PRIMENJENA PSIHOLOGIJA, 2011/4, STR. 317-333

UDC 159.9.072
Naučna polemika

Lazar Tenjović[1]
Odeljenje za psihologiju,
Filozofski fakultet,
Univerzitet u Beogradu

Snežana Smederevac
Odsek za psihologiju,
Filozofski fakultet,
Univerzitet u Novom Sadu

MALA REFORMA U STATISTIČKOJ ANALIZI PODATAKA U PSIHOLOGIJI: MALO P NIJE DOVOLJNO, POTREBNA JE I VELIČINA EFEKTA

Rezime

Osnovni cilj ovog rada je ukazivanje na ograničenja i probleme koji se javljaju pri oslanjanju na konvencionalno testiranje statističke značajnosti u prikazivanju rezultata empirijskih istraživanja. U radu su prikazane sledeće pogrešne interpretacije *p* vrednosti: *a) p vrednost predstavlja verovatnoću da su dobijeni rezultati posledica greške uzorkovanja; b) p vrednost predstavlja verovatnoću pogrešne odluke u slučaju odbacivanja tačne nulte hipoteze; c) p vrednost predstavlja verovatnoću da je nulta hipoteza tačna ako se dobijeni rezultati uzmu u obzir; d) 1-p daje direktno meru verovatnoće da će se dobijeni rezultat ponoviti ako bi se istraživanje ponovilo u istim uslovima i e) 1-p predstavlja verovatnoću da je, kada se uzmu u obzir dobijeni podaci, alternativna hipoteza tačna.* Kao dopunski pokazatelji u procesu statističkog zaključivanja mogu poslužiti mere *veličine efekta* i *interval pouzdanosti*. Veliki broj pokazatelja *veličine efekta* mogu se svrstati u pokazatelje veličine razlika između aritmetičkih sredina, poput Koenovog *d*, Hidžisovog *g*, Glasove δ i Koenovog *f* i pokazatelje proporcije objašnjene varijanse, poput R^2, $\acute{\eta}^2$ i $\acute{\eta}p^2$. Date su okvirne sugestije za vrednovanje pojedinih pokazatelja, kao i za način njihove interpretacije u kontekstu specifičnih istraživačkih nacrta.

Ključne reči: statističko zaključivanje, p vrednost, veličina efekta, interval pouzdanosti

[1] Adresa za korespodenciju:
ltenjovi@f.bg.ac.rs

Primljeno: 28.12.2011.
Prihvaćeno za štampu: 21.01.2012.

Uvod

Ovaj rad napisan je s ciljem da: ukaže na ograničenja i probleme koji postoje u uobičajenom isključivom oslanjanju na konvencionalno testiranje statističke značajnosti pri statističkoj analizi i prikazivanju rezultata empirijskih istraživanja; predloži korišćenje statističkih pokazatelja veličine efekta, opravda i ohrabri upotrebu ovih pokazatelja u izveštajima o rezultatima istraživanja; prikaže pokazatelje koji se najčešće koriste i ukaže na doprinos koji pokazatelji veličine efekta mogu dati u zaključivanju na osnovu rezultata empirijskih istraživanja.

Još od vremena 'inferencijalne revolucije' u psihologiji (Balluerka, Gomez, & Hidalgo, 2005) koja se desila 40-ih i 50-ih godina dvadesetog veka testiranje statističke značajnosti (TSZ) predstavlja jedno od osnovnih 'oruđa' u empirijskim istraživanjima u psihologiji i glavno sredstvo statističkog zaključivanja koje istraživači koriste. Testiranje statističke značajnosti, kao uobičajena praksa u analizi podataka psiholoških istraživanja, svodi se, najjednostavnije rečeno, na testiranje statističke nulte hipoteze računanjem vrednosti statistika za testiranje i određivanjem odgovarajuće p vrednosti. Pri tome, ukoliko je p vrednost, tj. verovatnoća koju ona predstavlja, manja od neke odabrane vrednosti (recimo 0.05) nulta hipoteza se odbacuje. U suprotnom, tj. ukoliko je statistik za testiranje 'ispao' previše mali, a njemu pridružena p vrednost suviše velika, istraživaču ne preostaje ništa drugo do da (često sa osećanjem očaja zbog truda uloženog u izvođenje istraživanja) konstatuje da nultu hipotezu nije moguće odbaciti.

Kontroverza u vezi sa testiranjem statističke značajnosti postojala je još u debatama vođenim između samih tvoraca matematičke teorije statističkog zaključivanja: Ser Ronalda Fišera, s jedne, i Egona Pirsona i Jirži Nojmana, s druge strane (Fisher, 1955; Pearson, 1955; Neyman, 1956, cf. Lenhard, 2006). Uprkos tome, paradigma TSZ koja je svih proteklih decenija predstavljala osnovni statistički inferencijalni oslonac istraživačima u psihologiji zapravo je svojevrstan hibrid dva, u mnogim aspektima inkompatibilna, teorijska pristupa testiranju statističkih hipoteza: Fišerovskog i Nojman-Pirsonovskog (Hubbard & Bayarri, 2003; veoma dobri prikazi ova dva pristupa mogu se naći u Macdonald, 1997 i Christensen, 2005). Ova paradigma TSZ izazivala je sve vreme njenog postojanja u psihologiji znatna sporenja i osporavanja. U svakoj deceniji, i to u najuglednijim psihološkim časopisima pojavljivali su se polemički članci na ovu temu (npr. Balluerka et al., 2005; Gonzales, 1994; Greenwald, 1975; Nickerson, 2000; Rozeboom, 1960). Mnogi autori su izražavali neslaganje s konvencionalnim pristupom statističkom zaključivanju, zasnovanom na interpretaciji statističke značajnosti, tj. p vrednosti, smatrajući da bi takav pristup trebalo odbaciti (Carver, 1978; Cohen, 1994) ili da ga ne bi trebalo uzimati previše ozbiljno (Guttman, 1985). Ipak, bilo je i umerenijih glasova koji su pre svega isticali neophodnost i

korisnost TSZ u psihološkim istraživanjima (Frick, 1996; Mulaik, Raju & Harshman, 1997; Wainer, 1999).

Dve organizujuće teme ovih polemika mogu se najsažetije opisati na sledeći način:

1. unutrašnja logika, ograničenja i dometi TSZ i
2. načini na koje se izvodi i tumači, tj. upotrebljava (i, posebno, zloupotrebljava) TSZ.

Gotovo isključivo oslanjanje istraživača na TSZ pri analizi i tumačenju rezultata dobijenih istraživanjima ponekad se smatra jednim od ključnih sistematskih faktora sporog rasta kumulativnih znanja u psihologiji (Schmidt, 1996). Drugim rečima, istraživači često pribegavaju trivijalnim nacrtima, pravdajući ih dobijanjem statistički značajnih rezultata, zanemarujući realan doprinos tih rezultata u procesu razvoja naučnih disciplina koje kreiraju referentni okvir za istraživanja. Ipak, bez obzira na ograničenja koja kao i svako drugo sredstvo ima, TSZ predstavlja korisno analitičko sredstvo i može imati važnu funkciju u empirijskim istraživanjima u psihologiji ukoliko se dobro razumeju njegova ograničenja, ako se pravilno koristi i, što je najvažnije, ukoliko se dopuni dodatnim informacijama koje istraživačima stoje na raspolaganju u podacima.

Najčešće i najštetnije pogrešne interpetacije p vrednosti

Jedan od ključnih izvora problema koji se u korišćenju TSZ javljaju jeste nedovoljno razumevanje suštine i pravog značenja p vrednosti koja istraživačima u psihologiji često predstavlja prevashodni oslonac u zaključivanju na osnovu dobijenih rezultata. Iz ne sasvim jasnih racionalnih razloga (a ponekad i iz nedovoljnog poznavanja matematičke osnove TSZ) mnogi istraživači pripisuju p vrednosti značenja koju ova vrednost nema. Najčešće (i najštetnije po zaključivanje) pogrešne interpretacije p vrednosti su sledeće:

a) p vrednost predstavlja verovatnoću da su dobijeni rezultati posledica greške uzorkovanja. Ovakvo tumačenje je pogrešno jer greška uzorkovanja stoji u osnovi računanja verovatnoće p. Ova greška je već uključena u logiku statističkog testa koja omogućuje računanje vrednosti p. Prema tome, verovatnoća greške uzorkovanja, kada se izvodi statistički test je 1, tj. ta greška sigurno postoji. Dakle, statistički značajan rezultat ne znači da takav rezultat izvesno nije posledica greške uzorkovanja;

b) p vrednost predstavlja verovatnoću pogrešne odluke u slučaju odbacivanja tačne nulte hipoteze. Ovo tumačenje pogrešno poistovećuje p vrednost iz Fišerovskog pristupa statističkom zaključivanju sa vrednošću α iz Nojman-Pirsonovskog pristupa. Naime, α je *apriorna* verovatnoća greške pri

odbacivanju tačne nulte hipoteze koja se određuje pre nego što su podaci prikupljeni i na osnovu koje se postavlja kriterijum za odbacivanje nulte hipoteze, dok je p *posteriorna* verovatnoća da statistik za testiranje nulte hipoteze na slučajnom uzorku određene veličine uzme vrednost jednaku onoj koja je dobijena u konkretnom istraživanju ili vrednost statistika koja je veća od one dobijene u datom istraživanju. Dakle, konkretna p vrednost je zasnovana na podacima a p inače teorijski predstavlja slučajnu varijablu koja ima pod nultom hipotezom uniformnu raspodelu na intervalu [0,1] (Hubbard & Bayarri, 2003).

c) *p vrednost predstavlja verovatnoću da je nulta hipoteza tačna ako se dobijeni rezultati uzmu u obzir*. Ovakva pogrešna interpretacija verovatno je posledica snažne potrebe istraživača da na osnovu prikupljenih podataka proceni verovatnoću da je nulta hipoteza tačna. Međutim, p vrednost predstavlja uslovnu verovatnoću dobijanja rezultata kakav je dobijen u datom istraživanju ili još ekstremnijih rezultata ako je nulta hipoteza tačna. Da bismo na osnovu ishoda statističkog testa mogli dobiti verovatnoću nulte hipoteze pod dobijenim podacima trebalo bi da važi sledeća jednakost:

$$P(T \geq T_{\text{dobijeno}} \left| H_0 \text{ tačna}) = P(H_0 \text{ tačna} \right| T \geq T_{\text{dobijeno}})$$

Rečima iskazano, uslovna verovatnoća dobijanja statistika T za testiranje nulte hipoteze jednakog ili većeg od onoga koji je dobijen u datom istraživanju (ako je nulta hipoteza tačna) trebalo bi da bude jednaka uslovnoj verovatnoći tačnosti nulte hipoteze ako je statistik jednak ili veći od dobijenog. Takva jednakost, naravno, generalno ne važi. (Verovatnoću da je nulta hipoteza tačna, kada se uzmu u obzir dobijeni podaci moguće je izračunati u okviru tzv. Bejzijanskog pristupa statističkom zaključivanju ali je za to neophodno imati ocenu verovatnoće tačnosti nulte hipoteze pre nego što se prikupe podaci u datom istraživanju). Statistička istraživanja u okviru tzv. Bejzijanskog pristupa statističkom zaključivanju pokazuju da p vrednost teži da preceni dokaze protiv nulte hipoteze: simulacionim eksperimentima pokazano je da je, za p = 0.05, posteriorna verovatnoća da je nulta hipoteza tačna za razumne vrednosti apriornih verovatnoća njene tačnosti znatno veća od 0.05 (cf. Hubbard & Bayarri, 2003).

a) 1-p daje direktno meru verovatnoće da će se dobijeni rezultat ponoviti ako bi se istraživanje ponovilo u istim uslovima. Istraživači često veruju da mala p vrednost koja vodi odbacivanju nulte hipoteze daje direktnu meru verovatnoće repliciranja rezultata u ponovljenim istraživanjima i da je ova mera jednaka 1-p. Premda je pokazano (cf. Greenwald, Gonzalez, Harris, & Guthrie, 1996) da između p vrednosti i statističke snage repliciranja rezultata postoji veza, verovatnoću repliciranja rezultata nije moguće izračunati na pomenuti način (Kline, 2004).

b) 1-p predstavlja verovatnoću da je, kada se uzmu u obzir dobijeni podaci, alternativna hipoteza tačna. Vrednost 1-p predstavlja verovatnoću da se dobije manje ekstremna vrednost statistika za testiranje nulte hipoteze, ako je nulta hipoteza tačna i ne može, prema tome, pokazivati posteriornu verovatnoću tačnosti alternativne hipoteze (Kline, 2004).

Dakle, u korišćenju TSZ neophodno je voditi računa o njegovim ograničenjima i dometima koji proističu iz pravog značenja p vrednosti: p vrednost predstavlja verovatnoću da se, *ako je nulta hipoteza tačna*, na uzorku iste veličine kakav je korišćen u datom istraživanju dobije takva ili ekstremnija vrednost statistika za testiranje nulte hipoteze. I ništa više od toga! Veoma mala p vrednost (na primer $p \leq 0.05$) može poslužiti istraživačima kao empirijski argument za opravdanost njihove sumnje u tačnost nulte hipoteze, tj. za njeno odbacivanje.

Osnovni problem koji se uočava u procesu statističkog zaključivanja na osnovu visine p vrednosti leži u mogućnosti da neki rezultat, iako statistički značajan, ima trivijalne implikacije ili praktičnu relevantnost (Kirk, 2001).

Posebno je važno u tom smislu uočiti ograničenost informacija sadržanih u p vrednosti: uprkos pretpostavci (koja je ponekad prisutna među istraživačima) da statistička značajnost i relevantnost rezultata konvergiraju (Thisted, 1998), p vrednost ne govori neposredno o praktičnoj relevantnosti, tj. važnosti dobijenih rezultata. Ilustrujmo to na jednom vrlo jednostavnom (izmišljenom) primeru.

U tabeli 1 nalaze se 'podaci' o razlikama između muškaraca i žena u pogledu broja seksualnih partnera koje bi želeli da imaju u narednih godinu dana (radi jednostavnosti željeni broj seksualnih partnera sveli smo na samo dve kategorije: jedan i više od jedan). U levom delu tabele prikazane su učestalosti dveju kategorija odgovora u pogledu broja željenih seksualnih partnera za 'uzorak' od 500 muškaraca i 500 žena.

Tabela 1.
Hipotetički primer za ilustraciju osetljivosti p vrednosti na veličinu uzorka

	Broj željenih partnera		Broj željenih partnera	
	Jedan	Više od 1	Jedan	Više od 1
Muškarci	250 (50.0%)	250 (50.0%)	25 000 (50.0%)	25 000 (50.0%)
Žene	260 (52.0%)	240 (48.0%)	26000 (52.0%)	24 000 (48.0%)

U desnom delu tabele prikazane su iste te učestalosti pomnožene sa 100. Dakle, u drugom slučaju 'uzorak' bi obuhvatao 50 000 muškaraca i 50 000 žena. Hi-kvadrat

test značajnosti na osnovu podataka prikazanih u levom delu tabele daje Pirsonov Hi-kvadrat statistik jednak 0.4 i p vrednost jednaku 0.527. Međutim, isti ovaj test značajnosti na osnovu podataka prikazanih u desnom delu tabele, *bez obzira na to što su razlike u procentima između muškaraca i žena u oba dela tabele iste*, daje Pirsonov Hi-kvadrat statistik 40.02 i p vrednost daleko manju od 0.0001! Praktična (pa i teorijska, na primer u smislu provere određenih postavki evolucione psihologije) relevantnost rezultata i veličina efekta su isti u oba slučaja. Međutim, p vrednosti i zaključci o statističkoj značajnosti su u ova dva slučaja bitno različiti. Ono što ovaj izmišljeni primer takođe dobro ilustruje jeste osetljivost p vrednosti (i posledično statističke značajnosti) na veličinu uzorka. Naime, statistici koji se koriste u TSZ i na osnovu čije distribucije uzorkovanja se određuje p vrednost mogu se generalno predstaviti kao funkcija dveju veličina–veličine uzorka (ili neke funkcije veličine uzorka) i veličine efekta (tj. statističkog pokazatelja stepena u kojem se rezultati dobijeni na uzorku razlikuju od onoga što je specifikovano nultom hipotezom). Dakle, statistik za testiranje nulte hipoteze, u oznaci T, može se predstaviti na sledeći način:

$$T = f(n)*IVE$$

pri čemu je f(n) funkcija veličine uzorka, a IVE indeks veličine efekta. Stoga je, na primer, i za trivijalnu veličinu efekta, menjanjem veličine uzorka moguće dobiti veoma malu p vrednost. S druge strane, istraživanje koje je urađeno na malom uzorku, a u kojem je dobijen visok indeks veličine efekta, može rezultirati visokom p vrednošću. Na taj način, poređenje rezultata različitih istraživanja u kojima je isti problem ispitivan na uzorcima različite veličine postaje prilično komplikovan poduhvat.

Problemi sa TSZ koje smo naveli (i mnogi drugi koji su opisani u velikom broju članaka u psihološkim časopisima) nameću potrebu da se dobro razumeju ograničenja TSZ kako bi se ovi postupci pravilno koristili. Isključivo oslanjanje na vrednost statistika za testiranje značajnosti i njemu pridruženu p vrednost u statističkom zaključivanju je nedovoljno. Rezultati koji se dobijaju TSZ pružaju, dakle, ograničene informacije koje se često pogrešno razumeju i, stoga, zloupotrebljavaju. Zato je važno pri statističkom zaključivanju na osnovu rezultata dobijenih istraživanjem prikazati i uzeti u razmatranje, osim deskriptivnih statističkih mera, i druge bitne statističke informacije koje se mogu dobiti iz podataka. To se, pre svega odnosi, na *intervale pouzdanosti* (ili, kako se još često zovu, intervale poverenja) i pokazatelje *veličine efekta*. *Intervali pouzdanosti* daju veoma važnu informaciju koja se ne može dobiti na osnovu p vrednosti – o preciznosti ocena odgovarajućih parametara na osnovu statistika dobijenih na uzorku i, posledično, o poverenju koje možemo imati u ove ocene. S druge strane, važna prednost pokazatelja *veličine efekta* nad standardnim pokazateljima statističke značajnosti je lakše sprovođenje meta-analiza na osnovu dobijenih rezultata. Podaci o *veličini*

efekta omogućuju stvaranje osnove za kreiranje specifičnijih hipoteza od strane budućih istraživača. Na primer, omogućuju evaluaciju podudaranja rezultata različitih istraživanja u određenoj oblasti koja su posvećena rešavanju istog problema. Pored toga, sama priroda statističkih informacija sadržanih u pokazateljima *veličine efekta* pomaže da se osvetle upravo oni aspekti podataka koji su važni za substantivne teorijske rasprave u psihologiji. Ishodi TSZ korisni su pre svega za to da se pokaže da je malo verovatno da su dobijeni rezultati posledica slučaja, dok se pokazatelji *veličine efekta* mogu veoma korisno upotrebiti u odgovorima na supstantivna pitanja, tj. za poređenje dobijenih rezultata sa predviđanjima određene psihološke teorije za datu oblast istraživanja.

Radna grupa za statistiku (Task Force on Statistical Inference - TFSI; Wilkinson & TFSI, 1999) u okviru Američke psihološke asocijacije (APA) je 1998. godine uvela nove kriterijume za prikazivanje rezultata statističkih analiza. Ova odluka bila je rukovođena potrebom da se uvaži sve veći broj dokaza koji su ukazivali na to da klasični pristup TSZ ima ozbiljne nedostatke.

Čime se može dopuniti p vrednost?

Kao što smo već istakli u prethodnom delu ovog rada, problemi koje generiše isključivo oslanjanje na p vrednost kao osnov za statističko zaključivanje mogu biti barem delimično umanjeni uvođenjem dodatnih statističkih pokazatelja koji je moguće odrediti na osnovu raspoloživih podataka, kao što su *veličina efekta* (effect size) i *interval pouzdanosti* (confidence interval).

Veličina efekta se u najširem smislu može definisati kao bilo koji statistički pokazatelj koji kvantifikuje stepen u kojem se rezultat dobijen na uzorku razlikuje od očekivanja specifikovanih nultom hipotezom (Sun, Pan, & Wang, 2010). Postoji veliki broj pokazatelja *veličine efekta*, a u najopštijem smislu mogu se podeliti u dve velike grupe. Prvu grupu čine pokazatelji veličine razlika između aritmetičkih sredina, poput Koenovog *d*, Hedžisovog *g*, Glasove δ i Koenovog *f* (za više od dve grupe). Drugu grupu čine pokazatelji proporcije varijanse obuhvaćene korelacijom između varijabli i predstavljeni su pokazateljima R^2 i η^2.

Jedan od najčešće korišćenih pokazatelja *veličine efekta* u poređenju dveju grupa je Koenovo *d*;

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

gde je $\sigma = \sqrt{[\Sigma(X - M)^2 / N}$; X - sirovi skor; M - aritmetička sredina a N je broj rezultata.

Koen (Cohen, 1988) je definisao *d* kao razliku između aritmetičkih sredina podeljenu standardnom devijacijom bilo koje grupe, ako je varijansa grupa homogena. U praksi se, međutim, najčešće koristi kombinovana standardna devijacija, koja predstavlja koren proseka varijansi (Cohen, 1988, p. 44).

Kao mera *veličine efekta* za poređenje dveju grupa koristi se i Hedžisovo *g*. Obično se izračunava tako što se koristi kvadratni koren prosečnog kvadrata greške (Mean Square Error) u proceduri ANOVA kada se ona primenjuje za ispitivanje razlika između dve grupe.

$$g = \frac{M_1 - M_2}{S_{pooled}}$$

gde je $S_{pooled} = \sqrt{MS_{within}}$, tj. kvadratni koren varijanse unutar grupa.

Glasova δ je definisana kao razlika aritmetičkih sredina između eksperimentalne i kontrolne grupe podeljena sa standardnom devijacijom kontrolne grupe:

$$g = \frac{M_1 - M_2}{\sigma_{control}}$$

Mera veličine efekta u analizi varijanse predstavlja stepen povezanosti između sistematskog izvora variranja (glavni efekat, interakcija, linearni kontrast) i zavisne varijable. Najčešći pokazatelj *veličine efekta* u dizajnu u kojem se rezultati obrađuju analizom varijanse je kvadrirana eta, u oznaci η^2:

$$\eta^2 = \frac{\sigma^2_{source}}{\sigma^2_{total}}$$

pri čemu je σ^2_{source} varijabilitet obuhvaćen određenim izvorom variranja, a σ^2_{total} ukupni varijabilitet na zavisnoj varijabli. Ovaj pokazatelj se uobičajeno definiše kao proporcija ukupne varijanse zavisne varijable koja je objašnjena svakim posebnim izvorom varijacija u istraživačkom dizajnu (Richardson, 2011). Parcijalizovana η^2 (η_p^2) predstavlja pokazatelj proporcije ukupne varijanse zavisne varijable koja se može pripisati određenom izvoru variranja ukoliko se kontroliše efekat ostalih sistematskih izvora variranja u dizajnu:

$$\eta^2_p = \frac{\sigma^2_{source}}{\sigma^2_{source} + \sigma^2_{error}}$$

Parcijalizovana η^2 se najčešće koristi kao pokazatelj *veličine efekta* u situacijama u kojima se koristi dodatna manipulacija ili kontrola varijabli koje su uključene

u istraživački dizajn i uobičajeno pokazuje nešto više vrednosti od $\acute{\eta}^2$ (Richardson, 2011). U slučaju jednostavnih istraživačkih nacrta, ove dve vrednosti se ne razlikuju. Međutim, što je složeniji nacrt (više faktora u analizi), veće su razlike između $\acute{\eta}^2$ i $\acute{\eta}_p^2$. Takođe, $\acute{\eta}_p^2$ je neaditivan pokazatelj, čija kumulativna vrednost može biti veća od 1, za razliku od $\acute{\eta}^2$, koji predstavlja aditivnu meru i čija kumulativna vrednost ne može biti veća od 1. Drugim rečima, potreban je oprez prilikom interpretacije $\acute{\eta}_p^2$, posebno u slučaju složenih istraživačkih nacrta.

U SPSS-u, $\acute{\eta}_p^2$ za jednofaktorijalni nacrt sa nezavisnim grupama se može izračunati u proceduri *Means*, ali ne i u T-Test-u ili One-Way ANOVA. Međutim, $\acute{\eta}_p^2$ je dostupna *u General Linear Model* proceduri (*Options*, *Estimation Effect Size*).

Pored $\acute{\eta}_p^2$, pokazatelji *veličine efekta* za dizajn sa neponovljenim merenjima u kojem se koristi analiza varijanse su i Koenovo f, kvadrirani epsilon (ε^2), kvadrirana omega (ω^2) i parcijalizovana ω^2.

U većini korelacionih nacrta kao pokazatelj *veličine efekta* koristi se R^2. R^2 i $\acute{\eta}^2$ su kompatibilne mere. R^2 se odnosi na proporciju predviđene varijanse kriterijumske varijable na osnovu prediktorskih varijabli, a u SPSS-u i drugim statističkim paketima automatski je uključen u ispis za *linearnu regresiju* i nije ga potrebno dodatno specifikovati.

Za dizajn s ponovljenim merenjima preporučuje se intraklasni koeficijent korelacije, u oznaci ρ_I:

$$\rho_I = \frac{\sigma^2_{source}}{\sigma^2_{source} + \sigma^2_{error}}$$

pri čemu je σ^2_{source} varijansa koja otpada na određeni izvor variranja, a σ^2_{error} varijansa greške.

Ovaj pokazatelj se odnosi na obuhvat varijanse koji je definisan specifičnim eksperimentalnim faktorom i konceptualno je sličan ω^2.

Najčešće korišćeni pokazatelji *veličine efekta* za tabele kontingencije 2x2 su fi-koeficijent i količnik šansi (odds ratio), a za tabele kontingencije većih dimenzija Kramerov V koeficijent. Ove pokazatelje moguće je u programu SPSS/PASW dobiti u proceduri Crosstabs.

U multivarijacionoj analizi varijanse i deskriptivnoj diskriminacionoj analizi kao pokazatelji *veličine efekta* uobičajeno se koriste kvadrirana Mahalanobisova distanca (kada postoje samo dve grupe) i multivarijaciona $\acute{\eta}^2$.

Kvadrirana Mahalanobisova distanca, u oznaci D^2, definisana je na sledeći način:

$$D^2 = (\mathbf{M}_1 - \mathbf{M}_2)^t \mathbf{S}^{-1} (\mathbf{M}_1 - \mathbf{M}_2)$$

pri čemu su $\mathbf{M}_1$ i $\mathbf{M}_2$ vektori aritmetičkih sredina grupa, a **S** kombinovana matrica kovarijansi obeju grupa. Kvadrirana Mahalanobisova distanca očigledno predstavlja standardizovanu razliku između vektora aritmetičkih sredina dveju grupa, tj. kvadrirano uopštenje Koenovog *d* na multivarijacioni slučaj.

Multivarijaciona $\acute{\eta}^2$ je definisana na sledeći način:

$$\text{Multiv. } \acute{\eta}^2 = 1\text{-}\Lambda$$

pri čemu je Λ Wilksova lambda.

Kao pokazatelj *veličine efekta* u klasifikacionoj diskriminacionoj analizi i logističkoj regresiji može se koristiti indeks I koji su definisali Huberty i Lowman (2000):

$$I = \frac{H_o - H_e}{1 - H_e}$$

pri čemu su Ho empirijski dobijena, a He očekivana (na osnovu slučaja) tačnost klasifikovanja ispitanika u grupe. Ovaj indeks pokazuje koliko je bolja uspešnost klasifikovanja ispitanika u grupe na osnovu klasifikacione diskriminacione funkcije ili logističke regresije u odnosu na uspešnost koja bi se dobila slučajnim klasifikovanjem.

U Modelima strukturalnih jednačina *veličina efekta* je indeks podesnosti (GFI - goodness of fit), ali se u poslednje vreme preferira kvadratni koren iz prosečne kvadrirane greške aproksimacije, ili skraćeno RMSEA (*Root Mean Square Error of Approximation*). Ovaj pokazatelj ukazuje na diskrepancu između opažene i pretpostavljene matrice prema broju stepeni slobode pretpostavljenog modela. U pitanju je parsimonijski pokazatelj, osetljiv na kompleksnost modela, kao i korigovano R^2 u višestrukoj regresionoj analizi.

Kako bi trebalo da izgleda prikaz rezultata istraživanja

Veoma je važno da psiholozi počnu da pri statističkim analizama i tumačenju rezultata pomere naglasak sa dihotomnog načina razmišljanja, karakterističnog za testiranje značajnosti nulte hipoteze, na procenu *veličine efekta*, što znači da u prikazu rezultata *veličina efekta* mora uvek biti navedena u vidu nekog prikladnog pokazatelja, a kad god je to moguće, neophodno je navesti i *interval pouzdanosti* i, naravno, interpretirati ga. Da bi se taj cilj ostvario, neophodno je da istraživači imaju na raspolaganju adekvatna uputstva o načinu na koji se ovi pokazatelji prikazuju i interpretiraju (Cumming, Fidler, Leonard, Kalinowski, Christiansen,

Kleining, Lo, McMenamin, & Wilson, 2007). Osnovne sugestije za tumačenje pojedinih pokazatelja *veličine efekta* prikazane su u Tabeli 2. (Za pokazatelj I koji nije sadržan u toj tabeli Huberty i Lowman (2000) su predložili da $I \leq 0.10$ govori o 'malom' a $I \geq 0.35$ o velikom efektu. Za Mahalanobisovo D^2, koje takođe nije prikazano u Tabeli 2, Stevens (1980) je predložio da se vrednosti oko 0.20 tretiraju kao da govore o malom efektu, oko 0.5 o srednjem efektu, a vrednosti ≥1 kao da govore o velikom efektu.

Tabela 2.
Sugestije za interpretaciju veličine efekta

Tip procene veličine efekta	Indikatori veličine efekta	PMVE*	Osrednji efekat	Jak efekat
Standardizovane razlike između grupa	d, g, δ, f	.2	.5	.8
Kvadrirani indikatori jačine povezanosti	$r^2, R^2, \acute{\eta}^2$	.01	.05	.13

*PMVE = preporučeni minimum veličine efekta

Na koji način se interpretiraju ove vrednosti u kontekstu prikaza rezultata istraživanja? Pretpostavimo da je u jednofaktorijalnom eksperimentu kao zavisna varijabla korišćena mera fiziološke reakcije (na primer pulsa), a da su osnovu eksperimentalne manipulacije činile tri nezavisne situacije u kojima su ispitanici posmatrali fotografije dopadljivih, manje dopadljivih i nedopadljivih pripadnika suprotnog pola. Dobijene razlike između ove tri grupe ispitanika u izraženosti fizioloških reakcija su statistički značajne, a $\acute{\eta}_p^2$ je 0.08. *To znači da, u odnosu na veličinu efekta, 8% varijabiliteta u merama zavisne varijable može biti objašnjeno, ili predviđeno pripadnošću grupi, odnosno izloženošću različitim vrstama stimulusa.* Naravno da postoje istraživački nacrti u kojima, iako osrednji, ovakav efekat može biti trivijalan, ali u nekim istraživačkim nacrtima ovakav efekat može biti vredan pomena.

Iako se *veličina efekta* smatra pokazateljem koji je veoma koristan u statističkom zaključivanju, važno je naglasiti da na njegovu vrednost utiču neki aspekti istraživačkog dizajna, kao što su karakteristike uzorka i merenja. Ako je uzorak suviše mali, ili nije slučajan, može doprineti iskrivljenoj vrednosti veličine efekta, koja bi trebalo biti interpretirana s izvesnim oprezom. Određeni metodološki aspekti istraživanja takođe mogu uticati na vrednost pokazatelja veličine efekta. Različi-

ti aspekti procene mogu imati nisku pouzdanost, koja doprinosi nižoj vrednosti pokazatelja veličine efekta. Nasuprot tome, loše standardizovane mere mogu dovesti do njegove precenjene procene. *Veličina efekta* može varirati i u zavisnosti od primenjenog statističkog postupka. Na primer, ako se istraživač opredeli za postupak dihotomizacije kontinuiranih varijabli, procena *veličine efekta* će biti niža. Takođe, ako se relevantne varijable u istraživanju ne kontrolišu na adekvatan način, *veličina efekta* može biti veštački povišena.

U izboru pokazatelja *veličine efekta* potrebno je voditi računa o karakteristikama distribucije i drugim svojstvima podataka jer, kao i svi drugi statistici, pokazatelji *veličine efekta* podrazumevaju određene uslove za adekvatnu primenu. Budući da pokazatelji *veličine efekta* predstavljaju ocene odgovarajućih parametara (*veličine efekta* u populaciji) veoma je korisno pri njihovom prikazivanju dati ne samo njihovu vrednost već i *intervale pouzdanosti* kako bi se mogla proceniti preciznost ovih ocena. Nažalost, u programu SPSS uglavnom nije moguće automatski dobiti ove intervale. Postupak konstruisanja egzaktnih *intervala pouzdanosti* za većinu pokazatelja *veličine efekta* koje smo prikazali u ovom radu nije sasvim jednostavan i podrazumeva korišćenje parametara necentralnosti koji se određuju na osnovu necentralnih distribucija a koje su kao funkcije sadržane u programu SPSS u okviru komande Compute (detalji ovog postupka mogu se naći u Odgaard & Fowler, 2010). *Intervale pouzdanosti* za jedan broj pokazatelja veličine efekta moguće je dobiti korišćenjem modula Power analysis statističkog paketa STATISTICA 10. Takođe, korišćenjem MBESS programskog paketa napisanog u R okruženju, a koji je od novembra 2011. godine dostupan na URL adresi http://nd.edu/~kkelley/site/MBESS.html moguće je konstruisati *intervale pouzdanosti* za veliki broj pokazatelja veličine efekta (Kelley, 2007).

U tumačenju pokazatelja *veličine efekta* veoma je važno imati u vidu prirodu pojave koja se ispituje jer u određenim slučajevima i veoma male *veličine efekata* mogu biti i teorijski i praktično relevantne (npr. u istraživanjima faktora preživljavanja ili obolevanja od određenih bolesti). Isto tako, pokazatelje *veličine efekta* uvek treba tumačiti u kontekstu srodnih ranijih istraživanja, direktnim poređenjem dobijenih pokazatelja sa onima koji su dobijeni u prethodnim srodnim istraživanjima (Vacha-Haase & Thompson, 2004). Okvirne sugestije za klasifikovanje *veličine efekta* koje su date u Tabeli 2 treba koristiti pre svega u situacijama kada ne postoje srodna prethodna istraživanja.

Statistička reforma u psihologiji

APA Priručnik iz 1998. godine sadrži preporuku da bi statističke analize trebalo, pored izračunavanja *p vrednosti*, da obuhvate i *veličinu efekta* i/ili *interval pouz-*

danosti, a u najnovijem, šestom izdanju Priručnika (APA, 2010), naglašeno je da su ovi elementi minimum koji se u vezi s prikazom rezultata istraživanja očekuje u svim APA časopisima.

Rezultati istraživanja koja su imala za cilj ispitivanje trenda usvajanja novih APA preporuka u vodećim psihološkim časopisima pokazuju da je testiranje značajnosti nulte hipoteze nesumnjivo duboko ukorenjeno u način razmišljanja psihologa (Cumming et al., 2007).

Naime, u istraživanju u kojem su analizirana četiri APA časopisa u 1995. godini, rezultati pokazuju da se procenat članaka u kojima su navedeni pokazatelji veličine efekta kreću od 12% u *Journal of Experimental Psychology do 77% u Journal of Applied Psychology* (Kirk, 1996). Međutim, najčešći postupak za analizu podataka u *Journal of Applied Psychology* je regresiona analiza, a svi statistički paketi prikazuju jedan od pokazatelja *veličine efekta*, R^2, kao standardni deo prikaza rezultata. To nije slučaj sa analizom varijanse, koja se najčešće primenjuje kao postupak izbora u eksperimentalnim istraživanjima. Statistički paketi nemaju pokazatelje *veličine efekta* koji bi bili sastavni deo standardnog prikaza rezultata. Stoga ovaj rezultat predstavlja artefakt uobičajene prakse u različitim psihološkim disciplinama.

Situacija je za nijansu drugačija kada se ispituje učestalost prikazivanja *intervala pouzdanosti* kao sastavnog dela rezultata. Naime, od ispitanih autora članaka publikovanih u periodu od 2003. do 2004. godine u časopisima *Acta Psychologica*, *Child Development*, *Cognition*, *Journal of Abnormal Child Psychology*, *Journal of Abnormal Psychology*, *Journal of Consulting and Clinical Psychology*, *Journal of Experimental Psychology: General, Journal of Personality and Social Psychology*, *Psychological Science*, *Quarterly Journal of Experimental Psychology, 55*% je smatralo da bi *interval pouzdanosti* trebalo češće koristiti, 75% je smatralo da *je interval pouzdanosti* informativniji od p vrednosti, a 55% je smatralo da je klasično testiranje statističke značajnosti sasvim zadovoljavajuće. Pokazatelj *p vrednosti* se koristi skoro u svim člancima. Samo 24,1% članaka u kojima je naveden *interval pouzdanosti* sadrži i njegovu interpretaciju, ali najčešće u terminima karakterističnim za klasično testiranje značajnosti. Čak i u medicinskim časopisima, u kojima je tradicija prikazivanja *intervala pouzdanosti* prilično duga, ovaj pokazatelj retko biva interpretiran (Fidler, Thomason, Cumming, Finch, & Leeman, 2004).

Domaći časopisi iz oblasti psihologije nemaju u uputstvima autorima zahtev da prikaz rezultata mora uključivati neki od pokazatelja *veličine efekta* i *intervale pouzdanosti* i njihovu adekvatnu interpretaciju. Takvi zahtevi su jedan od osnovnih preduslova dobre istraživačke prakse i jedini način koji bi autore prinudio da svojim rezultatima posvete nešto više od rutinske pažnje, koja se često manifestuje isključivim oslanjanjem na p vrednost u statističkom zaključivanju na osnovu dobijenih rezultata. Nadamo se da će ovaj rad biti podsticaj za postavljanje takvih zahteva.

Literatura

American Psychological Association. (2010). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Balluerka, N., Gomez, J., & Hidalgo, D. (2005). The Controversy over Null Hypothesis Significance Testing Revisited. *Methodology*, *1*, 55–70.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review*, *48*, 378–399.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillside, NJ: Erlbaum.

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

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician*, *59*, 121–126.

Cumming, G., Fidler, F., Leonard, M., Kalinowski, P., Christiansen, A., Kleining, A., Lo, J., McMenamin, N., & Wilson, S. (2007). Statistical reform in psychology: Is anything changing? *Psychological Science, 18*(3), 230–232.

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

Fisher, R. (1955). Statistical methods and Scientific Induction. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*, 69–78.

Frick, R. W. (1996). The Appropriate Use of Null Hypothesis Testing. *Psychological Methods*, *1*, 329–390.

Gonzalez, R. (1994). Tha Statistics Ritual in Psychological Research. *Psychological Science*, *5*, 321–328.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Greenwald, A. G., Gonzalez, R., Harris, R. J., & Guthrie, D. (1996). Effect Sizes and p-values: What should be reported and what should be replicated? *Psychophysiology*, *33*, 175–183.

Guttman, L. (1985). The illogic of statistical inference for cumulative science. *Applied Stochastic Models and Data Analysis*, *1*, 3–10.

Hubbard, R., & Bayarri, M. J. (2003). Confusion Over Measure of Evidence (p's)

Versus Errors (α's) in Classical Statistical Testing. *The American Statistician*, *57*, 171–182.

Huberty, C. J., & Lowman, L. L. (2000). Group overlap as a basis for effect size. *Educational and Psychological Measurement,* 60, 543–563.

Kelley, K. (2007). Confidence Intervals for Standardized Effect Sizes: Theory, Application, and Implementation. *Journal of Statistical Software*, *20*, Issue 8. Dostupno na http: www.jstatsft.org/.

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

Kirk, R. E. (2001). Promoting Good Statistical Practices: Some Suggestions. *Educational and Psychological Measurement,* 61, 213–218.

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

Lenhard, J. (2006). Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson. *British Journal of Philosophy of Science*, 57, 69–91.

Levine, T., & Hullett, C. (2002). Eta squared, partial eta squared and misreporting of effect size in communication research. *Human Communication Research, 28*, 612–625.

Macdonald, R. R. (1997). On statistical testing in psychology. *British Journal of Psychology*, *88*, 333–347.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a Time and Place for Significance Testing. In L.L.Harlow, Mulaik, S.A. & Steiger, J.H. (Eds). *What if There Were No Significance Tests?* (pp. 68–116). Mahwah, NJ: Erlbaum.

Neyman, J. (1956). Note on an Article by Sir Ronald Fisher. *Journal of the Royal Statistical Society. Series B (Methodological)*, *18*, 288–294.

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

Odgaard, E. C., & Fowler, R. L. (2010). Confidence Intervals for Effect Sizes: Compliance and Clinical Significance in the *Journal of Consulting and Clinical Psychology*. *Journal of Consulting and Clinical Psychology*, *78*, 287–297.

Pearson, E. S. (1955). Statistical Concepts in the Relation to Reality. *Journal of the Royal Statistical Society. Series B (Methodological)*, *17*, 204–207.

Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review*, *6*, 135–147.

Rozeboom, W. W. (1960). The fallacy of the null-hypotheses significance test. *Psychological Bulletin*, *57*, 416–428.

Schmidt, F. L. (1996). Statistical Significance Testing and Cumulative Knowledge in Psychology: Implications for Training of Researchers. *Psychological Methods*, *1*, 115–129.

Stevens, J. P. (1980). Power of the multivariate analysis of variance tests. *Psychological Bulletin*, *88*, 728–737.

Sun, S. S., Pan, W., & Wang, L. L. (2010). A Comprehensive Review of Effect Size Reporting and Interpreting Practices in Academic Journals in Education and Psychology. *Journal of Educational Psychology, 102*, 989–1004.

Thisted, R. A. (1998). What is a p-value? Dostupno na http://galton.uchicago.edu/~thisted/ Distribute/ pvalue.pdf (pristupljeno 7.12.2011.).

Vacha-Haase, T., & Thompson, B. (2004). How to estimate and Interpret Various Effect Sizes. *Journal of Counseling Psychology*, *51*, 473–481.

Wainer, H. (1999). One Cheer for Null Hypotheses Significance Testing. *Psychological Methods*, *4*, 212–213.

Wilkinson, L., & the Task Force on Statistical Inference (1999). Statistical methods in psychology journals. *American Psychologist, 54*, 594–604.

Lazar Tenjović
Department of Pscychology,
Faculty of Philosophy,
University of Belgrade

Snežana Smederevac
Department of Pscychology,
Faculty of Philosophy,
University of Novi Sad

A SMALL REFORM IN THE DATA ANALYSIS IN PSYCHOLOGY: A SMALL P IS NOT ENOUGH, EFFECT SIZE IS NEEDED TOO

Abstract

The main objective of this paper is to point out the limitations and problems that occur when relying on conventional tests of statistical significance in presenting the results of empirical research. The following misinterpretations of p values are emphasized in the paper: *a) p value is the probability that the result was due to sampling error; b) p value represents the probability of wrong decisions in the event of rejecting the true null hypothesis; c) p value is the probability of the null hypothesis being true given the data d) 1-p is the probability that a replication attempt will also reject the null hypothesis, and e) 1-p is the probability that the alternative hypothesis is true given the dat*a. *Effect sizes and confidence intervals* can be used as additional indices in the process of statistical inference. A large number of the effect size indexes can be classified as standardized mean difference indices, such as Cohen's d, Hedges' g, Glass's δ and Cohen's *ƒ*2, and variance-accounted-for indices, such as R^2, and $\acute{\eta}^2$ and $\acute{\eta}p^2$. Suggestions for approximate evaluations of certain indicators, as well as the manner of their interpretation in the context of specific research designs are given.

Key words: statistical inference, p value, effect size, confidence interval