Československá psychologie 2017 / ročník LXI / číslo 5

Diskuse

O VÝZNAME P-HODNÔT: REFLEXIA NA SILNEJÚCU KRITIKU TESTOV VÝZNAMNOSTI

IVAN ROPOVIK
Prešovská univerzita, Prešov

ABSTRACT

On the meaning of p-values: criticism of significance tests revisited

I. Ropovik

In the light of the reproducibility crisis, there is a growing criticism of null hypothesis significance testing (NHST) and its output, the p-values. In fact, the intuitive interpretations of p-values are usually misleading because the p-values do not answer the questions that researchers tend to ask and their role in the context of statistical inference is typically grossly overestimated. The aim of this article is to discuss the most pervasive misinterpretations of p-values as well as the statistical reasons for the weak reproducibility of findings based on crossing the $p < .05$ threshold. On the other hand, the article provides a reconsideration of the arguments against the p-values, while offering a pragmatic, but still rigorous perspective on their logic and usefulness.

key words:
p-values,
null hypothesis significance testing,
statistical inference,
hypothesis testing,
reproducibility crisis

kľúčové slová:
p-hodnoty,
testovanie významnosti nulovej hypotézy,
štatistická inferencia,
testovanie hypotéz,
kríza reprodukovateľnosti

V psychológii rovnako ako v mnohých iných vedných odboroch obvyklým cieľom výskumu býva odhaliť princípy a tendencie, ktoré sú platné v istej populácii. Formulovanie akýchkoľvek interpretácií a záverov vzťahujúcich sa k populácii na základe dát z výberového súboru je však podmienené konceptom tzv. štatistickej významnosti, ktorá je vyjadrená vo forme známych *p*-hodnôt. Na základe *p*-hodnôt zvykneme rozhodovať o tom, či je možné pripísať existenciu pozorovaných vzťahov či rozdielov medzi skupinami náhodným faktorom, alebo či daný výsledok reflektuje nami predpokladaný systematický jav. Cieľom testovania štatistickej významnosti je pritom konfrontovať pozorovaný výsledok s nulovou hypotézou (*H*0), ktorú predstavuje teoretická (napr. *z*, *t*, *F*, χ^2) pravdepodobnostná distribúcia, ktorá je realizáciou náhodného procesu. Z tohto dôvodu zvykneme hovoriť o testovaní významnosti nulovej hypotézy (známe ako NHST; null hypothesis significance testing). Ak je náš výsledok vzhľadom na stred nulovej distribúcie výrazne deviantným (napr. > 1,96 *SD* v prípade normálnej distribúcie, s asociovaným $p = 0{,}05$), je praxou zamietnuť *H*0 a zároveň akceptovať platnosť alternatívnej hypotézy *Ha* postulujúcej existenciu predpokladaného efektu. Táto a rôzne ďalšie variácie procedúr štatistickej inferencie

Došlo: 2. 11. 2015; I. R., Prešovská univerzita, Katedra predškolskej a elementárnej predagogiky a psychológie, Ul. 17. novembra 1, 081 16 Prešov, Slovenská republika; e-mail: ivan.ropovik@unipo.sk

sú natoľko späté s ideou psychológie ako kvantitatívnej vedy, že o ich spochybnení sa pri analýze výsledkov zväčša ani neuvažuje.

V prvom kvartáli roku 2015 však vo svete vedy silne zarezonoval úplný zákaz *p*-hodnôt, ako aj asociovaných nástrojov štatistickej inferencie v impaktovanom psychologickom časopise *Basic and Applied Social Psychology* (Trafimow, Marks, 2015). Spolu s *p*-hodnotami boli zakázané aj ostatné súvisiace nástroje inferenčnej štatistiky ako intervaly spoľahlivosti, štandardné chyby, *t*-, *F*-hodnoty a podobne. Editori zdôvodnili svoje rozhodnutie tým, že testy významnosti sú nevalidné a vyžadujú ich nahradenie výlučne deskriptívnou štatistikou a veľkosťami efektov (effect sizes). Hoci sa zaiste nedá hovoriť o trende, tento časopis nie je vo svojej snahe odpútať sa od *p*-hodnôt sám. Časopis *Epidemiology* zakázal používanie *p*-hodnôt pred viac ako dekádou. *Psychological Science*, najvplyvnejší empirický časopis v odbore psychológie zverejnil v roku 2014 novú redakčnú politiku vyžadujúcu nahradenie *p*-hodnôt intervalmi spoľahlivosti. Neuspokojivá prax v používaní *p*-hodnôt dokonca viedla najvplyvnejšiu štatistickú organizáciu American Statistical Association, aby vôbec prvýkrát vo svojej 177-ročnej histórii zaujala stanovisko k otázke štatistickej praxe s cieľom jasne pomenovať viaceré interpretačné problémy *p*-hodnôt (Wasserstein, Lazar, 2016). Ak k tomu pridáme fakt, že veľká väčšina publikovaných prác na tému NHST je značne kritická, musí byť zrejmé, že NHST nie je bez problémov.

P-hodnoty ako primárny výstup NHST sú totiž často považované za niečo, čím nie sú, keďže v skutočnosti odpovedajú na otázku, ktorú sa výskumníci nezvyknú pýtať. Naviac, ich korektná interpretácia zďaleka nie je natoľko intuitívna, ako by sa mohlo zdať, a je pomerne náročné korektne vysvetliť ich úlohu v kontexte vedeckej inferencie. Existuje viacero logicko-interpretačných problémov *p*-hodnôt. Totiž, (1) *p* nevyjadruje pravdepodobnosť, že príčinou výsledku bola náhoda, (2) *p* nie je pravdepodobnosťou platnosti nulovej hypotézy vzhľadom na pozorovaný výsledok, rovnako ako nie je pravdepodobnosťou platnosti alternatívnej hypotézy, (3) *p* nie je pravdepodobnosťou chyby I. typu a (4) *p* nie je ani doplnkom pravdepodobnosti, že výsledok je reprodukovateľný. Význam *p*-hodnôt je tak zvyčajne výrazným spôsobom preceňovaný a ich zautomatizovaná chybná interpretácia neoprávnene nahrádza potrebu hlbšieho vhľadu do výsledkov.

Cieľom tohto príspevku je analyzovať najviac pervazívne chybné interpretácie a zároveň korektne vysvetliť skutočný význam *p*-hodnôt. Na druhej strane kompilácia kritických ohlasov týkajúcich sa používania *p*-hodnôt nie je vzhľadom na hojnosť takto orientovaných zdrojov náročnou úlohou. Niektoré kritické argumenty proti používaniu *p*-hodnôt však nemajú absolútnu platnosť a ich opodstatnenosť je pri hlbšom skúmaní ich premís spochybniteľná. Druhým aspektom tejto analýzy je teda ponúknuť racionálny, pragmatický, ale zároveň rigorózny pohľad na logiku *p*-hodnôt a v závere syntetizovať predkladané fakty a argumenty do formy praktických odporúčaní pre empirický výskum v psychológii, ako aj v iných sociálnych a behaviorálnych vedách.

Vo svetle aktuálne silno rezonujúcej krízy reprodukovateľnosti – kde sa napríklad v oblasti psychológie nepodarilo reprodukovať takmer 2/3 z výskumných zistení publikovaných vo vysoko impaktovaných časopisoch (Open Science Collaboration, 2015) – v prípade *p*-hodnôt už nejde iba o tému technickej povahy ale práve naopak, o problematiku zásadného významu. Na *p*-hodnotách je totiž najmä v sociálnych vedách podložené vedecké poznanie vôbec, resp. formálne rozhodovanie o platnosti hypotéz, a nesprávne používanie *p*-hodnôt je preto jedným z hlavných dôvodov slabej reprodukovateľnosti psychologických výskumov (Johnson, 2013), a teda preukázateľne značným ohrozením integrity psychologických vied.

V súčasnosti používaná procedúra NHST je v skutočnosti zmesou dvoch konceptuálne nekompatibilných prístupov k štatistickej inferencii (Perezgonzalez, 2015). Koncom 20. rokov minulého storočia totiž začal byť pôvodný prístup Ronalda Fishera vytláčaný novým prístupom Jerzyho Neymana a Egona Pearsona, ktorého cieľom bolo testovanie kvantitatívnych hypotéz čo najviac objektivizovať a umožniť rozhodovanie o ich konečnej platnosti. Pre Fishera *p*-hodnoty slúžili iba na predbežné rozlíšenie systematických vplyvov od tých nesystematických, pričom poslaním každého experimentu bolo dať dátam šancu poukázať na neplatnosť *H*0 (Fisher, 1966, s. 16). Fisher pritom považoval testovanie *H*0 za jednu z hierarchicky najelementárnejších procedúr, ktorá mala svoje opodstatnenie iba pri riešení nových a pre výskumníka dosiaľ neznámych problémov (Gigerenzer et al., 1989). Pre Fishera výsledok jediného štatistického testu nebol zďaleka dostatočným empirickým dôkazom pre definitívne rozhodnutie o akejkoľvek hypotéze, pričom vravel, že „vedecký fakt by mal byť považovaný za experimentálne preukázaný, iba ak náležite naplánovaný experiment zriedka nedosiahne danú úroveň významnosti“ (Fisher, 1926, s. 504). Vo Fisherovom systéme štatistickej indukcie však nebola žiadna zmienka o zamietaní, či potvrdzovaní hypotéz, *Ha*, chybách I. a II. radu či štatistickej sile. Rovnako, pravidlo $p < 0,05$ bolo spomenuté iba ako návrh, bez nároku na univerzálnu platnosť. Nízka *p*-hodnota pre Fishera znamenala, že daný experimentálny výsledok je hodný bližšej pozornosti, najmä vo forme replikácie, no nič viac. Neyman a Pearson naproti tomu prišli s deduktívnym, skôr mechanickým a teoreticky štruktúrovaným prístupom (Neyman, 1942), ktorý pre svoju interpretáciu vyžadoval formulovanie hypotetickej série replikácií daného experimentu, obsahoval už explicitnú definíciu alternatívnej hypotézy (*Ha*) a koncepty ako hladina α, β či štatistická sila, a na druhej strane marginalizoval význam samotných *p*-hodnôt. Tento formálny aparát následne umožňoval s presne definovateľnými mierami chybovosti rozhodovať o platnosti hypotéz (Neyman, 1942) – niečo, čo Fisher striktne odmietal a považoval za nevedecké, keďže pre neho boli testy významnosti iba „dočasné a prehodnotiteľné“ (Fisher, 1956, s. 99).

Konceptuálne rozpory v daných prístupoch v krátkom čase vyústili do ostrého osobného sporu (Nuzzo, 2014). Napríklad Fisher považoval používanie konvenčnej hladiny významnosti (α) za absurdne akademické. Argumentoval pritom, že dlhá séria opakovania toho istého experimentu je možno realizovateľná v priemyselných procedúrach stanovovania kvality (čo bol aj pôvodný zámer Neymana a Pearsona), ale nie vo vede (Fisher, 1956). Pravdivosť špecifickej hypotézy totiž nie je náhodnou premennou. Tá je buď pravdivá, alebo nepravdivá a nie je tu možné z hľadiska frekventistickej štatistiky hovoriť o dlhej sérii identických experimentov. Podľa Fishera je úlohou výskumníka komunikovať exaktnú *p*-hodnotu dosiahnutú v danom experimente, nie však zamietať či potvrdzovať akékoľvek hypotézy presahujúce rámec aktuálneho experimentu.[1] Na pozadí tohto vlečúceho sa konfliktu však začali vznikať učebnice štatistiky, ktoré tieto konceptuálne nekompatibilné prístupy spojili do formy problematickej procedúry, ktorú dnes poznáme ako NHST (Nuzzo, 2014). V súčasnosti je teda NHST fúziou, ktorá z procedurálneho hľadiska nasleduje prístup Neyman-Pearson, no s interpretačnou logikou prístupu Ronalda Fishera (Hubbard, 2004). Hoci oba tieto prístupy samy osebe napĺňajú podmienku logickej a formálnej integrity, v prípade ich

[1] Dichotomické uvažovanie vychádzajúce z pravidla $p = 0,05$ je samo osebe problematické. S *p*-hodnotami 0,04 a 0,001 sa nakladá rovnako, hoci sú veľmi rozdielne. To isté platí napr. v prípade *p*-hodnôt 0,06 a 0,60. Na druhej strane, hodnoty 0,04 a 0,06 sú vnímané ako kvalitatívne úplne rozdielne, hoci sú takmer identické (Cortina, Landis, 2011; Frick, 1996). Často tak rozdiel medzi „signifikantný“ a „nesignifikantný“ sám osebe signifikantný nie je (Gelman, Stern, 2006).

zlúčenia do procedúry NHST to neplatí. Nesprávne chápanie niektorých princípov týchto prístupov v kombinácii so zamieňaním nesúrodých konceptov vyprodukovalo na poli aplikovaných disciplín viacero intuitívnych, interpretačne lákavých, no chybných presvedčení, ktoré podkopávajú pravdivosť, oprávnenosť a reprodukovateľnosť vedeckých výsledkov.

Kritika procedúry NHST trvá už od jej počiatkov (Perezgonzalez, 2015), pričom sa zameriava na dva aspekty: interpretačný a formálno-logický. Cieľom nasledujúcej diskusie bude primárne analyzovať otázky interpretácie *p*-hodnôt a zároveň konvergovať k hodnotiacim záverom ohľadom užitočnosti a korektného použitia *p*-hodnôt.

P-hodnota = $p(\mathrm{D}|H0) \neq p(H0|\mathrm{D})$

Najrozšírenejším a značne pervazívnym nesprávnym presvedčením týkajúcim sa *p*-hodnôt je, že *p*-hodnota vyjadruje pravdepodobnosť platnosti nulovej hypotézy $p(H0)$. Poznať túto nepodmienenú pravdepodobnosť ale samozrejme nie je možné, keďže pravdivosť $H0$ nie je priamo pozorovateľná. Pravdepodobnosť platnosti $H0$ je tak vždy podmienená platnosťou nášho pozorovania (t. j. dát), a preto naše očakávanie, že *p*-hodnota vyjadruje pravdepodobnosť $H0$, sa v skutočnosti vzťahuje na tzv. podmienenú pravdepodobnosť $p(H0|\mathrm{D})$, t. j. pravdepodobnosť platnosti $H0$ za predpokladu platnosti našich dát (resp. daného výsledku). Takéto očakávanie je však nesprávne, keďže *p*-hodnota nie je pravdepodobnosťou vzťahujúcou sa na $H0$ či akúkoľvek inú hypotézu, ale vzťahuje sa na inverznú pravdepodobnosť, t. j. na hypotetickú frekvenciu výskytu pozorovaných a extrémnejších dát (D) za predpokladu platnosti $H0$, ďalej zjednodušene označovanú ako $p(\mathrm{D}|H0)$. Pravdepodobnosť platnosti nulovej hypotézy $p(H0)$ sa teda pri kalkulácii *p*-hodnôt z definície rovná 1.[2] Hoci sa na prvý pohľad môže zdať, že rozdiel medzi $p(H0|\mathrm{D})$ a $p(\mathrm{D}|H0)$ je z významového a pragmatického hľadiska nepodstatný, opak je pravdou. *P*-hodnota, t. j. $p(\mathrm{D}|H0)$ je totiž spravidla odlišná od $p(H0|\mathrm{D})$, pričom nezriedka môže ísť o rádový rozdiel a sú situácie, keď $p(\mathrm{D}|H0)$ konverguje k 1, zatiaľ čo $p(H0|\mathrm{D})$ sa blíži k hodnote 0 (Lindley, 1957). Napríklad predpokladajme, že D = *úmrtie* a $H0$ = *pád lietadla*. Aká je pravdepodobnosť, že človek zahynie za predpokladu pádu lietadla, $p(\mathrm{D}|H0)$? Značne vysoká. Na druhej strane, aká je pravdepodobnosť, že človek sa stal obeťou pádu lietadla, ak vieme, že zomrel, t. j. $p(H0|\mathrm{D})$? Táto pravdepodobnosť je takmer 0, keďže drvivá väčšina ľudí zomiera inak ako pri páde lietadla.

Na ilustráciu nezmyselnosti predpokladu, že $p(\mathrm{D}|H) = p(H|\mathrm{D})$ bývajú často uvádzané príklady z diagnostickej praxe. Napríklad, spomedzi všetkých ľudí, ktorí prichádzajú na zdravotnú diagnostiku metódou „D“, 1 % má chorobu „H“ [$p(H1)$]. Výsledok diagnostiky je pozitívny u 99 % ľudí s chorobou H [$p(\mathrm{D}|H1)$], ale aj u 4,9 % zdravých ľudí [$p(\mathrm{D}|H0)$]. Predstavme si, že človek prichádza na diagnostiku a výsledok je pozitívny. Aká je potom pravdepodobnosť, že má chorobu H, resp. $p(H1|\mathrm{D})$?

Keďže pravdepodobnosť diagnostikovania choroby H u zdravého človeka je 0,049, čiže $p < 0{,}05$, a výsledok testu bol pozitívny, na základe bežne aplikovanej logiky NHST môžeme chybne konštatovať, že osoba má chorobu H, a to viac než s 95% pravdepodobnosťou. Hodnota $p(H1|\mathrm{D})$ sa však v skutočnosti neblíži hodnote $p(\mathrm{D}|H1)$, t. j. 0,99. Dokonca ani len nie je pravdepodobné, že daný človek je chorý, ak je nález pozitívny. V skutočnosti je táto pravdepodobnosť $p(H1|\mathrm{D})$ na úrovni 0,17. Môže sa to javiť málo uveriteľné, ale napriek – na pomery sociálnych vied – solídnej senzitivite a špecificite (0,99, resp. 0,951) daná diagnostická procedúra produkuje v dôsledku

[2] *P*-hodnoty teda nemôžu formálne vyjadrovať pravdepodobnosť, že daný výsledok je dielom náhody, t.j. $p(H0|\mathrm{D})$, keďže ich samotná kalkulácia je postavená na predpoklade náhodnej distribúcie.

nízkej celkovej prevalencie choroby H v populácii (t. j. nízkej pravdepodobnosti $H1$) veľký počet falošne pozitívnych prípadov, čo je ľahko demonštrovateľné pri premietnutí pravdepodobností do formy počtov (viď napr. Cohen, 1994).

Jediný z uvedených údajov, ktorý máme na základe NHST k dispozícii, je $p(D|H0)$, t. j. p-hodnota, ktorá je v našom príklade na úrovni 0,049. Presvedčenie, že za predpokladu pozorovania takýchto deviantných dát (pozitívny nález, vysoká hodnota testovej štatistiky) svedčí tento údaj o nízkej pravdepodobnosti $p(H0|D)$, môže byť celkom ľahko úplne chybné. V našom príklade pravdepodobnosť, že človek je v skutočnosti zdravý, ak je výsledkom diagnostiky pozitívny nález, je $p = 0{,}83$ a nie $p = 0{,}049$. Poznanie $p(H0|D)$ je ale prístupné iba prostredníctvom Bayesovej teorémy (Bayes, 1763), ktorá definuje, akým smerom a o koľko máme revidovať náš odhad pravdepodobnosti vo svetle nových empirických dôkazov. P-hodnoty však iba vyjadrujú pravdepodobnosť dát za predpokladu špecifickej $H0$, no neumožňujú spätné závery o pozorovanej realite (Nuzzo, 2014).

Formálne tak iba na základe p-hodnôt nie je v rámci daného experimentu možné potvrdiť platnosť $H0$ ani Ha. Ak $H0$ neplatí, tak distribúcia p-hodnôt získaných mnohonásobným replikovaním experimentu je pozitívne zošikmená. Pritom platí, že čím väčšia je štatistická sila, tým je zošikmenie distribúcie výraznejšie. V prípade platnosti $H0$ je distribúcia p-hodnôt uniformná, kde všetky hodnoty v intervale (0,1) sú rovnako pravdepodobné (Hung et al., 1997). Každá p-hodnota, ktorú získame v rámci reálneho výskumu, teda môže pochádzať z oboch distribúcií ($H0$, alebo Ha), rozdiel je len v tom, že výskyt nižších p-hodnôt je pravdepodobnejší v pozitívne zošikmenej distribúcii (ak $H0$ neplatí). Zároveň platí, že testová štatistika (napr. t, F) vo všeobecnosti narastá spolu s množstvom empirických dôkazov (primárne s veľkosťou vzorky), čo v konečnom dôsledku znižuje výslednú p-hodnotu.[3] Zvyšovaním N sa tak zvyšuje pravdepodobnosť oprávneného zamietnutia $H0$, čo poskytuje istú empirickú podporu pre vyhlásenie jej neplatnosti.

Ak však $H0$ platí, tak testová štatistika nekonverguje k limite ani v prípade limitnej veľkosti N a štatistický test produkuje bez ohľadu na N náhodnú p-hodnotu z uniformnej distribúcie. Zvyšovaním N v prípade platnosti $H0$ tak nie je možné poskytnúť dôkaz o platnosti $H0$, lebo N v tom prípade nie je v žiadnom vzťahu k distribúcii p-hodnôt. To je ten dôvod, prečo už Fisher trval na tom, že v rámci frekventistickej štatistiky nie je možné poskytnúť dôkaz pre potvrdenie $H0$ (zvykne sa iba konštatovať neschopnosť zamietnutia $H0$). Ak aj na základe p-hodnoty zamietneme možnosť, že pozorovaný výsledok je dôsledkom náhody, tak interpretácia predpokladaného kauzálneho činiteľa je možná iba po vylúčení iných systematických efektov. To už ale nie je otázkou štatistiky, ale vecou zodpovedajúceho experimentálneho plánu (Chow, 1998).

Vzťah medzi $p(D|H0)$ a $p(H0|D)$

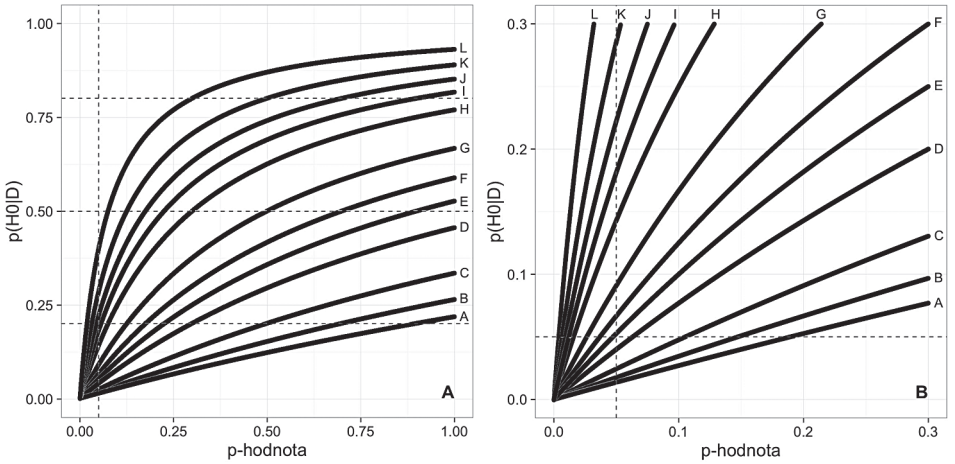
Je pravdou, že neexistuje priamočiary, formálne definovateľný vzťah medzi $p(D|H0)$ (p-hodnotami) a $p(H0|D)$ (tým, čo chceme zistiť) a existuje množstvo situácií, kde nízka hodnota $p(H0|D)$ nie je asociovaná s nízkou hodnotou $p(D|H0)$. Fakt, že nízka hodnota $p(D|H0)$ vo väčšine prípadov znižuje $p(H0|D)$ totiž neznamená, že $p(H0|D)$ musí byť nízka. Z tohto dôvodu viacerí autori (viď Falk, Greenbaum, 1995) považujú NHST za nevalidnú a zavádzajúcu procedúru. Na základe výsledkov simulačných štú-

[3] NHST de facto penalizuje za slabiny vo výskumnom pláne (Cortina, Landis, 2011). Totiž odhad efektu, ktorý je založený na relatívne malej vzorke, nie je považovaný za dôveryhodne preukázaný, keďže pri nízkom N je od vzorky k vzorke možné očakávať prílišnú náhodnú variabilitu.

dií je však možné konštatovať, že akékoľvek kategorické zamietnutie validity *p*-hodnôt pre posúdenie $p(H0|D)$ je do istej miery tendenčným. Vzťah medzi $p(D|H0)$ a $p(H0|D)$ totiž preukázateľne existuje napriek tomu, že nie je formálne definovateľný. Nakoľko sú teda *p*-hodnoty dobrým indikátorom $p(H0|D)$, ak náhodne variujú ostatné relevantné hodnoty, t.j. apriórna pravdepodobnosť nulovej hypotézy $p(H0)$ a populačná hodnota štatistickej sily? Trafimow a Rice (2009) v Monte-Carlo simulačnej štúdii naprieč generovanou sadou 65000 rôznych dátových setov zistili, že korelácia medzi $p(H0|D)$ a $p(D|H0)$ je na úrovni $r = 0{,}396$, čo predstavuje cca 16 % zdieľanej variancie. Na základe tohto výsledku daní autori konštatovali neužitočnosť *p*-hodnôt pre odhad $p(H0|D)$. Problémom takéhoto odhadu (lineárnej korelácie) však je, že neberie do úvahy nelineárnosť vzťahu daných dvoch premenných. Odhalenie predmetného vzťahu tak nie je veľmi výpovedné, ak ostatné relevantné premenné $p(H0)$ a štatistická sila $p(D|Ha)$ náhodne variujú.

S cieľom zistenia vzťahu medzi *p*-hodnotami a $p(H0|D)$, ale už pre rôzne vopred definované kombinácie apriórnej pravdepodobnosti $p(H0)$ a štatistickej sily $p(D|Ha)$, boli v prostredí softwaru R (Lakens, 2015) realizované simulácie 50000 dátových setov osobitne pre každú z 12 kombinácií – $p(H0) = [0{,}2; 0{,}5; 0{,}8]$ a $p(D|Ha) = [0{,}3; 0{,}5; 0{,}7; 0{,}9]$. Ako je možné vidieť na grafe 1A, vzťah medzi *p*-hodnotami a $p(H0|D)$ má pri fixných úrovniach $p(H0)$ a štatistickej sily logaritmický charakter. Zároveň je pri priblížení toho istého grafu (1B) zrejmé, že *p*-hodnoty sa blížia obdobnej podmienenej pravdepodobnosti nulovej hypotézy $p(H0|D)$, ak je (1) $H0$ nepravdepodobná už pred začiatkom experimentu (krivky A–D) alebo (2) v prípade rovnakej apriórnej pravdepodobnosti $H0$ a Ha a zároveň dostatočnej štatistickej sily (krivky E–F). Ak však chceme rigoróznejšie testovať existenciu istého efektu a z hľadiska vedeckého skepticizmu apriórne predpokladáme vyššiu pravdepodobnosť $H0$, tak $p(H0|D)$ býva podstatne vyššia ako získaná *p*-hodnota (krivky I–L), a to aj pri vysokých hodnotách štatistickej sily.

Rovnako aj iné simulačné štúdie ukazujú, že naprieč širokou distribúciou možných dát najpravdepodobnejšia hodnota $p(H0|D)$ pre $p = 0{,}05$ variuje medzi 0,13 a 0,29, pričom pre $p = 0{,}001$ je to medzi 0,04 a 0,09 (Sellke, Berger, 1987). *P*-hod-



Graf 1 Vzťah medzi *p*-hodnotami a $p(H0|D)$ pre kombinácie 3 úrovní $p(H0)$ a 4 úrovní štatistickej sily
$p(H0) = 0{,}2$ a sila (A) 0,9 (B) 0,7 (C) 0,5 (D) 0,3. $p(H0) = 0{,}5$ a sila (*E*) 0,9 (*F*) 0,7 (*G*) 0,5 (*H*) 0,3. $p(H0) = 0{,}8$ a sila (*I*) 0,9 (*J*) 0,7 (*K*) 0,5 (*L*) 0,3.

nota 0,05 tak zodpovedá obdobnej pravdepodobnosti *H*0 za predpokladu platnosti dát iba v prípade takmer limitných, veľmi silných dát, pričom vo veľkej väčšine prípadov $p = 0{,}05$ bude asociované s podstatne vyššou hodnotou $p(H0|D)$ (Dickey, 1977; Lindley, 1993).

Aj keď vzťah medzi *p*-hodnotami, t.j. $p(D|H0)$, a silou empirických dôkazov svedčiacich proti *H*0 nie je konzistentne a formálne definovateľný, vo väčšine reálnych aplikácii možno *p*-hodnoty považovať za nepriamy, no dobrý indikátor $p(H0|D)$. Ich potenciál na dokladovanie nízkej podmienenej pravdepodobnosti *H*0 je ale podstatne slabší, ako sa vo všeobecnosti predpokladá. Najmä za predpokladu vysokej štatistickej sily ale možno považovať nízku *p*-hodnotu za pomerne spoľahlivý indikátor relatívne nízkej hodnoty $p(H0|D)$ (Baril, Cannon, 1995; McGraw, 1995). Tento predpoklad je v mnohých prípadoch možné považovať za rozumný. Stále však platí, že *p*-hodnoty nie sú spoľahlivým formálnym dôkazom pre vyvodenie akýchkoľvek záverov o $p(H0|D)$ (Nickerson, 2000).

P-hodnoty, chyba I. radu a hladina α

Nie je to len samotný koncept *p*-hodnôt, ktorý býva chybne interpretovaný, ale problémom je aj jeho odlíšenie od na prvý pohľad príbuzných konceptov, a to hladina α a chyba I. radu. V samotnej štatistickej literatúre sa často uvádza, že hladina α je matematickým zápisom chyby I. radu, čo nie je pravda. Ako je dobre známe, chyba I. radu je chybným zamietnutím pravdivej *H*0. Naproti tomu, α je teoretickou pravdepodobnosťou dopustenia sa chyby I. radu *za predpokladu, že H*0 *je v dlhej sérii opakovaní vždy pravdivá*. Teda, α je fixne stanovená hodnota, ktorá sa nikdy nemení, a to bez ohľadu na to, koľkokrát je *H*0 v sérii experimentov pravdivá, alebo nepravdivá. Ak stanovíme, že $\alpha = 0{,}05$, tak tým vyjadrujeme akceptáciu nasledujúcej miery chybovosti: ak v každom z napr. 100 realizovaných výskumov (je jedno, či rovnakých, alebo rozdielnych) bude *H*0 pravdivá, tak očakávame, že v 5 prípadoch túto hypotézu chybne zamietneme. Naproti tomu pravdepodobnosť chyby I. radu hovorí o niečom inom. Tá už nie je fixnou hodnotou, ale skutočnou proporciou chybovosti s ohľadom na pravdivosť *H*0. Ide tak o produkt hodnoty α a pravdivosti *H*0. Totiž, nie je možné dopustiť sa chyby I. radu, ak *H*0 neplatí. Keď vezmeme do úvahy, že v značnej proporcii psychologických výskumov *H*0 neplatí, tak skutočná frekvencia chýb I. radu je o mnoho nižšia ako hladina α (Cortina, Dunlap, 1997). Stanovená hladina α je tak vždy hornou nominálnou hranicou pravdepodobnosti chyby I. radu.

Spojenie dvoch konceptuálne rozdielnych prístupov k štatistickej inferencii (Fisher a Neyman-Pearson) spôsobilo, že *p*-hodnota býva často interpretovaná ako flexibilná hladina α a následne ako doplnok pravdepodobnosti, že výsledok je reprodukovateľný. Na rozdiel od α, ktorá je fixnou vlastnosťou štatistického testu, *p*-hodnota je však náhodnou premennou vzťahujúcou sa na konkrétne dáta (Hubbard, 2004). Napríklad, ak výsledkom experimentu je $p = 0{,}03$, vôbec to neznamená, že ak by sme daný experiment opakovali 100-krát, tak v 97 prípadoch zaznamenáme signifikantný efekt. Rovnako, $p = 0{,}05$ neznamená, že daný experimentálny efekt sa neprejaví pravdepodobne iba v 1 z 20 replikácií. Pravdepodobnosť reprodukovateľnosti akéhokoľvek pozorovaného efektu totiž nie je vecou ani *p*-hodnôt, ani hladiny α, ale empirickou otázkou, pričom neexistuje žiadna štatistická procedúra, ktorá by dokázala nahradiť potrebu replikácie.

P-hodnoty a reprodukovateľnosť

Hoci sú *p*-hodnoty zväčša primárnym deskriptorom zozbieraných dát, nebývajú sprevádzané žiadnym odhadom variability (Johnson, 2013), ako je to zvykom v prípade

priemerov, korelácií a podobne. Aj za predpokladu absencie akýchkoľvek nedostatkov výskumného plánu a dodržania všetkých predpokladov použitého štatistického modelu prirodzená variabilita vyplývajúca z podstaty *p*-hodnôt (najmä v prípade skúmania slabších efektov) býva príčinou slabej reprodukovateľnosti výskumných zistení.

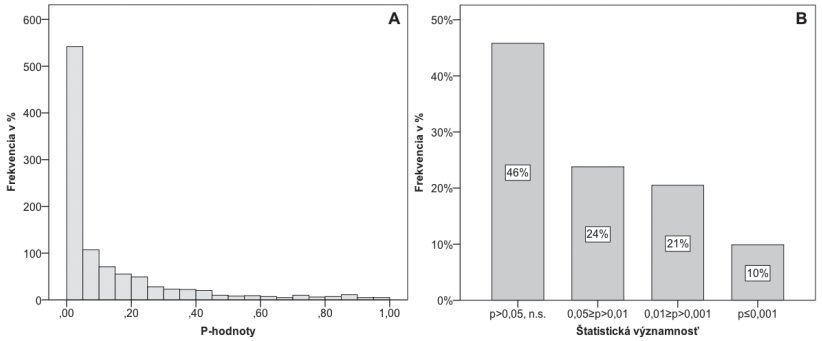
Na ilustráciu nestáleho charakteru *p*-hodnôt a jeho efektu na reprodukovateľnosť je možné uviesť nasledujúci empirický príklad. Predpokladajme, že istý experimentálny zásah vedie k populačnej hodnote priemerného skóre v istej premennej na úrovni $\mu = 57$ ($\sigma = 16$). Zároveň, bez daného experimentálneho zásahu populačná hodnota sledovanej premennej je $\mu = 50$ ($\sigma = 16$). Sila experimentálneho efektu (v jednotkách σ, ekvivalent Cohenovho *d*) je tak na úrovni $\delta = 0{,}44$, čiže existuje 62% pravdepodobnosť[4], že náhodne vybraný subjekt z „experimentálnej populácie" bude mať vyššie skóre v porovnaní s náhodne vybraným subjektom z „kontrolnej populácie". Ak by uvedené populačné parametre zhodou okolností exaktne opisovali charakteristiku dvoch nezávislých výberov, každý o veľkosti $N = 45$, tak daný experimentálny efekt by bolo možné označiť za štatisticky signifikantný; $t(88) = 2{,}1$; $p = 0{,}04$. Štatistická sila pre detekciu signifikantného efektu by v tomto prípade bola na úrovni $1 - \beta = 0{,}54$. Z hľadiska veľkosti efektu a štatistickej sily uvedený príklad predstavuje typickú situáciu v oblasti sociálnej psychológie (Maxwell, 2004; Richard, Bond, Stokes-Zoota, 2003).

Ako je známe, štatistiky výberových súborov sú náhodnými premennými a vzorkovaním populácií spravidla získame hodnoty odlišné od populačných parametrov. Pritom platí, že zvyšovaním počtu replikácií uvedeného experimentu sa bude distribúcia výberových parametrov (*M, SD, t, d*) približovať normálnej distribúcii. Rovnako náhodnými premennými sú preto aj *p*-hodnoty. Raz totiž bude pomer efektu k odhadovanej chybovej variabilite väčší, inokedy menší. Ich distribúcia naprieč množstvom replikácií už avšak nie je normálna, ale jej tvar závisí od populačnej veľkosti štatistickej sily vyplývajúcej z použitého výskumného plánu. Čím sú populačná veľkosť efektu a veľkosť *N* väčšie, tým je distribúcia *p*-hodnôt výraznejšie pozitívne zošikmená, čo značí vyššiu pravdepodobnosť nízkej *p*-hodnoty, ak sú dve populácie vzhľadom na sledovanú premennú skutočne odlišné (ak *H*0 neplatí). Aká je však pravdepodobnosť, že náhodným vzorkovaním a následným porovnaním priemerov získame v našom modelovom prípade $p = 0{,}05$ správne indikujúc prítomnosť experimentálneho efektu? Alebo, ak by sa štatistiky prvého výberového súboru zhodovali s populačnými parametrami a výsledkom by bolo dosiahnutie štatistickej významnosti (v tomto konkrétnom prípade $p = 0{,}04$), aká je pravdepodobnosť reprodukovateľnosti takéhoto zistenia v ďalšom výskume?

Na zodpovedanie týchto otázok bolo simulovaných 1000 párov výberových súborov náhodným vzorkovaním ($N_{ES} = N_{KS} = 45$) z apriórne definovaných populácií ($\mu_{ES} = 57$, $\sigma = 16$; $\mu_{KS} = 50$, $\sigma = 16$). Ak výsledky simulácie premietneme do distribúcie veľkostí efektov vyjadrených Cohenovým *d*, môžeme pozorovať normálnu distribúciu so značnou variabilitou, keďže v rámci $\pm$ 1*SD* (cca 68 % teoretickej *z*-distribúcie) sa nachádzajú veľkosti efektov *d* v intervale (0,22; 0,64). Z hľadiska štatistickej inferencie je však zaujímavejšia distribúcia asociovaných *p*-hodnôt (Graf 2A). V súlade s očakávaním rozdielu medzi skupinami v populačnom parametri priemeru v kombinácii s dostatočnou štatistickou silou generoval model výrazne pozitívne zošikmenú distribúciu *p*-hodnôt s centrálnou tendenciou na úrovni $Mdn(p) = 0{,}038$. Ak však túto

[4] Ide o mieru veľkosti efektu zvanú „common language effect size" (*CL*), ktorá sa vypočíta ako kumulatívna distribučná funkcia (CDF) *z*-distribúcie pre hodnotu $(\frac{\delta}{\sqrt{2}})$, kde δ je odhad populačnej hodnoty Cohenovho *d* (McGraw, Wong, 1992).

spojitú distribúciu *p*-hodnôt rozkategorizujeme v súlade s konsenzuálne používanými hladinami významnosti (graf 2B), môžeme konštatovať nasledujúce. Skutočnosti adekvátny empirický dôkaz, t. j. *p*-hodnotu v intervale (0,01; 0,05), získame v 24 % replikácií experimentu. V cca 30 % (20,5 + 9,9) realizácií takéhoto experimentu získame dáta, ktoré taktiež poskytnú pre daný experimentálny efekt empirickú podporu (*p*-hodnotu v intervale (0; 0,01]), no výrazne silnejšiu, ako by bolo vzhľadom na skutočnosť adekvátne. Vo zvyšných 46 % prípadov sa konštatovaním neexistencie experimentálneho efektu (*p* v intervale (0,05; 1)) dopustíme chyby II. radu. Hoci tento výsledok je zrejmý už z výpočtu štatistickej sily, môže byť zarážajúce, že veľká proporcia (cca 73 %) nesignifikantných výsledkov je výraznejšie (*p* v intervale [0,1; 1)) vzdialená od hladiny $p = 0,05$. Napríklad, hodnotu $p > 0,20$ získame v 47 % nesignifikantných výsledkov, čo predstavuje viac ako 1/5 všetkých replikácií experimentu.



Graf 2 Distribúcia *p*-hodnôt (2A) a distribúcia hladín *p*-hodnôt (2B)

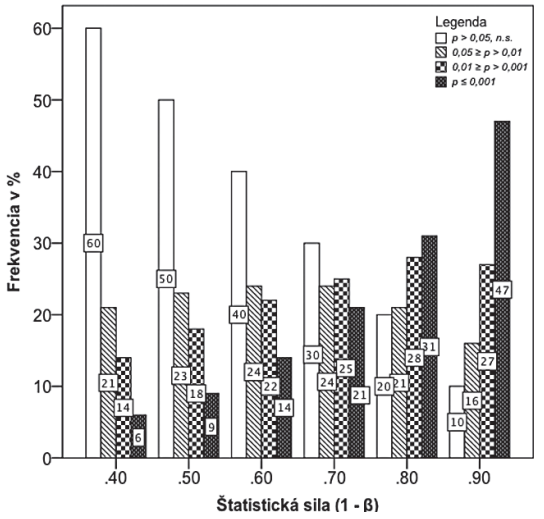
Z uvedeného príkladu vyplýva, že dosiahnutie hladiny $p = 0,05$ predstavuje výrazne slabší empirický dôkaz v prospech experimentálneho efektu, ako sa zvykne predpokladať. Dané konštatovanie platí najmä v prípade výskumných plánov disponujúcich nízkou štatistickou silou, ako je bežné napríklad v oblastiach klinickej psychológie či neuropsychológie.

To však nie je jediným problémom, ktorý vyplýva z interpretácie výsledkov založených na kritériu $p = 0,05$ v štúdiách s nízkou štatistickou silou. Ak totiž v prípade štúdie, ktorá nemá adekvátnu štatistickú silu, zistíme signifikantný efekt, ten bude pravdepodobne umelo „nafúknutý“ (Ioannidis, 2008). Napríklad v prípade našej štúdie s pravou hodnotou štatistickej sily $1 - \beta = 0,54$ priemerná výberová hodnota rozdielu medzi experimentálnou a kontrolnou skupinou naprieč 1000 replikáciami bola na úrovni $\bar{x}(\Delta) = 9,34$, hoci populačná hodnota rozdielu je iba $\Delta = 7$ (57 – 50). To značí 34% infláciu, pričom odhadovaná veľkosť efektu bola v 91 % replikácií väčšia ako populačný efekt. V prípade sily na úrovni 0,20 (bežné v oblasti neuropsychológie, pozri Button et al., 2013), priemerná odhadovaná sila efektu by bola takmer dvojnásobná. Dôvodom je fakt, že štatistický test bude s väčšou pravdepodobnosťou signifikantný v prípade menšieho počtu experimentov, ktorých výsledkom bola nadpriemerná veľkosť efektu. Pri nízkej štatistickej sile tak dáta často nie sú schopné dokladovať existenciu skutočného efektu, a ak aj áno, jeho veľkosť bude veľmi pravdepodobne „nafúknutá“ (Colquhoun, 2014; Ioannidis, 2008).

Vo všeobecnosti platí, že reprodukovateľnosť *p*-hodnôt rastie spolu so štatistickou silou výskumného plánu. Na grafe 3 je možné vidieť teoretické distribúcie hladín *p*-hodnôt pre rôzne úrovne štatistickej sily. Tieto distribúcie platia univerzálne pre všetky kombinácie veľkosti vzorky *N* a veľkosti efektu *d*. Problémom ale je, že po-

zorovaná (tzv. post-hoc) štatistická sila je iba prostou matematickou transformáciou pozorovanej *p*-hodnoty. Pravá štatistická sila býva neznáma, keďže nepoznáme populačnú hodnotu veľkosti efektu. Preto nie je formálne možné na základe pozorovanej sily kalkulovanej na základe dát z výberového súboru vyslovovať akékoľvek závery týkajúce sa pravdepodobnosti rozloženia *p*-hodnôt v prípade konkrétnej výskumnej hypotézy. Naviac, uvedené teoretické distribúcie *p*-hodnôt platia iba vtedy, ak sa dodržia predpoklady použitého štatistického modelu, z ktorých azda najvýznamnejší je predpoklad náhodného vzorkovania (Cuberek, Frömel, 2011). Akékoľvek iné ako náhodné vzorkovanie totiž môže do dát vniesť zdroj neželanej systematickej variancie, ktorý následne skresľuje získané výsledky.

Vo vzťahu k pravdepodobnosti reprodukcie experimentálneho efektu je tak výskumník odkázaný iba na nespoľahlivý induktívny úsudok (ak je založený na jedinej *p*-hodnote), pričom skutočná reprodukovateľnosť je empirickou otázkou, na ktorú inferenčné štatistické postupy nedokážu dať spoľahlivú odpoveď (Halsey et al., 2015).



Graf 3 Teoretické distribúcie hladín p-hodnôt pre rôzne úrovne štatistickej sily

ODPORÚČANIA

Tie najzásadnejšie argumenty proti *p*-hodnotám sa spravidla vzťahujú na prax v ich používaní, nie na koncept ako taký. Za predpokladu korektnej interpretácie a nevyhnutnej akceptácie vyššej miery neistoty, *p*-hodnoty dokážu byť užitočným formálnym nástrojom vedeckej inferencie. Cieľom tejto podkapitoly je preto ponúknuť sadu všeobecne platných rámcových odporúčaní týkajúcich sa používania *p*-hodnôt.

Najdôležitejšou nutnou podmienkou korektnosti akejkoľvek interpretácie *p*-hodnôt je dodržanie predpokladu náhodného priradenia experimentálnych podmienok a predpokladu absencie systematických faktorov produkujúcich chýbajúce dáta (Greenland et al., 2016). Ešte dôležitejší je však fakt, že používanie *p*-hodnôt na rozhodovanie o platnosti hypotéz (pozri Mayo, Spanos, 2001) plní svoju funkciu, t.j. umožňuje udržať kontrolovanú mieru chybovosti v dlhej sérii takýchto rozhodnutí (napr. 5 %), iba za predpokladu, že (a) je dodržaný apriórne stanovený plán vzorkovania populácie, (b) žiadne hypotézy nie sú selektívne formulované až po preskúmaní dát a zároveň (c) arbitrárne rozhodnutia v rámci analýzy a reportovania dát nie sú determinované výsledkami (tzv. *p*-hacking). V prípade eventuálneho uplatnenia rôznych „flexibil-

ných“ ad hoc postupov[5] s cieľom dosiahnutia $p < 0,05$ skutočná miera chýb I. typu v sérii rozhodnutí o platnosti hypotéz prestáva byť známa, jej hodnota však bude spravidla rádovo vyššia ako maximálna nominálna miera chybovosti stanovená hladinou (viď Simmons, Nelson, Simonsohn, 2011).

Transparentnosť a úplné reportovanie dát s explicitným odlíšením konfirmačných analýz (testovanie apriórne stanovených hypotéz) od tých exploračných je základným predpokladom dôveryhodnosti publikovaných výsledkov, pričom sprostredkovanie dát, materiálov či kódu analýz, ako aj predregistrácia výskumného plánu a analytického postupu sa, nasledujúc psychológiu, stáva štandardom v mnohých vedných odvetviach (Nosek et al., 2015; https://cos.io/top). Nad rámec vyššie uvedených nutných podmienok je dobré pridržiavať sa nasledujúcich odporúčaní:

(1) *P*-hodnoty nevyjadrujú pravdepodobnosť platnosti testovanej nulovej hypotézy (*H*0), ale pravdepodobnosť výskytu dát (výsledku), ak je *H*0 pravdivá a všetky predpoklady daného štatistického modelu opisujúceho proces generovania dát sú splnené. Za predpokladu, že nie je dôvod apriórne pokladať *H*0 za veľmi pravdepodobnú, možno *p*-hodnoty považovať za nepriamy indikátor dátami podmienenej pravdepodobnosti existencie efektu. Ak však chceme vzťahovať *p*-hodnoty na platnosť hypotéz, simulačné štúdie (Johnson, 2013; Sellke, Berger, 1987) ukazujú, že $p = 0,05$ nezodpovedá šanci 1 : 19, že pozorovaný efekt je dielom náhody (výberovej chyby), ale má približne šanci jedna ku trom až piatim, 1:3–1:5 (pomer *H*0/*Ha*, tzv. Bayesov faktor). Obdobne v prípade $p = 0,01$ to nie je 1:99, ale ~ 1:12–1:20. Hranične signifikantné *p*-hodnoty ($0,05 > p > 0,01$) spravidla predstavujú výrazne slabší empirický dôkaz proti *H*0 (a to najmä v prípade výskumných plánov disponujúcich vysokou štatistickou silou). Z tohto dôvodu býva navrhované, aby hranica potrebná na vyhlásenie štatistickej významnosti bola na úrovni $p = 0,005$, čo reálne zodpovedá nominálnej šanci ~ 1:25–1:50, že sa výskumník konštatovaním významnosti efektu dopustil chyby I. radu. Konštatovanie existencie testovaných efektov (*Ha*) by tak vyžadovalo viac empirických dôkazov (takmer zdvojnásobenie potrebnej *N*)[6], redukovalo by to však mieru nereprodukovateľnosti zistených efektov z dôvodu výberovej chyby až o 80 % (Johnson, 2013). Samozrejme je úplne korektné aplikovať aj menej striktné kritérium ($0,05 > p > 0,005$). V prípade vzťahovania *p*-hodnôt na pravdepodobnosť hypotéz je však korektné čitateľa explicitne upozorniť na vyššiu pravdepodobnosť chybovosti takéhoto empirického poznania.

(2) *P*-hodnoty je možné s istou rezervou považovať za indikátor sily empirických dôkazov, ich účelom však nie je slúžiť ako ukazovateľ sily, resp. praktickej významnosti či klinického významu efektu vôbec. Výsledok štatistických testov tak musí byť vždy sprevádzaný veľkosťou efektu (napr. *r*, *d*, η^2), čo je dôrazne akcentované aj v rámci aktuálnych štandardov APA. Naviac, keďže populačná hodnota veľkosti efektu je iba odhadom, býva odporúčané taktiež uvádzať niektorý z druhov intervalov spoľahlivosti. Intervaly konštruované na základe frekventistickej teórie intervalov spoľahlivosti (Neyman, 1937) sú však rovnako ako *p*-hodnoty predmetom viacerých chybných presvedčení, no implikácie vychádzajúce z ich formálne korektnej definície[7] sú ešte viac kontraintuitívne (viď Morey et al., 2016).

---

[5] Relevantné je nielen to, aké postupy spracovania dát boli realizované na aktuálnych dátach, ale aj to, aké by boli realizované, ak by dáta boli odlišné (Gelman, Loken, 2014).

[6] Alternatívnym riešením by bolo použitie meracích nástrojov produkujúcich menej chybovej variancie alebo použitie rigoróznejšieho výskumného plánu (vnútrosubjektový plán, lepšia kontrola interferujúcich premenných).

[7] *X*% interval spoľahlivosti parametra je intervalom generovaným procedúrou, ktorá má v dlhej sérii opakovaní vzorkovania populácie [najmenej] *X*% pravdepodobnosť, že bude zahŕňať pravú hodnotu parametra (Neyman, 1937).

(3) *P*-hodnota je kombinovaným ukazovateľom, keďže v zásade vyjadruje pomer veľkosti efektu ku očakávanej štatistickej chybe. Ako bolo konštatované vyššie, sama osebe má teda skôr limitovanú informačnú hodnotu, a preto je dobrou praxou vždy realizovať analýzu *apriórnej* štatistickej sily, pričom cieľom by malo byť odhadnúť schopnosť výskumného plánu detegovať najnižšiu veľkosť efektu, ktorá môže byť považovaná za relevantný empirický dôkaz v prospech formulovanej teórie a na základe toho vopred určiť veľkosť potrebnej vzorky. Úroveň štatistickej sily zároveň umožňuje odhadnúť mieru očakávanej inflácie veľkosti efektu v prípade zistenia štatisticky významného efektu (pozri Colquhoun, 2014). Výpočet pozorovanej (post-hoc) štatistickej sily je zbytočnou matematickou procedúrou, keďže ide iba o monotónnu transformáciu pozorovanej *p*-hodnoty.

(4) Za predpokladu, že konkrétny výsledok štatistického testu je z hľadiska interpretácie dôležitý, je potrebné uvádzať exaktnú *p*-hodnotu. Keďže interval $p = (0; 0{,}05]$ zahŕňa v praxi rádovo rozdielne hodnoty, vágne konštatovanie, že $p < 0{,}05$ nestačí.

(5) Keďže *p*-hodnota (ako všetky štatistiky výberového súboru) je náhodnou premennou, opakovaným výskumom, resp. vzorkovaním základného súboru a následným testovaním tej istej hypotézy dospejeme k rozdielnym *p*-hodnotám. Slabá reliabilita *p*-hodnôt spojená so slabou reprodukovateľnosťou psychologických efektov je tak do značnej miery prirodzeným dôsledkom ich výberovej variability. Pravdepodobnosť reprodukovateľnosti výskumného zistenia je preto empirickou otázkou, ktorá nie je v kompetencii jedinej *p*-hodnoty, ale vecou reálnej replikácie.

(6) Alternatívou k *p*-hodnotám (frekventistická štatistika) je Bayesov faktor, ktorý je možné intuitívne a priamočiaro interpretovať ako relatívnu vierohodnosť pozorovaných dát v prípade dvoch konkurujúcich hypotéz *Ha* a *H*0, pričom na rozdiel od *p*-hodnôt indikuje mieru empirických dôkazov v prospech *Ha* vs *H*0 (Jeffreys, 1961). Bayesov faktor tak vyjadruje, ktorým smerom a o koľko je na základe pozorovaných dát rozumné revidovať naše presvedčenie o platnosti istej hypotézy. Použitím Bayesovských inferenčných metód pritom odpadáva väčšina formálno-logických a interpretačných problémov asociovaných s používaním *p*-hodnôt (Dienes, 2011). Dnes sú pritom voľne k dispozícii verejne prístupné softwarové nástroje zamerané na Bayesovskú inferenciu, ako napr. *JASP, STAN* či rôzne nástroje v rámci programovacieho jazyka *R*.

ZÁVER

Množstvo kritických argumentov proti formálnej platnosti a obvyklej interpretácii NHST (a jeho produktu *p*-hodnôt) je len ťažko spochybniteľných. Totiž, induktívny sylogizmus, na ktorom je NHST postavené, je z dôvodu absencie deterministických zákonitostí, resp. nedostatočnej experimentálnej kontroly z formálneho hľadiska neplatný[8] (Cohen, 1994) a väčšina intuitívnych interpretácií *p*-hodnôt výrazným spôsobom preceňuje ich význam. V skutočnosti formálna validita NHST je obhájiteľná skôr iba pragmaticky a *p*-hodnoty iba do istej miery a nepriamo odpovedajú na otázky, ktoré sa zväčša pýtame, pričom zďaleka nie sú také reliabilné a objektívne, ako sa zvykne predpokladať (Nuzzo, 2014), čo je zdôrazňované aj v rámci oficiálneho stanoviska Americkej štatistickej asociácie k problematike *p*-hodnôt (Wasserstein, Lazar, 2016).

V súčasnosti silnejú hlasy za zákaz používania *p*-hodnôt, ako aj asociovaných ná-

[8] Logické pravidlo modus tollens, $((P \rightarrow Q) \wedge \neg Q) \rightarrow \neg P$, t. j., ak *H*0 *platí,* potom $p > 0{,}05$; zároveň $p < 0{,}05$; teda *H*0 *neplatí*. Ak má konsekvent uvedeného výroku, a teda aj záver pravdepodobnostnú povahu, výroková schéma sa stáva formálne chybnou (pozri Hagen 1997).

strojov inferenčnej štatistiky a ich nahradenie postupmi estimácie (interpretujúc výlučne deskriptívne štatistiky a veľkosti efektov). Eventuálnym zákazom NHST však nezmiznú problémy spojené s chápaním konceptu sily empirických dôkazov či chápaním konceptu výberovej chyby a chýb merania (Cortina, Dunlap, 1997; Savalei, Dunn, 2015). Naviac, nahradenie NHST estimáciou (veľkosti efektu) možno nemusí byť vzhľadom na aktuálny vývojový stupeň psychologickej vedy adekvátne. Hypotézy testované v rámci psychologického výskumu totiž majú spravidla ordinálny charakter, keďže predmety skúmania v psychológii majú natoľko komplexnú povahu, že neumožňujú špecifikovať kvantitatívne hypotézy postulujúce exaktnú veľkosť efektu (Frick, 1996). Z dôvodu arbitrárnosti vo vyvodzovaní záverov (Aký výsledok možno očakávať, ak je teória pravdivá, a aký, ak je nepravdivá?) postupy estimácie nie sú vhodným nástrojom na verifikáciu hypotéz vyplývajúcich z psychologických teórií, ale skôr na ich formuláciu.

Napriek množstvu logických a interpretačných problémov *p*-hodnôt, ktoré boli počas mnohých dekád cieľom ostrej kritiky, NHST stále predstavuje primárny nástroj formálnej inferencie. Práve úzka kompatibilita medzi povahou výskumných otázok v psychológii a procedúrou dichotomického rozhodovania o platnosti hypotéz je pravdepodobne hlavným dôvodom, prečo kritické hlasy nedokázali oslabiť pozíciu NHST v kontexte štatistickej inferencie.

Vo vedných odboroch, ktoré v súčasnom štádiu vývoja nie sú schopné formulovať teórie postavené na exaktnej kvantitatívnej predikcii, nie je veľa formálne ukotvených referenčných bodov, na ktorých možno stavať. Znalosť teoretickej distribúcie náhodnej premennej (t.j. znalosť pravdepodobnosti dát, ak je *H*0 pravdivá) je však určite jedným z nich. Procedúra NHST nedokáže zodpovedať otázky, na ktorých riešenie nebola navrhnutá. *P*-hodnoty tak majú svoj špecifický účel, pričom ich nemožno viniť, že nedokážu suplovať iné štatistické nástroje. V prípade akceptovania rozumných predpokladov a korektného použitia ale dokážu v súčinnosti s ďalšími štatistickými nástrojmi (veľkosti efektov, Bayesovské metódy) poskytnúť dobrý základ pre racionálny induktívny úsudok.

LITERATÚRA

Baril, G. L., & Cannon, J. T. (1995). What is the probability that null hypothesis testing is meaningless? *American Psychologist*, *50*(12), 1098–1099.

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, *53*, 370–418.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376.

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist*, *49*(12), 997–1003.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, *1*(3), 140216.

Cortina, J. M., & Dunlap, W. P. (1997). On the logic and purpose of significance testing. *Psychological Methods*, *2*(2), 161–172.

Cortina, J. M., & Landis, R. S. (2011). The earth is not round (p=.00). *Organizational Research Methods*, *14*(2), 332–349.

Cuberek, R., & Frömel, K. (2011). K problematice výzkumného výběru a testování nulové hypotézy. Československá psychologie, *55*(5), 468–477.

Dickey, J. (1977). Is the tail area useful as an approximate Bayes factor? *Journal of the American Statistical Association*, *72*(357), 138–142.

Dienes, Z. (2011). Bayesian versus orthodox statistics: which side are you on? *Perspectives on Psychological Science*, *6*(3), 274–290.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of a probabilistic misconception. *Theory & Psychology*, *5*(1), 75–98.

Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture for Great Britain*, *33*, 503–513.
Fisher, R. A. (1956). *Statistical methods and scientific inference*. Edinburgh: Oliver and Boyd.
Fisher, R. A. (1966). *The design of experiments (8th ed.)*. Edinburgh: Oliver and Boyd.
Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, *1*(4), 379–390.
Gelman, A., & Loken, E. (2014). The statistical crisis in science. *American Scientist 102*, 460–465.
Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.
Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Kruger, L. (1989). *The empire of chance.* Cambridge: Cambridge University Press.
Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology, 31*(4), 337–350.
Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist*, *52*(1), 15–24.
Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle p-value generates irreproducible results. *Nature Methods*, *12*(3), 179–185.
Hubbard, R. (2004). Alphabet soup: Blurring the distinctions between p's and α's in psychological research. *Theory & Psychology*, *14*(3), 295–327.
Hung, H. M., O'Neill, R., Bauer, P., & Kohne, K. (1997). The Behavior of the p-value when the alternative hypothesis is true. *Biometrics, 53*(1), 11–22.
Chow, S. L. (1998). Précis of statistical significance: rationale, validity, and utility. *The Behavioral and Brain Sciences*, *21*(2), 169–239.
Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*, 640–648.
Jeffreys, H. (1961). *Theory of probability*. Oxford: Oxford University Press.
Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, *110*(48), 19313–19317.
Lakens, D. (2015). pvalue_P(H0|D)_relationship. *R code*. Vyhledané na https://gist.github.com/Lakens/bb060a6044650cd30c8e.
Lindley, D. V. (1957). A statistical paradox. *Biometrika*, *44*, 187–192.
Lindley, D. V. (1993). The analysis of experimental data: The appreciation of tea and wine. *Teaching Statistics*, *15*(1), 22–25.
Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: causes, consequences, and remedies. *Psychological Methods*, *9*(2), 147–163.
Mayo, D., & Spanos, A. (2011). Error statistics. In P. S. Bandyopadhyay & M. R. Forster (Eds.), *Handbook of the philosophy of science, Vol. 7: Philosophy of statistics* (pp. 153–198). London: Elsevier.
McGraw, K. O. (1995). Determining false alarm rates in null hypothesis testing research. *American Psychologist*, *50*(12), 1099–1100.
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*(2), 361–365.
Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.
Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. Philosophical Transactions of the Royal Society of London. Series A. *Mathematical and Physical Sciences*, *236*, 333–380.
Neyman, J. (1942). Basic ideas and some recent results of the theory of testing statistical hypotheses. *Journal of the Royal Statistical Society*, *105*(4), 292–327.
Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301.
Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425.
Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, *506*(7487), 150–152.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716.
Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*, *6*, 223.
Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One hundred years of social psychology quantitatively described. *Review of General Psychology*, *7*(4), 331–363.
Savalei, V., & Dunn, E. (2015). Is the call to abandon p-values the red herring of the replicability crisis? *Frontiers in Psychology*, *6*, 245.

Sellke, T., & Berger, J. O. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, *82*(397), 112–122.
Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.
Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, *37*(1), 1–2.
Trafimow, D., & Rice, S. (2009). A test of the null hypothesis significance testing procedure correlation argument. *The Journal of General Psychology*, *136*(3), 261–270.
Wasserstein, R. L, Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. *The American Statistician, 70*(2), 129–133.

SÚHRN

Vo svetle aktuálnej krízy reprodukovateľnosti vedeckého poznania silnie kritika na adresu p-hodnôt ako výstupu procedúry testovania významnosti nulovej hypotézy (tzv. NHST). Totiž intuitívne interpretácie významu p-hodnôt sú spravidla zavádzajúce, keďže p-hodnoty neposkytujú odpovede na otázky, ktoré sa výskumníci zvyknú pýtať, a ich úloha v kontexte štatistickej inferencie je preto výrazne preceňovaná. Príspevok poukazuje na najviac pervazívne nesprávne interpretácie p-hodnôt, ako aj na štatistické dôvody slabej reprodukovateľnosti zistení postavených na prekročení konsenzuálnej hladiny $p < 0,05$. Zároveň však ponúka kritickú reflexiu na argumenty proti používaniu p-hodnôt a prezentuje pragmatický, no stále rigorózny pohľad na logiku a užitočnosť p-hodnôt.

Reproduced with permission of copyright owner. Further reproduction prohibited without permission.