in glede na to, da je to Bernoullijev univerzum, sta obe verjetnosti enaki 0,5. Verjetnost, povezano s pojavljanjem k pozitivnih in $n - k$ negativnih dogodkov, dobimo z naslednjo enačbo:

$$\binom{n}{k}\left(\frac{1}{2}\right)^{k}\left(\frac{1}{2}\right)^{n-k} \tag{36}$$

Fisher (1935) meni, da mora raziskovalec izračunati vse verjetnosti, povezane z vsemi možnimi izidi eksperimenta. Te verjetnosti mora urediti po vrstnem redu pripisovanja k izidom, ki so progresivno vedno bolj oddaljeni od ničelne hipoteze. Najti mora točko, v kateri je verjetnost opazovanega dogodka. Ta razdeli na dva dela prostor verjetnosti vseh izidov, ki se jih da realizirati, in sicer je na eni strani opazovana verjetnost in vse verjetnosti, povezane z bolj neugodnimi izidi za ničelno hipotezo (vsota, ki jo moramo podvojiti za dvosmerne hipoteze), na drugi strani pa so verjetnosti, ki prištete k prejšnjim dajo 1.

Primer lahko razširimo na bolj kompleksne situacije. Vzemimo, da imamo dva neodvisna vzorca in dve ravni obravnave, in da lahko za primerjavo povprečij uporabimo t. Vzorci so majhni, učinek in moč nista velika in zato tudi pomembna vrednost t pušča dvome o možnosti dokončne odločitve. Zato kot naključno spremenljivko obravnavamo uporabljeno statistiko (v našem primeru t), tako da s podatki, ki jih imamo na razpolago,

povečamo pare vzorcev za primerjavo. Tako ustvarimo veliko *t*-jev, preučimo njihovo porazdelitev in pogledamo, kje se nahaja na začetku opazovani *t*. Vzamemo torej prvi vzorec in zamenjamo eno njegovo vrednost z vrednostjo iz drugega vzorca. Tako bomo imeli dva nova vzorca, za katera bomo lahko izračunali *t*. Nato sistematično zamenjamo s prvo vse ostale vrednosti drugega vzorca, eno naenkrat, tako da vsako vsakokrat postavimo na svoje mesto in vsakokrat izračunamo ustrezni *t*. Potem naredimo enako z drugo, tretjo vrednostjo itd. do zadnje vrednosti. Če ima vsak vzorec *n* oseb, bomo dobili populacijo z n^2 statistikami, katerih porazdelitev bomo preučevali in znotraj nje pogledali intervale zaupanja ter kje se nahaja originalni *t*.

Neeksaktni testi randomizacije uporabljajo Fisherjevo intuicijo, ampak omejujejo število vzorcev, uporabljenih za preučevanje porazdelitve uporabljene statistike. Dovolj naj bi bilo 1000 vzorcev (vzetih slučajno), če hočemo postaviti intervale zaupanja na ravni 0,05, in vsaj 5000 vzorcev, če hočemo preiti na raven 0,01 (Manly, 1997).

Cross-validation (navzkrižna validacija)

Prvo praktično realizacijo povedanega predstavlja metoda *cross-validation*, ki so jo razvijali od štiridesetih let naprej na korelacijskih načrtih in ki ima izvor v študijah zanesljivosti (*reliability*) testov. Prvo sistematizacijo metode je podal Mosier (1951). V enostavni navzkrižni validaciji originalen vzorec naključno razdelimo na dva podvzorca enake velikosti. Izračunamo parametre regresijske premice v prvem podvzorcu, s katerimi izračunamo napovedane vrednosti drugega podvzorca. Tako dobimo koeficient navzkrižne validacije – koeficient korelacije med dejanskimi in napovedanimi vrednostmi. V dvojni navzkrižni validaciji izračunamo regresijsko premico in napovedane vrednosti v obeh podvzorcih. Napovedane vrednosti v prvem podvzorcu izračunamo prek regresijske premice, dobljene na drugem vzorcu, in nasprotno naredimo za drugi podvzorec. Koeficient navzkrižne validacije je torej koeficient korelacije med opazovanimi in napovedanimi vrednostmi za celoten vzorec. Pomemben napredek pa pomeni uporaba večkratne navzkrižne validacije. V tem primeru vzorec razdelimo na dva slučajna podvzorca. Postopek nadaljujemo kot pri dvojni navzkrižni validaciji in ga nato velikokrat ponovimo. Koeficient navzkrižne validacije je povprečje dobljenih korelacijskih koeficientov. Poleg tega lahko preučujemo porazdelitev dobljenih koeficientov.

Navzkrižna validacija ima še vedno določene privržence, kot je Thompson (1993). Vendar pa je zaradi verjetnosti artefaktov, ki so posledica majhnosti vzorcev, veliko raziskovalcev previdnih in predlagajo raje uporabo tehnike *jackknife* (npr. Ang, 1998).

Jackknife

V tehniki *jackknife*, ki sta jo predlagala Quenouille (1949) in Tukey (1958), imamo vzorec **x**, na katerem računamo našo statistiko θ. Če vzorec **x** sestavlja *n* vrednosti, ustvarimo *n* podvzorcev vzorca **x**, vsakega z $n - 1$ vrednostmi, kot če bi s

pipcem izrezali (jackknife) vsakokrat po eno vrednost iz originalnega vzorca. Na teh vzorcih računamo statistiko θ', in za vsak vzorec dobimo t. i. psevdovrednost (*PV*):

$$PV = n \times q - (n-1)\breve{q} \tag{37}$$

Tako bomo imeli en vzorec z *n* psevdovrednostmi, katerega povprečje je *jackknifed coeficient* (*jc*). Če *jc* delimo s standardno napako ocene, dobimo *t*, katerega pomembnost bomo lahko preverjali.

Tehnika *jackknife* minimizira učinek osamelcev. Diskusija o primernosti izločitve teh točk iz podatkov je vedno aktualna, tudi zato, ker je proizvedeno popačenje tesno povezano z velikostjo vzorca (prim. Miller, 1991). *Jackknife* nam ta problem reši brez stranskih učinkov, ker učinek osamelcev izgine v velikem številu uporabljenih vzorcev.

Bootstrap

Bistvena razlika med tehnikama *bootstrap*, ki sta jo neodvisno razvila Simon (1966; prim. Simon, 1969) in Efron (1979), in *jackknife* je v tem, da pri slednji uporabljamo vzorčenje brez vračanja, pri tehniki *bootstrap* pa vzorčenje z vračanjem. Pri tehniki *bootstrap* vsak element, ki ga izvlečemo, takoj nato spet vrnemo v originalno celoto, zato se lahko zgodi, da je ponovno izvlečen. To znatno poveča univerzum možnih vzorcev, medtem ko smo pri tehniki *jackknife* omejeni na število vzorcev, ki je lahko zmanjšano.

Po našem mnenju je tehnika *bootstrap* zelo učinkovita pri neodvisnih vzorcih, medtem ko je lahko pri paroma odvisnih vzorcih precej manj zanesljiva kot randomizacija. To velja predvsem za majhne vzorce (ki po navadi psihologa bolj zanimajo), medtem ko moramo pri velikih vzorcih na problem verjetno gledati na drugačen način.

Zaključki

Naše delo lahko zaključimo na sledeč način: v inferenčnih analizah upoštevajmo okvir, ki je na razpolago raziskovalcu, v vsej njegovi širini in kompleksnosti ter (skladno s priporočili APA) naj pomembnost vedno spremljajo indeksi, o katerih smo govorili (velikost, intervali zaupanja in moč). K temu dodajamo priporočilo, naj analizo spremlja katera od mer prevzorčenja.

Dodatno, četudi le na kratko, moramo preudariti še tri probleme: (i) v kolikšni meri so to razpravo ustvarili prav raziskovalci; (ii) kako učimo inferenčno statistiko; (iii) ali je res treba prekomerno uporabljati te tehnike za analizo podatkov.

Metoda TNH v objavljenih raziskavah

Analize v glavnih psiholoških revijah so pokazale, da so raziskovalci neobčutljivi na problem, in predvsem to, da v zadnjih letih, kljub obširni razpravi o tej temi, ni prišlo do izboljšanja. Skoraj trideset let od prve pionirske raziskave Cohena (1962) v reviji *Journal of Personality and Social Psychology* je analiza Sedlmeierja in Gigerenzerja (1989) dala skoraj enake rezultate: videti je, da raziskovalci (in uredništva znanstvenih revij) v tridesetih niso opazili problema in so nadaljevali z delom na enak način. Pozornost raziskovalcev je bila predvsem osredotočena na moč uporabljenih statističnih testov, vendar je ta običajno zelo nizka, s sredino manjšo od 0,5, torej, kot poudarja Hunter (1997), na ravni slučaja. Pomembnost je splošno privzeta kot potrditev alternativne hipoteze. Pri tem raziskovalci ne razmišljajo o moči (zgoraj omenjene vrednosti so rezultati post hoc analiz, ki so jih izvedli citirani raziskovalci in ne avtorji raziskav), izjemo pa predstavlja prisotnost drugih indeksov, od tistih glede velikosti učinka do intervalov zaupanja.

Tudi v Italiji je imela diskusija glede metode TNH slab odmev. En prispevek (ki se nanaša na sociologijo) je dal Pisati (2002), na kongresu AIP iz Barija pa je zanimanje sprožilo poročilo Angolija (2002). V juniju 2003 je bil v Firenzah en dan posvečen obravnavi tega problema. Srečanja so se udeležili skoraj vsi italijanski učitelji disciplinskega združenja M.PSI/03, ki združuje psihometrijo in sorodne predmete. Diskusija je bila omejena na krog metodologov.

Ali je to, da raziskovalci ne zaznajo problema, lahko odvisno od načina, na katerega so se med usposabljanjem naučili analizirati podatke raziskav? To bomo zdaj skušali dognati.

Učenje statistike bodočih psihologov in metoda TNH

Kot opažata Haller in Krauss (2002), bi bilo »učenje metode TNH upravičeno le, če so študenti zmožni razumeti *pomen* tistega, kar počnejo« (str. 2). Vendar do zdaj opravljene raziskave kažejo, da po povprečnih predavanjih iz statistike študent nima najmanjše ideje o tem, kaj TNH dejansko pomeni (Falk in Greenbaum, 1995; Gigerenzer in Krauss, 2001) in je v najboljšem primeru osvojil postopek računanja (Gigerenzer, 1998).

Na žalost pa, kot je pokazal Oakes (1986), akademskim psihologom pogosto stvari niso bolj jasne kot študentom. Haller in Krauss (2002) sta želela ugotoviti, ali se taka zmotna razumevanja prenesejo pri učenju, prek napak ali zmede v priročnikih ali celo prek zmotnih razumevanj pri učiteljih. Glede priročnikov tako Haller in Krauss kot Sedlmeier in Gigerenzer (1989) predstavljajo »muzej grozot«, ki ne izvzema niti produktov take »svete krave« statistike, kot je Nunnally (1975). Le-ta je na treh straneh (194–196) podal kar osem različnih definicij pomembnosti, ki pa so vse napačne. Kar se tiče učiteljev statistike, predvsem praktikov, so podatki, ki jih

predstavljajo Haller in Krauss za Nemčijo in Oakes za ZDA, še bolj brezupni. Za Italijo podatkov nimamo na razpolago, vendar moramo upoštevati, da je razširitev univerzitetnih programov psihologije na nacionalnem območju močno povečala potrebo po psihometričnem poučevanju, ki se pogosto zaupa mladim izvrstnim raziskovalcem, ki pa žal nimajo zadovoljive specifične izobrazbe.

Je metoda TNH obvezna pot?

Na tej točki se zdi, da bi lahko sklenili tudi takole: TNH se zdi skoraj obvezna pot, ker je njena logika nedvoumna in linearna ter so za raziskovalca s slabim osnovnim znanjem kompleksnejši pristopi težko razumljivi. Poleg tega ta postopek obravnave problema dobiva neprestane podkrepitve. Tako narejen članek znanstvena revija prej sprejme, pri predstavitvi rezultatov na določenem kongresu uporaba te metode ne vzpodbudi ugovorov, temveč celo pohvale kolegov.

Vendar, ali je ta način postopanja nujen? Povedali smo že, da v splošnem znanost meni, da je ta način obdelovanja podatkov nenavaden in poleg psihologije ni veliko disciplin, v katerih ima privilegiran položaj. Toda tudi psihologija je že od začetka 20. st. naprej razvijala alternativne načine za obdelovanje svojih podatkov, npr. faktorsko analizo. Vendar je ta razumljena kot raziskovalna analiza, mati vseh multivariantnih analiz in na splošno vseh tistih analiz, ki temeljijo na strukturah kovariance, od konfirmatornih analiz do modelov strukturnih enačb. Na osnovi tega bi lahko rekli, da te analize zahtevajo velike vzorce, v primerjavi s TNH, ki lahko dela na majhnih vzorcih (čeprav ne tako majhnih, kot jih pogosto vidimo objavljene). Res je, vendar pa obstaja skupen vidik teh in drugih vedno bolj uporabljenih analiz (kot npr. log-linearni modeli ali logistična regresija), pri katerem se je vredno ustaviti, in sicer preverjanje modelov.

Če izvzamemo posebna področja (kot je npr. psihofizika), je preverjanje modelov glavna aktivnost pri analizi podatkov večjega dela naravoslovnih znanosti. Znanstvenik na osnovi literature ali lastnih inovativnih hipotez in na koncu tudi po pregledu podatkov predpostavlja, da je narava urejena tako, da podatki sledijo določenemu modelu in da lahko njihovo pojavljanje določimo na osnovi določene matematične funkcije ali skupka funkcij. Model je lahko Fechnerjev zakon, a tudi strukturna enačba pri faktorski analizi ali pa skupek enačb multiple regresije, ki v modelu strukturnih enačb povezujejo med seboj eksogene in endogene spremenljivke ter napake.

Preverjanje modelov v priročnikih ni zelo prisotno in je omejeno na povzete informacije o oceni *goodness of fit* testa prileganja podatkov predpostavljeni funkciji. Ta nepozornost pripelje raziskovalce do napak pri formulaciji modelov, ki jih je treba preveriti. Najbolj pogosta napaka je, da raziskovalci poskušajo razložiti pojav s funkcijo, ki vsebuje preveč parametrov. To povzroči, da funkcija natančno sledi modelu in je ne moremo zavrniti. Vendar pa zaradi prevelike količine parametrov funkcija izgubi svoj pomen (Forster in Sober, 1994). Toda pomislimo na to, kako koristno bi bilo boljše poznavanje različnih pristopov preverjanja modelov – kot npr. tistega, ki ga je trideset

let nazaj predlagal Akaike (1973; prim. Bozdogan, 1987) – ki dovoljujejo na informaciji utemeljeno preverjanje (po Kullback in Leibler, 1951), kot npr. *razdaljo* med opazovanimi podatki in modelom.

Vse omenjeno pa se v psihologijo vpeljuje s težavo. Veliko lažje je zastaviti raziskave v smislu eksperimentalne in kontrolne skupine ali podobnih načrtov ter interpretirati podatke s pomočjo varnih postopkov metode TNH. Prednost je tudi ta, da celo ni nujno, da razumemo, kar delamo! (klicaj dodala prev.)

Literatura

Aaron, B., Kromrey, J. D. in Ferron, J. M. (1998, November). *Equating r-based and d-based effect size indices: Problems with a commonly recommended formula.* Prispevek, predstavljen na letnem srečanju Florida Educational Research Association, Orlando, FL. (ERIC Document Reproduction Service No. ED 433 353).

Abelson, R. P. (1997). On the surprising longevity of flogged horses: Why there is a case for the significance test. *Psychological Science, 8*, 12–15.

Agnoli, F. (2002). *Come presentare i risultati di ricerche sperimentali: ambiguita, errori, controversie*. Congresso AIP, Sez. Psicologia Sperimentale, Bari.

Akaike, H. (1973). Information Theory and an extension of the Maximum Likelihood Principle. V B. N. Petrov in F. Csaki (ur.), *2nd International Symposium on Information Theory* (str. 267–281). Budapest: Akademiai Kiado.

American Psychological Association (1994). *Publication Manual of the American Psychological Association* (4. izd.). Washington, DC: Author.

American Psychological Association (2001). *Publication Manual of the American Psychological Association.* Washington, DC: Author.

Ang, R. P. (1998). Use of Jackknife statistic to evaluate result replicability. *Journal of General Psychology, 125*, 218–228.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*, 423–437.

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.

Bland, J. M. (2000). *An Introduction to Medical Statistics.* New York: Oxford University Press.

Bonferroni, C. E. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni R. Istituto Superiore Scienze Economiche e Commerciali di Firenze*, *8*, 3–62.

Borestein, M., Cohen, J. in Rothstein, H. (1997). *Power and precision*. Dataxiom, Inc., [Online] Dostopno na URL: http://www.dataxiom.com.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370.

Burstein, H. (1975). Finite population correction for binomial confidence limits. *Journal of American Statistic Association*, *70*, 67–69.

Cahan, S. (2000). Statistical significance is not a "kosher certificate" for observed effects: A critical analysis of the two-step approach to the evaluation of empirical results.

Educational Researcher, *29*, 31–34.
Carver, R. P. (1978). The case against significance testing. *Harvard Educational Review, 48*, 378–399.
Carver, R. P. (1993). The case against statistical significance testing, revisited. *Journal of Experimental Education, 61*, 287–292.
Chow, S. (1996). *Statistical Significance: Rationale, Validity, and Utility*. London: Sage.
Chow, S. (1998). Précis of statistical significance: Rationale, validity, and utility. *Behavioral and Brain Sciences, 21*, 169–240.
Chow, S. (1999). In defense of significance tests. Commentary on Krüeger on social bias. *Psycoloqui, 10* (006). Dostopno na: http://www.cogsci.soton.ac.uk/cgi/psyc/newpsy? 10.006.
Cliff, N. in Charlin, V. (1991). Variances and covariances of Kendall's tau and their estimation. *Multivariate Behavioural Research, 26,* 693–707.
Clipper, C. J. in Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of binomial. *Biometrika*, *26*, 404–413.
Cohen, J. (1962). The statistical power of abnormal social psychology research. *Journal of Abnormal and Social Psychology*, *65* (3), 145–153.
Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. izd.). New York: Academic Press.
Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2. izd.). Hillsdale, NJ: Erlbaum.
Cohen, J. (1992a). A power primer. *Psychological Bulletin*, *112*, 155–159.
Cohen, J. (1992b). Statistical Power Analysis. *Current Directions in Psychological Science, 1,* 98–105.
Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997–1003.
Cooper, H. in Findley, M. (1982). Expected effect sizes: Estimates for statistical power analysis in social psychology. *Personality and Social Psychology Bulletin, 8*, 168–173.
Daniel, L. G. (1998). Statistical significance testing: A historical overview of misuse and misinterpretation with implications for the editorial policies of educational journals. *Research in the School*, *5* (2), 23–32.
Dunlap, W. P., Cortina, J. M., Vaslow, J. B. in Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1,* 170–177.
Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*, 1–26.
Falk, R. in Greenbaum, C. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology, 5,* 75–98.
Fidler, F. in Thompson, B. (2001). Computing correct confidence intervals for ANOVA fixed- and random-effects effect sizes. *Educational and Psychological Measurement, 61,* 575–604.
Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver & Boyd.
Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh: Oliver & Boyd.
Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of Royal Statistic Society, B, 17,* 69–78.
Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and

Boyd.
Forster, M. in Sober, E. (1994). How to tell when simpler, more unified, or less *ad hoc* theories will provide more accurate predictions. *British Journal for the Philosophy of Science, 45*, 1–35.
Friedman, H. (1968). Magnitude of experimental effect and a table for its rapid estimation. *Psychological Bulletin, 70,* 245–251.
Friendly, M. (1991). *SAS macro programs: mpower*. [On-line] Dostopno na URL: http: // www.math.yorku.ca/SCS/sasmac/mpower.html.
GPower (2000). [Online]. Dostopno na URL: http://www.psychologie.uni-trier.de:8000/ projects/ gpower.html.
Gigerenzer, G. (1998). We need statistical thinking, not statistical rituals. *Behavioral and Brain Sciences, 21*, 199–200.
Gigerenzer, G. in Krauss, S. (2001). Statistisches Denken oder statistische Rituale? Was sollte man unterrichten? V M. Borovcnik, J. Engel in D. Wickmann (ur.), *Anregungen zum Stochastikunterricht: Die NCTM-Standards 2000, Klassische und Bayesische Sichtweise im Vergleich* (str. 53–62). Franzbecker: Hildesheim.
Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5***,** 3–8.
Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–20.
Haase, R. F., Waechter, D. M. in Solomon, G. S. (1982). How significant is a significant difference? Average effect size of research in Counseling Psychology. *Journal of Counseling Psychology, 29,* 58–65.
Haller, H. in Krauss, S. (2002). Misinterpretation of significance: a problem students share with their teachers? *Methods of Psychological Research Online, 17(1),* 1–20.
Harlow, L. L., Mulaik, S. A. in Steiger, J. H. (ur.). (1997). *What if there were no significance tests?* Mahwah, NJ: Erlbaum.
Hays, W. L. (1963). *Statistics for psychologists*. New York: Holt, Rinehart & Winston.
Hedges, L. V. (1981). Distributional theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6,* 107–128.
Hedges, L. V. in Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego; Academic Press.
Hubbard, R. in Armstrong, J. S. (1992). Are null results becoming an endangered species in marketing? *Marketing Letters*, *3*, 127–136.
Hubbard, R. in Armstrong, J. S. (1994). Replicationss and extensions in marketing: Rarely published but quite contrary. *International Journal of Research in Marketing*, *11*, 233–248.
Hubbard, R. in Armstrong, J. S. (1997). Publication bias against null results. *Psychological Reports, 80*, 337–338.
Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317–333.
Hunter, J. E. (1997). Needed: A ban on the significance test. *Psychological Science, 8*, 3–7.
Kerlinger, F. N. (1979). *Behavioral Research: A conceptual approach*. New York, NY: Holt, Rinehart and Winston.
Kraemer, H. C. in Thiemann. S. (1987). *How Many Subjects?* London, UK: Sage Publica-

tions.
Kuhn, T. S. (1962). *The Structure of Scientific Revolutions,* Chicago: Chicago University Press.
Kullback, S. in Leibler R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics 22,* 79–86.
Levin, J. R. in Robinson, D. H. (1999). Further reflections on hypothesis testing and editorial policy for primary research journals. *Educational Psychology Review, 11*, 143–155.
Lindsay, R. M. (1994). Publication system biases associated with the statistical testing paradigm. *Contemporary Accounting Research, 11*, 33–57.
Lipsey, M. W. (1990). *Design Sensitivity: Statistical Power for Experimental Research*. Newbury Park, California: Sage.
Loftus, G. R. (1996). Psychology will be much better science when we change the way we analyze data. *Current Directions in Psychological Science, 5*, 161–170.
Long, J. S. (1983). *Covariance Structure Models*. Newbury Park, CA: Sage.
Long, J. S. in Cliff, N. (2004). Confidence intervals of Kendall's tau. *British Journal of Mathematical and Statistical Psychology, 57,* 31–41.
Lyons, L. C. in Woods, P. J. (1991). The efficacy of rational emotive therapy: A quantitative review of the outcome research.. *Clinical Psychology Review, 11,* 357–369.
Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology.* New York, NY: Chapman & Hall.
Maxwell, S. E. in Delaney, H. D. (1990). *Designing experiments and analyzing data. A model comparison perspective*. Belmont, CA: Wadsworth.
Meehl, P. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. V L. L. Harlow, S. A. Mulaik in J. H. Steiger (ur.), *What if there were no significance tests?* (str. 393–426). Mahwah, NJ: Erlbaum.
Melton, A. (1962). Editorial. *Journal of Experimental Psychology, 64*, 553–557.
Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156–166.
Miller, J. (1991). Reaction time analysis with outliers exclusion: Bias varies with sample size. *Quarterly Journal of Experimental Psychology, 43A*, 907–912.
Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement,* 11, 5–11.
Mulaik, S. A., Raju, N. S. in Harshman, R. A. (1997). There is a time and a place for significance testing. V L. L. Harlow, S. A. Mulaik in J. H. Steiger (ur.). *What if there were no significance tests?* (str. 65-115). Mahwah, NJ: Lawrence Erlbaum.
Muller, K. E. in Lavange, L. M. (1992). Power calculations for general linear multivariate models including repeated measures applications. *Journal of the American Statistical Association, 87,* 1209–1216.
NCSS Statistical Software (1999). *PASS*. [On-line] Dostopno na: http://www.ncss.com.
Neyman, J. in Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses, *Philosophical Transactions of the Royal Society of London, A 231*, 289–337.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5,* 241–301.
Nix, T. W. in Barnette J. J. (1998). The data analysis dilemma: ban or abandon. A review of

null hypothesis significance testing. *Research in the Schools, 5(2),* 3–14.
Nunnally, J. C. (1960). The place of statistics in psychology. *Education and Psychological Measurement*, *20*, 641–650.
Nunnally, J. C. (1975). *Introduction to statistics for psychology and education*. New York: McGraw-Hill.
Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. New York: Wiley.
Ottenbacher, K. J. (1996). The power of replications and replications of power. *American statistician, 50*, 271–275.
Perneger, T. V. (1998). What's wrong with Bonferroni adjustments? *British Medical Journal, 316*, 1236–1238.
Pisati, M. (2002). Nelle stime non c'č certezza. Uso, abuso e non uso dell'inferenza statistica nella ricerca sociale. *Rassegna Italiana di Sociologia, 63* (1)*,* 115–141.
Quenouille, M. (1949). Approximate tests of correlation in time series. *Journal of the Royal Statistical Society, Soc. Series B,* 11, 18–84.
Rindskopf, D. M. (1997). Testing "small," not null, hypotheses: classical and Bayesian approaches. V L. L. Harlow, S. A. Mulaik in J. H. Steiger (ur.), *What if there were no significance tests?* (str. 319–332). Mahwah, NJ: Lawrence Erlbaum.
Robinson, D. in Levin, J. R. (1997). Reflections on statistical and substantive significance, with a slice of replication. *Educational Researcher, 26* (5)*,* 21–26.
Rosenthal, R. (1991). Effect sizes: Pearson's correlation, its display via the BESD and alternative indices. *American Psychologist*, *46*, 1086–1087.
Rosenthal, R. (1993). Cumulating evidence. V G. Keren in C. Lewis (ur.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (str. 519–559). Hillsdale, NJ: Erlbaum.
Rosenthal, R. in Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2. izd.). New York: McGraw Hill.
Rosenthal, R., Rosnow, R. L. in Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research*. Cambridge University Press.
Rosenthal, R. in Rubin, D. B. (1982). A simple, general purpose display of magnitude of experimental effect. *Journal of Educational Psychology, 74,* 166–169.
Rosnow, R. L. in Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276–1284.
Rosnow, R. L. in Rosenthal, R. (1996). Computing contrasts effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods, 1*, 331–340.
Schafer, J. P. (1993). Interpreting statistical significance and nonsignificance. *Journal of Experimental Education*, *61* (4), 383–387.
Sedlmeier, P. in Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, *105* (2), 309–316.
Siegel, S. (1956). *Nonparametric Methods for Behavioral Sciences.* New York, NY: McGraw-Hill.
Simon, J. L. (1969). *Basic Research Methods in Social Science.* New York, NY: Random House.
Simon, J. L., Atkinson, D. T. in Shevokas, C. (1976). Probability and Statistics: Experimental Results of a Radically Different Teaching Method. *The American Mathematical*

Monthly, *83*, 733–739.
Simon, S. (1999, 7. januar). *Re: Type I and Type II error. Educational Statistics Discussion List (EDSTAT-L).* [Online]. Dostopno na E-mail: edstat-l@jse.stat.ncsu.edu [1999, January 7].
Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational and Psychological Measurement*, *61*, 605–632.
Steiger, J. H. in Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. V L. L. Harlow, S. A. Mulaik in J. H. Steiger (ur.), *What if there were no significance tests?* (str. 221–257). Mahwah, NJ: Erlbaum.
Sterling, R. (1959). Publications decisions and their possible effects on inferences drawn from tests of significance – or vice versa. *Journal of the American Statistical Association, 54*, 30–34.
Sterling, T. D., Rosenbaum, W. L. in Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician*, *49*, 108–112.
Sullivan, J. R. (2000). *A Review of Post-1994 Literature on Whether Statistical Significance Tests Should be Banned*. Prispevek, predstavljen na letnem srečanju Southwest Educational Research Association, Dallas, TX, 29.1.2000.
Thompson, B. (1993). The use of statistical significance tests in research: Bootstrap and other alternatives. *Journal of Experimental Education*, *61*, 361–377.
Thompson, B. (1996). AERA editorial policies regarding statistical significance testing: Three suggested reforms. *Educational Researcher, 25* (2)*,* 26–30.
Thompson, B. (1997). Editorial policies regarding statistical significance tests: Further comments. *Educational Researcher, 26,* 29–32.
Thompson, B. (2001). Significance effect sizes, stepwise methods, and other issues: strong arguments move the field. *Journal of Experimental Education, 70,* 80–93.
Thompson, B. (2002). What future quantitative social science research could look like: confidence intervals for effect sizes. *Educational Researcher, 31* (3)*,* 24–31.
Thompson, B. (2003). "Statistica", "pratica", "clinica": quanti tipi di significativita deve considerare chi opera nel counseling? *Bollettino di Psicologia applicata, 240,* 3–13.
Thompson, K. N in Schumacker, R. E. (1997). An evaluation of Rosenthal and Rubin's binomial effect size display. *Journal of Educational and Behavioral Statistics, 22*, 109–117.
Tukey, J. W. (1958). Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*, 29, 614.
Tukey, J. W. (1991). The philosophy of multiple comparison. *Statistical Science, 6,* 100–116.
von Mises, R. (1928/1957). *Probability, statistics, and truth.* London: Macmillan.
von Mises, R. (1964). *Mathematical theory of probability and statistics.* New York, NY: Academic Press.
Welkowitz, J., Ewen, R. B. in Cohen, J. (1982). *Introductory statistics for the behavioral sciences.* San Diego, CA: Harcourt Brace Jovanovich, Publishers.
Wilkinson, L. in The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wilson, F. D., Smoke, G. L. in Martin, J. D. (1973). The replication problem in sociology: A report and a suggestion. *Sociological Inquiry*, *43*, 141–149.

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.

Zar, J. H. (1996). *Biostatistical analysis* (3rd Ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Priloga

Neymanova in Pearsonova (1933) lema

Izhajamo iz naključne zvezne spremenljivke ξ, ki se porazdeljuje na osnovi verjetnostne gostote oz. verjetnostne porazdelitve $z(\xi; \theta)$, kjer je θ parameter in pripada prostoru Ω. Razdelimo množico Ω v podmnožici C in A, kjer je A komplement podmnožice C znotraj Ω (A = Ω - C). Podmnožica A se lahko ponovno razdeli na nove podmnožice C*'*. Če predpostavimo, da *z* velja v obeh primerih, parameter pa se lahko spreminja, je naša naloga določiti, ali vzorec $\mathbf{x} = [X_1, X_2, \ldots, X_n]$, za katerega velja funkcija verjetnostne gostote $z(\mathbf{x}; \theta)$, pripada populaciji.

V primeru, da velja H_0, potem ko smo določili funkcijo delitve množice Ω, $t(\mathbf{x})$, je pogojna verjetnost α, da θ pripada podmnožici C, enaka verjetnosti, da θ pripada katerikoli drugi podmnožici C*'* iz Ω:

$$p\left[t(\mathbf{x}) \in \mathrm{C} \middle| H_0\right] = p\left[t(\mathbf{x}) \in \mathrm{C}' \middle| H_0\right] = \alpha \tag{A1}$$

V primeru, ko pa je resnična alternativna hipoteza H_1, bo imel parameter večjo verjetnost, da bo padel v podmnožico C kot pa v katerokoli od podmnožic C*'*:

$$p\left[t(\mathbf{x}) \in \mathrm{C} \middle| H_1\right] > p\left[t(\mathbf{x}) \in \mathrm{C}' \middle| H_1\right] \tag{A2}$$

Ko smo definirali funkcijo verjetja *L* kot funkcijo skupne (združene) verjetnostne gostote *n*-tih naključno vzetih (torej neodvisnih) in enako porazdeljenih spremenljivk iz univerzuma, je naša naloga oceniti parameter, za katerega je verjetnost, da bodo komponente vektorja, ki predstavlja vzorec **x,** zavzele točno določene vrednosti (x_1, x_2, …, x_n), največja. Ta vrednost, označena kot $\hat{\theta}$, se imenuje »ocena največjega verjetja« in se razlikuje od katerekoli druge možne vrednosti parametra θ. Kot smo videli pri analizi, ta vrednost ustreza točki, v kateri je vrednost njenega prvega odvoda enaka nič, hkrati pa je njen drugi odvod negativen.

Pri določanju C predpostavimo, da lahko določimo tako konstanto k_a, da velja:

$$p\left[\frac{L(\mathbf{x};\theta_0)}{L(\mathbf{x};\theta_1)} < \mathrm{k}_\alpha \middle| H_0\right] = \alpha \tag{A3}$$

V enačbi A3 je θ_0 vrednost parametra za H_0 in θ_1 vrednost parametra za H_1. Kritično območje C prostora parametrov lahko določimo tako:

$$C = \left\{\mathbf{x} : \frac{L(\mathbf{x};\theta_1)}{L(\mathbf{x};\theta_0)} \geq k_\alpha\right\} \tag{A4}$$

Podobno lahko območje sprejetja določimo na naslednji način:

$$A = \left\{\mathbf{x} : \frac{L(\mathbf{x};\theta_0)}{L(\mathbf{x};\theta_1)} > k_\alpha\right\} \tag{A5}$$

To je znana Neymanova in Pearsonova (1933) lema.

Prispelo/Accepted: 01.08.2005
Sprejeto/Received: 30.01.2005