Nástroje na zlepšenie štatistickej inferencie – analýza *p*-kriviek a ekvivalenčné testovanie

Milan Fico[1]
Inštitút pre výskum práce a rodiny, Bratislava

Tools for Better Statistical Inference. Analysis of *P*-Curves and Equivalence Testing. File-drawer effect, selective data reporting and additional data adjustment often lead to publication bias in social sciences. This text deals with two techniques diminishing consequences of these negative phenomena. *P*-curves can distinguish relevant publications from non-significant ones and help us to select information base for formulation of next research goals. Another presented technique is equivalence testing. Focused on the falsification of hypotheses, it helps better explore trivial results and makes them more informative. Both techniques support informative value of results so that the conclusions of statistical inference are more valuable.
Sociológia 2020, Vol. 52 (No. 4: 323-353)
https://doi.org/10.31577/sociologia.2020.52.4.14

***Key words**: Null hypothesis significance testing; publication bias; file-drawer effect; selective data reporting; hypothesizing after results are known; p-curves, equivalence testing*

Úvod

Testovanie nulových hypotéz (*Null Hypothesis Significance Testing*, *NHST*) patrí medzi najrozšírenejšie postupy vyhodnocovania kvantitatívnych údajov. Štatistická inferencia, pomocou ktorej sa zo vzorky dozvedáme viac o skúmanej populácii, je dlhodobo zaužívanou praxou podstatnej časti publikovaných textov. Oživenie kritickej diskusie v *NHST* prístupe je reakciou najmä na situáciu v spoločenských vedách, kde sa nedarí s mnohými replikáciami (Dienes 2008; Ioannidis 2005; Morey – Lakens 2016). Replikácie sú opakovaním pôvodných textov s rovnakým dizajnom ale s novou vzorkou. Odpovedajú na otázku: *„Čo by sa stalo, ak by sme realizovali rovnakú štúdiu opäť?“.* Cieľom je preskúmať, či záver nebol iba šťastnou náhodou (Cumming 2012, Shmidt 2009). Ak sa zistenia nedarí zopakovať, vzniká podozrenie, že teórie, založené na vyhodnocovaní dát, nezodpovedajú realite. Výsledkom sú publikačné skreslenia. Zo sociálnych vied sa jedná najmä o problém behaviorálnych ekonómov alebo psychológie (Simmons et al. 2011). V sociológii sa publikačné skreslenia môžu vyskytnúť, ak sa na otázky odpovedá experimentálnym dizajnom (Jackson – Cox 2013), alebo ak používame rôzne typy vzoriek bez priameho prístupu k celej populácii. Reprodukcia mylných záverov je potom dôsledkom nedodržiavania pravidiel, za akých *NHST* testovanie poskytuje správne odhady. Jedná sa najmä o používanie postupov, ktoré dodatočne menia

[1] Korešpondencia: Milan Fico, Inštitút pre výskum práce a rodiny, Župné nám. č. 5-6, 812 41 Bratislava, Slovensko. E-mail: Milan.Fico@ivpr.gov.sk

pravidlá hry s cieľom prezentovať zistenia v lepšom svetle (Nosek et al. 2018). Najznámejšími sú selektívne vykazovanie údajov (*p*-hacking) a úprava hypotéz po získaní výsledkov (*HARK*ing – *Hypothesizing After the Results are Known*). Pod paľbu sa dostáva aj nesprávne používanie *p*-hodnôt. Ich chybnej aplikácii sa pripisuje časť viny za reprodukciu nespoľahlivých zistení (American Statistical Association, ASA, 2016; Greenland et al. 2016; Ziliak – McCloskey 2011).

Vyrovnávanie sa s dôsledkami publikačných skreslení prebieha na viacerých frontoch. Hovorí sa o nových postupoch transparentnosti, o preregistračnej revolúcii (Munafó et al. 2017; Nosek et al. 2018). Ak sú kroky preregistrácie dodržané, niektoré odborné časopisy sa vopred zaväzujú zverejniť výsledky. Objektivita poznania sa zvyšuje (Earp 2017; Ioannidis 2005). Vznikajú replikačné projekty, v ktorých sa viacerými nezávislými tímami opakovane realizuje rovnaká štúdia (Social Science Replication Project 2016, Center for Open Science). Do popredia sa dostávajú alternatívy k frekvenčnému spracovaniu údajov, najmä bayesovský prístup, ktorý namiesto *p*-hodnôt využíva bayes faktory a – na rozdiel od *NHST* testovania – formuluje pravdepodobnostné závery priamo o hypotézach (Benjamin et al. 2017, Krushke 2015; Van De Schoot et al. 2014).

V texte prezentujeme techniky, umožňujúce znížiť reprodukciu chybných štúdií a zlepšiť závery štatistickej inferencie. Ako prvé uvádzame najčastejšie používané postupy nekorektného vyhodnocovania údajov. Popisom princípov *NHST* prístupu a používania *p*-hodnôt ilustrujeme, v čom spočíva porušenie pravidiel testovania a ako mu najlepšie predchádzať. Ťažisko je kladené na dve techniky, z ktorých každá predstavuje odlišnú cestu zlepšovania inferenčných odhadov.

Prvou je analýza *p*-kriviek (Simonsohn et al. 2014a, 2015b, 2016). Ide o techniku, ktorá dokáže rozlíšiť štúdie s dôkaznou hodnotou. Ak pri formulácii zámeru hľadáme relevantnú literatúru, na ktorú by bolo možné nadviazať, potom ide o pomôcku, ktorou sa možno riadiť.

Druhou technikou je ekvivalenčné testovanie (*ET*) (Lakens 2017b, 2018a). Jedná sa o štatistický test, ktorý – na rozdiel od *NHST* prístupu – kladie dôraz na lepšie preskúmanie triviálnych zistení. Ide o „kontru“ na preteky za štatisticky významnými výsledkami s cieľom rozhodnúť, či je rozdiel/efekt príliš malý na to, aby bol predmetom nášho záujmu.

Cesty publikačných skreslení a dôležitosť transparentného postupu

Kým sa výsledky dostanú k čitateľom, ide o dlhú cestu s mnohými zákutiami. Netransparentnosť realizovaných krokov znamená väčšie pochybnosti o spoľahlivosti zistení. Riziko publikačných skreslení sa zvyšuje aj v dôsledku existencie filtrov nad databázami publikovaných prác, ktoré môžu nasmerovať

pozornosť nesprávnym smerom. Mylná dôvera v hypotézy a teórie, ktoré vychádzajú z dát, neodrážajúcich dostatočne empirickú realitu, má za následok, že ostatní nedokážu zopakovať merania s rovnakými závermi. Hlavná príčina sa nachádza vo využívaní postupov, ktoré porušujú pravidlá, za akých *NHST* prístup poskytuje spoľahlivé odhady.

Prvým postupom je tzv. *šuflíkový efekt* (*„file-drawer effect"*), spájaný s nadmerným uverejňovaním štatisticky významných výsledkov. Môže vzniknúť nielen v dôsledku editorských kritérií časopisov, ktoré zvyknú zamietať triviálne alebo štatisticky nevýznamné poznatky, ale aj rozhodnutím samotných autorov/riek. Ak sa – v rozpore s predpokladmi - dospelo k negatívnym záverom, text sa odloží do „šuflíka". *Šuflíkový efekt* blokuje prístup k textom, neodporujúcim nulovej hypotéze. O kvalitných štúdiách s triviálnymi, ale dôležitými zisteniami, sa nedozvieme (Earp 2017). Vzniká selektívne vykazovanie, nárast zverejňovania štúdií, napĺňajúcich pôvodné predpoklady.

V priebehu rokov 1990 – 2007 bolo analyzovaných 4600 publikácií (Fanelli 2012). Autor uvádza, že podpora nenulových hypotéz sa zvýšila o 22 %, a to najmä v sociálnych a biomedicínskych disciplínach. V porovnaní s USA, kde je prírastok menší, sa to výraznejšie prejavuje v ázijských (Japonsko) a európskych krajinách (Veľká Británia). Na záver konštatuje: *„Nie je jasné, či za uvedeným trendom stojí priekopnícky prieskum alebo pokles objektivity zverejňovania výsledkov"*. Franco et al. (2014) upozorňujú, že štatisticky významné závery majú o 40 % väčšiu pravdepodobnosť publikovania, ako negatívne zistenia. Lakens (2017a) udáva, že až 90 % textov v karentovaných psychologických časopisoch podporuje stanovené predpoklady. Následne si kladie otázku: *„Načo zbierať údaje a testovať hypotézy, keď sa skoro vždy potvrdia? Alebo je za tým genialita ľudí, ktorí výskumné hypotézy stanovujú?"*. Podobne Gerber, Malhorta, (2008) konštatujú, že niektoré výsledky popredných časopisov v sociológii môžu byť zavádzajúce a nepresné. Riziko opisu iluzórneho plátna, omylom zameneného za celý obraz alebo jeho výsek, je veľké.

Publikačný tlak *(„publish or perish"*), či potreba získania grantov môžu viesť k pokušeniu modifikovať zozbierané údaje s cieľom rýchleho zverejnenia zásadných zistení (Nosek et al. 2018; Field 2018). Dvoma technikami, používanými na zvýšenie šancí zverejnenia textov sú *„p-hacking"* a *„HARKing"*.

„P-hacking" (selektívne vykazovanie zozbieraných dát), ignoruje, vylúči alebo neuvedie poznatky, spôsobujúce odchýlku od pôvodných predpokladov. Pozornosť sa zameria iba na to najsľubnejšie s cieľom vyladiť požadované zistenia (Cumming – Jageman 2017)[2]. Na rozdiel od *„šuflíkového efektu"*, sa

[2] Postupov *„p-hackingu"* môže byť viac: ad hoc spájanie viacerých súborov, zväčšovanie vzorky v rozpore s pôvodnými predpokladmi, dodatočné rozhodnutie o použití kovariátu až po získaní dát, rozhodovanie, ktoré údaje do analýzy zahrnúť, ich začlenenie/vylúčenie podľa vplyvu na *p*-hodnoty, iná transformácia alebo manipulácia s cieľom dosiahnuť štatisticky významné výsledky pri nízkej sile účinnosti testu (Simonsohn et al. 2014b; Field 2018).

selekcia netýka vykazovaných článkov, ale výberu výsledkov v rámci jednej štúdie (Simonsohn et al. 2014a). Problémom je, že výberové rozhodnutia o zverejnení iba toho sľubného (*„cherry picking results“*) sa urobí až po prečítaní dát v exploračnej analýze. Štatistici majú na to príslovie: *„Ak údaje dostatočne dlho vypočúvate, priznajú sa.“* (Cumming – Jageman 2017). Pri veľkom počte premenných, *p*-hacking zvyšuje pravdepodobnosť nájdenia akýchkoľvek štatisticky významných vzťahov (tzv. *„crud factor“ – kedy sa do analýzy zahrnú rôzne údaje, pôvodne mimo predmetu spracovania*), ktoré sú v skutočnosti chybou prvého druhu (α).

Ďalší postup sa nazýva *„HARKing“*. Jedná sa o dodatočnú zmenu hypotéz až po zistení, *„čo funguje“*. Hypotéza sa vyvodzuje a dokazuje na rovnakých údajoch a následne sa aj tvrdí, že ide o predpoklad, ktorého podpora sa v dátach hľadala. Je to rovnaké, ako si staviť na víťazného koňa až po dobehnutí do cieľa, alebo posunúť terč po hode šípkou tak, aby zasiahla stred.

„P-hacking“ aj „HARKing“ môžu byť použité jednotlivo alebo sa dopĺňať. Pre ilustráciu: v prevalencii násilia páchaného na deťoch sa ako východisko zvykne používať ekologický model (Belsky 1980) so štyrmi základnými úrovňami – individuálnou, rodinnou, komunitnou a spoločenskou. V každej z nich sa nachádza niekoľko ukazovateľov, ktoré v rôznej miere s násilím asociujú (European report on preventing child maltreatment, WHO 2013, Scannapieco et al. 2005). Ak sa po zozbieraní dát – v rozpore zo zvolenou hypotézou, zameranou na konkrétnu úroveň i ukazovateľ (napr. na rodinnú úroveň a rodinný stav) – ukážu sľubné výsledky pre odlišnú úroveň i ukazovateľ (napr. pre individuálnu úroveň a vek rodiča), môžu byť obidve techniky na úpravu výsledkov použité komplementárne. Aplikovaním *„HARKingu“* preformulujeme pôvodnú hypotézu a nesprávne uvedieme, že sa od začiatku hľadala dôkazná podpora pre individuálnu úroveň. Namiesto pravdivého reportovania, dodatočne zmeníme hypotézy tak, aby zapadli do získaných výsledkov (Weber – Popova, 2012). *„P-hackingom“* publikujeme ukazovateľ s $p < 0{,}05$, zameraný na vek rodiča, ktorý sa ukázal ako rozhodujúci determinant. Čitateľ sa nedozvie o zvyšných ukazovateľoch, ktoré boli z analýzy vylúčené. Rovnako nevie, že nájdená, štatisticky významná súvislosť nemusela byť medzi stanovenými hypotézami.

Odpoveď na otázku, prečo uvedené postupy spôsobujú publikačné skreslenia, sa nachádza v pravidlách, za akých testovanie nulových hypotéz *NHST* poskytuje správne inferenčné odhady. *NHST* prístup používame vtedy, ak nemáme prístup k celej populácii. Pomocou predpokladu nulovej a alternatívnej hypotézy (*H0*/*H1*) sa štatistickou inferenciou snažíme získať viac poznatkov o populačných parametroch (tabuľka č. 1). Na vzorku sa aplikuje vhodný test

a na zvolenej hladine významnosti sa podľa získanej *p*-hodnoty[3] rozhodne o ne/zamietnutí *H0* (Cumming 2008, 2012; Cumming – Jageman 2017; Soukup – Rabušic 2007; Soukup 2010, 2019; Rabušic et al. 2019). Daňou za testovanie je riziko chybovosti. Ak platí *H0*, potom má riziko podobu *falošne pozitívneho výsledku* (*FPV*), čo je chyba prvého druhu (α). Manipulácia s veľkosťou hranice (α) má vplyv na zastúpenie *falošne pozitívnych* (*FPV*) a *pravdivo negatívnych výsledkov* (*PNV*). Ak naopak platí *H1*, potom je rizikom *falošne negatívny výsledok* (*FNV*), t.j. chyba druhého druhu (β). Manipulácia s hranicou (β) ovplyvňuje pravdepodobnosť získania *falošne negatívnych* (*FNV*) *a pravdivo pozitívnych výsledkov* (*PPV*). Cieľom testovania je, pri rešpektovaní vopred stanovenej miery chybovosti (α, β), správne rozhodnúť o zamietnutí alebo nezamietnutí *H0*.

Tabuľka č. 1: **Pravidlá testovania nulových hypotéz** (NHST)

	pravdivá *H0*	pravdivá *H1*
nezamietnutie *H0*	Správne rozhodnutie o nezamietnutí *H0*. Pravdivý negatívny výsledok (*PNV*). $P = 1 - \alpha$	Mylné rozhodnutie o nezamietnutí *H0*. Falošne negatívny výsledok (*FNV*). *chyba druhého druhu* $P = \beta$
zamietnutie *H0*	Mylné rozhodnutie o preferencii *H1*. Falošne pozitívny výsledok (*FPV*). *chyba prvého druhu* $P = \alpha$	Správne rozhodnutie o preferencii *H1*. Pravdivo pozitívny výsledok (*PPV*). *sila testu* $P = 1 - \beta$

Poznámka: *H0* – nulová hypotéza (predpoklad o absencii vzťahu/súvislosti, rozdiely sa rovnajú nule), *H1* – alternatívna hypotéza (predpoklad o vzťahu/súvislosti, rozdiely sa nerovnajú nule), α – chyba prvého druhu, β – chyba druhého druhu, *P* – pravdepodobnosť dosiahnutia výsledku pre danú variantu (viď aj Soukup 2010; Rabušic et al. 2019).

Na aký kompromis chybovosti pristúpiť nemusí byť vždy zrejmé. V praxi je akceptovateľnejšia konfrontácia s chybou prvého druhu (α), pretože sa ľahšie identifikuje pri replikáciách, slúžiacich na jej kontrolu. Ak je slabá sila testu a v skutočnosti platí H1, konfrontácia s chybou druhého druhu (β) môže byť závažnejšia. Mylne sa potom domnievame, že asociácia nie je prítomná, hoci skutočnosť je iná. Pravdepodobnosť nepreskúmania nových hypotéz sa zvyšuje. Aby neprišlo k mylnému zisteniu o rozdiele/asociácií, dôležitou je vopred nastavená hladina významnosti (štandardne, $\alpha = 5$ %, často ale aj menšia, napr. $\alpha = 1$ %, $\alpha = 0{,}1$ %), ktorá hovorí o tom, ako často sme ochotní pripustiť, že sa mýlime, t.j. že z dlhodobého hľadiska (in the long run) získame falošne pozitívny výsledok (Field 2018). Hladina významnosti je pravdepodobnosť nájdenia štatisticky významného rozdielu, aj keď žiadny neexistuje pre sériu opako-

[3] V štatistickom jazyku sú *p*-hodnoty výsledkom porovnania vypočítanej testovej štatistiky s jej kvantilovým teoretickým rozdelením na zvolenej hladine významnosti (α) pri daných stupňoch voľnosti. Ak získanej hodnote testovej štatistiky prislúcha *p*-hodnota nižšia, ako je štandardne zvolená hladina významnosti (α), potom existuje max 5% pravdepodobnosť získania rovnakého alebo horšieho (väčšieho) rozdielu, za predpokladu platnosti nulovej hypotézy (*H0*) (Soukup 2010).

vaných výberov (nie pre konkrétnu štúdiu, v ktorej sa *H0* buď zamietne alebo nezamietne) (Greenland et al. 2016). *„P-hacking“*, *„HARKing“*, ako aj *„šuflíkový efekt“* tento odhad rušia dodatočnou zmenou pevných predpokladov. Aplikáciou *p-hackingu* narastá chybovosť v dôsledku voľby publikovania krajších výsledkov. Ak po exploračnej analýze vyberieme iba to, čo sa hodí a ostatné zamlčíme, môže ísť o náhodu ďaleko častejšie, ako si myslíme (Meehl 1990; Simmons et al. 2011). Použitie *HARK*ingu zvýši chybovosť formulovaním predpokladov až po analýze dát. Narastá riziko nesprávneho zamietnutia nulovej hypotézy. *Šuflíkový efekt*, v dôsledku neprítomnosti textov s triviálnymi výsledkami, zvýši pravdepodobnosť falošne pozitívnych zistení, čo vytvára skreslený obraz (nadmernej) veľkosti efektov (Kerr 1998). Všetky uvedené postupy zvyšujú vopred nastavenú chybu prvého druhu (najčastejšie α = 5 %) na neznámu úroveň a závery z *p*-hodnôt, zodpovedajúcich testovacej štatistike pre dané rozdelenie, prestávajú platiť. Pri veľkom množstve replikácií a za predpokladu platnosti *H0*, je potom očakávanie maximálne jedného chybného výsledku z každých dvadsiatich mylné. Zvýši sa nad stanovených 5 %.

Interpretácia *NHST* testovania trpí aj ďalšími obmedzeniami, týkajúcimi sa správneho používania *p*-hodnôt.[4] *P*-hodnoty sú iba prvotným nástrojom, ako na vzorke rozhodnúť, či sa v „náhodnom hluku“ nachádza „signál“. V ojedinelej štúdii majú iba malú výpoveď – ich zmysel spočíva v dlhodobej perspektíve opakovaní. V tejto súvislosti vydala príručku aj Americká štatistická asociácia (ASA, 2016), kde upozorňuje, že dobre odôvodnené štatistické argumenty obsahujú oveľa viac, ako osamelé číslo presahujúce arbitrárny, vopred stanovený prah chybovosti (0,05). *P*-hodnoty by nemali nahrádzať hlbšie úvahy a samotné na správne závery nestačia (ASA 2016). Podobne argumentuje Greenland: *„keď sa jediným výkladom stane oddeľovanie výsledkov na štatisticky významné/nevýznamné, môže to spôsobiť, že dôležitá práca nebude publikovaná iba kvôli vysokej p-hodnote* (Greenland, et al. 2016).

P-hodnoty disponujú aj slabou spoľahlivosťou: pri opakovanom náhodnom výbere rovnakej veľkosti a za rovnakých podmienok neexistuje žiadna záruka, že hľadaný výsledok bude opäť štatisticky významný. Je tomu tak preto, že ide o náhodnú veličinu (s vopred známym rozdelením pri platnosti *H0*), ktorej ďalší údaj nemožno v skutočnosti predpovedať – isté je len to, že bude z intervalu <0,1>. Ak chceme disponovať presnejšími závermi, tak je vhodnejšie opustiť čierno biele videnie „všetko alebo nič“ a využiť intervaly spoľahlivosti (Cumming 2008; Cumming – Jageman 2017; Field 2018; Soukup 2010, 2019). Tento postup upúšťa od mechanickej aplikácie hladiny významnosti

[4] Komplexný prehľad správneho používania *p*-hodnôt, vrátane 25 najčastejších nedorozumení, možno nájsť v texte (Greenland et al. 2016) alebo American Statistical Association ASA (2016). Pripomienky k *p*-hodnotám, aj ako reakcia na zákaz ich používania niektorými karentovanými zahraničnými časopismi, boli publikované aj v československom sociálno-vednom prostredí (Soukup 2010, 2019; Kanovský 2016, Ropovík 2018).

0,05), kedy p = 0,051 neodporuje *H0*, ale p = 0,049 už áno, hoci ide o takmer totožné zistenia.

Na veľkosť p-hodnoty má tiež vplyv dodržanie predpokladov testovania pre použitie konkrétnych testov, dizajn a veľkosť vzorky – čím je väčšia, tým sa ľahšie dosiahne signifikantnosť (Rabušic et al. 2019; Field – Hole 2003; De Vaus, 2002, 2014). Ak je použitá celá populácia alebo výber nie je náhodný/reprezentatívny/randomizovaný, údaj stráca zmysel. Nižšia hodnota (v celkovom intervale <0,1> neznamená kvalitnejší výsledok (môže ísť o vecne zanedbateľný rozdiel), iba vyššiu pravdepodobnosť nekompatibility s predpokladom *H0*. Najmä pri väčšom objeme dát je H0 ľahké zamietnuť. Niektorí autori preto konštatujú „ničotnosť nulových hypotéz" a ich slabú výpovednú silu v prospech hypotézy/teórie (Soukup 2010). A nakoniec: $p > 0,05$ nehovorí o pravdivosti nulovej hypotézy $P(H0/D)$, ale o nedostatku evidencie v dátach na jej zamietnutie $P(D/H0)$. $P < 0,05$ nevypovedá o pravdivosti *H1*, iba o evidencii dát v rozpore s *H0*. Jediné, čo frekvenčná štatistika dokáže, je preukázať evidenciu v neprospech *H0*, neodporujúcu *H1* (Kanovský 2016)[5].

Tabuľka č. 2: **Vplyv nastavenia parametrov testu a percenta apriori platných *H0*/*H1* na výsledky *NHST* testovania**

	ALTERNATÍVA 1 % apriori platných *H0*/*H1* 50/50			ALTERNATÍVA 2 % apriori platných *H0*/*H1* 50/50			ALTERNATÍVA 3 % apriori platných *H0*/*H1* 10/90		
	nastavenie testu	pravdivá *H0*	pravdivá *H1*	nastavenie testu	pravdivá *H0*	pravdivá *H1*	nastavenie testu	pravdivá *H0*	pravdivá *H1*
nsg. výsledok nezamietnutie *H0*	1- α = 99% β = 20%	*PNV* 49,5%	*FNV* 10%	1- α = 99% β = 1%	*PNV* 49,5%	*FNV* 0,5%	1- α = 95% β = 20%	*PNV* 9,5%	*FNV* 18%
sig. výsledok zamietnutie *H0*	α = 1% 1- β = 80%	*FPV* 0,5%	*PPV* 40%	α = 1% 1- β = 99%	*FPV* 0,5%	*PPV* 49,5%	α = 5% 1- β = 80%	*FPV* 0,5%	*PPV* 72%

Poznámka: *FPV* – falošne pozitívny výsledok, *PNV* – pravdivo negatívny výsledok, *PPV* – pravdivo pozitívny výsledok, *FNV* – falošne negatívny výsledok, α – chyba prvého druhu, β – chyba druhého druhu, (1 – α) – pravdivo negatívny výsledok, (1 – β) – pravdivo pozitívny výsledok (sila testu), výraznejšie šedá: zmena parametrov, šedá: vplyv zmien na pravdivo pozitívne výsledky (*PPV*), výpočet online: http://shinyapps.org/apps/PPV/, Lakens – Evers, 2014.

Za najlepšiu prevenciu publikačných skreslení je považovaná *preregistrácia štúdií*. Jedná sa o časový prehľad realizovaných krokov, vrátane prístupu k dátovému súboru. Online protokol a peer-review proces robia výskumný

[5] Ak napr. výsledku testovej štatistiky – pre dané rozdelenie - zodpovedá štatisticky významná hodnota p = 0,037, potom existuje iba 3,7 % pravdepodobnosť získania údajov (respektíve rozdielov rovnakých alebo väčších), ak v skutočnosti platí H0 P(D/H0). Ak je naopak údaj vyšší: p = 0,40, potom sú dáta viac v zhode/kompaktné/zlučiteľné s použitým modelom (rozdelením za predpokladu H0).

zámer a zber dát transparentným.[6] Štúdie, ktoré prešli preregistráciou a u ktorých sa nezíska podpora v prospech *H0*, sú najdôveryhodnejšími zdrojmi dôkaznej hodnoty (Lakens, Etz, 2017d). Ako ukazuje tabuľka č. 2 (alternatíva č. 3, % apriori platných *H0*/*H1* = 10/90 v 100 štúdiách), ak chceme preukázať rozdiely/efekt, tak najlepšou voľbou je nadviazať na preregistračné výsledky, v ktorých sa opakovane ukazuje preferencia *H1*.

V porovnaní s neznámym prostredím, kde jediné, čo vieme, je platnosť náhody (% apriori platných *H0*/*H1* = 50/50 v 100 štúdiách, alternatíva č. 1, a č. 2), preregistračné výsledky najviac zvyšujú pravdepodobnosť získania pravdivo pozitívnych výsledkov (*PPV*) (alternatíva č.3). Pre detailnejšiu ilustráciu a vysvetlenie k tabuľke: ak pred výskumným zámerom nemáme vopred vedomosti o predchádzajúcich textoch (tzn. % apriori platných *H0*/*H1* = 50/50) a v *NHST* testovaní je nastavená chyba prvého druhu na: α = 1 % a sila testu sa nachádza na štandardnej úrovni 80 %, potom získame najčastejšie pravdivo negatívny záver *PNV* = 49,5 % (alternatíva č. 1). Ak zvýšime silu testu na 99 %, pričom α a % apriori platných hypotéz zostanú nezmenené, potom pravdepodobnosť pravdivo negatívneho záveru sa vyrovná pravdivo pozitívnym: *PPV* = *PNV* = 49,5 %. *FNV* sa zníži z 10 % na 0,5 % (alternatíva č.2). Najväčšie zvýšenie *PPV* ilustruje alternatíva č. 3 – so štandardnou mieru chybovosti, kde je sila testu 80%, α = 5 %, ale s vyšším pomerom apriori pravdivých (a preregistrovaných) hypotéz v prospech *H1*: *H0*/*H1* = 10/90. Získanie *PPV* pri takomto nastavení sa zvýši až na 72 %. Znamená to, že ak plánujeme štúdiu alebo replikáciu, u ktorých chceme preukázať vzťah, dôležité je mať prístup k spoľahlivým, pravdivo pozitívnym výsledkom (v alternatíve č. 3 je to v predchádzajúcich štúdiách vysoké % apriori platných hypotéz *H1*). Jedná sa o oveľa lepšiu možnosť, ako prísnejšie nastavenie chybovosti testu manipuláciou s prednastavenými mierami (α, β). Analýza *p* kriviek – detailnejšie priblížená v nasledujúcej časti ukazuje, ako zistiť aby závery nepochádzali zo zverejnených štúdií s nedostatkom dôkaznej hodnoty.

P-krivky – ako v publikáciách odhaliť selektívne vykazovanie

Pomerne novou technikou, odlišujúcou dôkaznú hodnotu od falošne pozitívnych zistení, sú *p*-krivky (Simonsohn et al. 2014a, 2015b, 2016). Jedná sa o užitočnú, na vyhodnotenie nenáročnú pomôcku s vysokou presnosťou. Do analýzy môže byť zahrnutý ľubovoľný počet štúdií. Výsledky neskresľujú chýbajúce hodnoty $p > 0,05$. Použitie je možné v observačnom aj experimentálnom dizajne. *P*-krivky posúdia nezrovnalosti a vedú k lepšiemu výberu relevantnej literatúry (Simonsohn et al. 2016). Opodstatnenosť ich využitia je všade tam,

[6] Postupy pre preregistráciu výskumného zámeru dostupné na: https://cos.io/, https://osf.io/prereg/, https://aspredicted.org/

kde chceme preskúmať publikačné skreslenia a podozrenia z dodatočnej úpravy údajov (Lakens- Evers 2014).

Použitie *p*-kriviek má miesto v začiatočnej fáze výskumu, keď sa rozhoduje o výbere článkov, na ktoré sa bude nadväzovať. Podľa tvaru a zošikmenia *p* - hodnôt zahrnutých v rešerši, sa dostatočne včas dozvieme, na ktorých zisteniach možno stavať. Kandidátmi na preskúmanie sú štúdie s malými vzorkami, prekvapujúcimi závermi, či slabou silou testu, ktorý má za následok nízku pravdepodobnosť zachytenia rozdielov/efektu – ak sú tieto reálne prítomné. Závery *p*-kriviek umožňujú predchádzať reprodukcii nespoľahlivých výsledkov a šetria čas, investovaný do nesprávneho výskumného predmetu. Ak ich analýza nepodporuje evidenciu v prospech určitej hypotézy, mali by sme si rozmyslieť, či chceme na publikácie nadväzovať. Hrozí, že texty nie sú postavené na reálnych základoch – zistenia sú ťažko reprodukovateľné. Môžeme čeliť záverom z falošných hypotéz a literatúra obsiahnutá v rešerši nemusí byť najlepšou inšpiráciou (Lakens – Evers 2014).

P-krivky sú distribúciou štatisticky významných hodnôt $p < 0,05$, získaných z výsledkov testovania sady publikovaných štúdií. Otázka, ktorú si v analýze kladieme, sa týka porovnania ich tvarov: *„Je nameraný sklon štatisticky významných p-hodnôt – v porovnaní s dvoma očakávanými distribúciami, dostatočne zošikmený (vpravo) na vylúčenie selektívneho reportovania, alebo je tento sklon pre podporu dôkaznej hodnoty nepostačujúci?“*. Hlavný výstup tvoria tri krivky (graf č. 1 a č. 2).

Prvá (bodkovaná) krivka vyjadruje očakávanú distribúciu pri platnosti *H0* a na *x*-ovej osi štatisticky významných *p* hodnôt z intervalu $0 < p < 0,05$ má uniformný tvar. Ak platí *H0* (rozdiel/efekt neexistuje), tak v dlhodobej perspektíve opakovaní je rovnaká šanca získania celého spektra *p* hodnôt v uvedenom intervale. Vznikne priamka rovnobežná s osou *x*. Druhá (čiarkovaná) *p*-krivka je distribúciou *p* hodnôt pri 33 % sile testu a pravdivosti *H1*. Ak platí *H1* (rozdiel/efekt existuje), tak je viac štatisticky významne menších *p* hodnôt z intervalu $0,01 \leq p \leq 0,04$, ako väčších hodnôt z intervalu $0,04 < p < 0,05$[7] (Cumming 2008; Simonsohn et al. 2014a). Na *x*-ovej osi je potom distribúcia mierne zošikmená doprava. 33 % sila testu zároveň spôsobuje, že tvar krivky zostáva približne rovnaký, bez ohľadu na veľkosť vzorky resp. koeficient sily účinku (s. 668, simulačné grafy, Simonsohn et al. 2014b). Kľúč k vylúčeniu selektívneho vykazovania spočíva vo vzájomných porovnaniach s treťou, nameranou (plnou) *p*-krivkou, ktorá je výsledkom zahrnutých štúdií.

[7] Ak efekt reálne existuje, potom *p*-krivky sú vždy zošikmené doprava, bez ohľadu na veľkosť vzorky a silu testu. Napríklad, ak je Cohenovo $d = 0,42$ a $N = 20$, tak pre 38% *p*-hodnôt platí $p < 0,01$, ak je $d = 0,91$ a $N = 20$, potom až 71% *p*-hodnôt je menších ako 0,01, čo znamená výrazné pravé zošikmenie. Čím väčší efekt, tým väčšie zošikmenie doprava, ktoré narastá aj s vyššou silou testu (pravdepodobnosť odhalenia rozdielu, ak existuje). Známa veľkosť vzorky a známe percento *p*-hodnôt, spadajúcich pod určitú hranicu, umožňujú taktiež odhadnúť priemerný koeficient sily účinku. (Simulácie s. 537, grafy 1A – 1D, Simonsohn et al. 2014a alebo s. 668, obrázok č. 1, Simonsohn et al. 2014b)

Ak je nameraná *p*-krivka dostatočne zošikmená doprava, selektívne vykazovanie textov je vylúčené, a na súbor štúdií možno nadviazať.[8] Dôkazná hodnota sa ale nemusí nachádzať vo všetkých štúdiách (aj pri porovnávaní rozdielov, napr. medzi dvoma nezávislými skupinami sa niektoré zistenia v skupinách nemusia líšiť, hoci v priemere sa jedná o štatisticky významne rozdiely).

Ak je nameraná distribúcia menej strmá, ako *p*-krivka s očakávaným predpokladom platnosti *H1* pri 33 % sile testu, potom sa konštatuje nerozhodný výsledok. Hodnota dôkazu buď neexistuje, nie je dostatočná alebo je príliš malá na to, aby sme sa ňou mohli zaoberať a urobiť definitívne rozhodnutie. Je tiež možné, že je nutné zhromaždiť väčší počet textov a zahrnúť do analýzy väčší počet *p* hodnôt (Simonsohn et al. 2014; Simmonsohn – Nelson 2015b).

O *„p-hackingu"* svedčí prípad, kedy sa väčšina nameraných *p* hodnôt ukáže v hornej hranici x-ovej osi (0,04-0,05). Ak efekt/rozdiel reálne existuje, tak ľavotočivá distribúcia (tesne pod hranicou 0,05) je málo pravdepodobná. Rozloženie svedčí o prítomnosti selektívneho vykazovania (napr. pri malých vzorkách, v dôsledku jej postupného zväčšovania, až kým sa rozdiely nestanú štatisticky významné (obrázok 1E, s. 537 Simonsohn et al. 2014a). Ľavotočivý tvar sa očakáva pri platnosti *H0* a nie *H1*. Ak sa rozhodneme na zahrnuté štúdie nadviazať, hrozí reprodukcia nespoľahlivých zistení (Lakens – Evers 2014; viď tiež graf Masicampo – Lalande 2012; Field 2018, grafická ilustrácia rozloženia *p*-hodnôt pri *„p-hackingu"*).

Na konečné rozhodnutie o dôkaznej hodnote ale nestačí iba vizuálne odčítanie. Zošikmenie nameranej distribúcie je nutné štatisticky otestovať. V online aplikácii (www.p-curve.com), do ktorej sa zadávajú výsledky[9], a ktorá je hlavným vstupom analýzy *p*-kriviek, sa na to používajú dva testy. Prvým je *binomický test*, ktorý porovnáva zastúpenie nameraných štatisticky významných *p* hodnôt z intervalov $0 < p < 0{,}025$ a $0{,}026 < p < 0{,}05$ s dvoma očakávanými distribúciami – za predpokladu platnosti *H0* a za predpokladu platnosti *H1* pri 33 % sile testu (tabuľka č. 4, prvý stĺpec). Nevýhodou *binomického testu* je, že ignoruje variáciu *p* hodnôt v rámci uvedenej dichotomizácie a nemusí byť dostatočne účinný (Simonsohn et al. 2015b).

Vo výstupe online aplikácie je preto uvedený aj *spojitý test*, ktorý využíva robustnú *Stoufferovu metódu*[10], doplňujúcu *Fisherov prístup binomického testu* (tabuľka č. 4, druhý a tretí stĺpec). Ukázalo sa, že so zvyšujúcou sa intenzitou *p-hackingu* (najmä pod hranicu $p < 0{,}025$) sa zvyšuje pravdepodobnosť získa-

[8] Ak efekt reálne existuje, bez ohľadu na jeho veľkosť a veľkosť vzorky, očakávané distribúcie *p*-hodnôt sú vždy stočené vpravo (tzn. k nižším *p*-hodnotám). Korelácia koeficientu sily účinku s veľkosťou vzorky tu nehrá žiadnu rolu (simulácie a grafy s. 538, Simonsohn et al. 2014a).

[9] Ide o konkrétne hodnoty štatistických testov jednotlivých štúdií, napr. F (1,100) = 9,1, chi2 (2) = 8,74 a pod. Aplikácia zodpovedajúcu *p*-hodnotu prepočíta, tzn. nezadávajú sa *p*-hodnoty priamo.

[10] Detaily výpočtu viď. sprievodná tabuľka, ktorá je automaticky publikovaná online s výsledkami.

nia štatisticky významnej falošnej *p*-krivky stočenej vpravo, aj keď žiadny efekt v skutočnosti neexistuje. Táto pravdepodobnosť narástla i vtedy, ak sa v simuláciách do množiny štúdií s nulovou evidenčnou hodnotou pridal malý počet falošne pozitívnych zistení (simulačné grafy, s. 1149-1150, obr. č. 2: Simonsohn et al. 2015b). *Stoufferov prístup* znižuje riziko mylného stočenia vpravo a získanie mylnej dôkaznej hodnoty zmenší na prijateľnú úroveň. Z tohto dôvodu sú namerané *p*-krivky rozdelené na plnú ($p < 0,5$) a polovičnú *p*-krivku ($p < 0,025$). Druhá z nich umožňuje lepšie odhaliť závažnejší *p*-hacking - pri vyhodnotení obidvoch distribúcií pomocou Z-skóre (tabuľka č. 4) zabezpečí lepší odhad dôkaznej hodnoty, ako výlučné použitie úplnej *p*-krivky, zameranej na celé spektrum rozloženia získaných $p < 0,05$. Ak efekt existuje, (úplná aj polovičná p-krivka) poskytujú rovnako dobré výsledky i pri nízkom počte zahrnutých *p*-hodnôt. Ak ale efekt neexistuje a *p*-hacking je intenzívny (napr. $p < 0,025$), potom je polovičná *p*-krivka ďaleko presnejšia (s. 1150, obr. č. 2, grafy 1-6 Simonsohn et al. 2015b). Vďaka svojej robustnosti (na rozdiel od plnej netestuje *p*-hodnoty > 0,025) má však nižšiu štatistickú silu, t.j. vyššiu pravdepodobnosť neodhaliť dôkaznú hodnotu tam, kde sa v skutočnosti nachádza. Kombinácia vyhodnotenia *spojitým testom* so *Stouffer prístupom* a plnou krivkou jej nižšiu štatistickú silu koriguje (Simonsohn et al. 2015).

P-krivky sú v identifikácii *publikačných skreslení* vysoko presné. Aj s malým počtom zahrnutých *p* hodnôt a nízkou silou testu je možné získať výsledok, ktorý správne vyhodnotí absenciu dôkaznej hodnoty alebo naopak: jej podporu. Falošné negatívne a falošne pozitívne závery sú málo pravdepodobné[11] (s. 544, obr. č. 6. výsledné simulačné grafy, Simonsohn et al. 2014a).

Nie všetky *p*-hodnoty sú však vhodné pre zahrnutie do analýzy. Nesprávna voľba je najväčšou hrozbou pre validitu výsledkov. Dôležité je dodržať transparentné kroky, na ktoré by ostatní mohli ľahko nadviazať (detaily: Online Supplement For Better *P*-curves, Official User-Guide to *P*-curve, online na: www.p-curve.com). Prvým krokom je nasledovať vopred zdôvodnené a zverejnené pravidlo, zabraňujúce subjektívnemu výberu textov (*napr. päť najcitovanejších článkov v odbornom časopise v priebehu posledných piatich rokov)*. Druhým krokom je uvedenie sprievodnej tabuľky s detailmi štúdií (tabuľka č. 3).

[11] Napr. 20 štatisticky významných *p*-hodnôt (so skutočnou dôkaznou hodnotou) už pri 33 % sile testu dosahuje 85 % pravdepodobnosť získania správnej, štatisticky významnej pravotočivej *p*-krivky. Pravdepodobnosť získania falošne negatívnej *p*-krivky, menej strmej ako očakávaná distribúcia, je pri 33 % sile testu iba 5 % (alternatívy simulácií. s. 544, grafy 6A/6B/6C/6D, Simonsohn et al. 2014a).

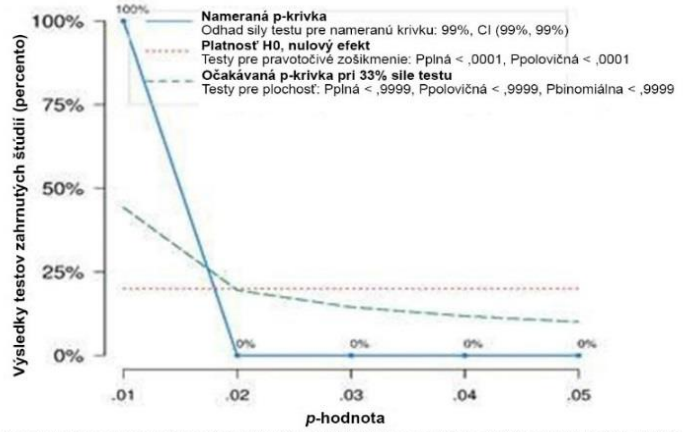
Tabuľka č. 3: **Sprievodné údaje zahrnutých štúdií**
(pre detaily viď tiež manuál ku *p*-krivkám, alebo Simonsohn et al. 2014a)

	autori/ky zdroj	zámer/hypotéza	dizajn	kľúčové zistenia	citácia z výsledkov v texte	štatistický test pre p-krivku
1.	Duchesne, G., A., Hébert, M., Daspe, M., E., (2017)	skúmanie rozdielnych dôsledkov sexuálneho násilia u chlapcov a dievčat	447 sexuálne zneužívaných detí (319 dievčat, 128 chlapcov)	rozdiely medzi percentami chlapcov a dievčat vystavených sexuálnemu zneužívaniu v miere po-straumatickej stresovej poruchy	...dievčatá so skúsenosťou sex. násilia dosahovali väčšie skóre v postraumatickej stresovej poruche (PTSD) $t(425) = 3,23$; $p = 0,001$ ...chlapci so skúsenosťou sexuálneho násilia čelia závažnejším „intrusive acts“ $\chi2\ (2) = 12,64$; $p = 0,002$	$t(425) = 3,23$ $p = 0,001$ $\chi2\ (2) = 12,64$ $p = 0,002$
2.	Negriff, S., Schneider-man, J.U., Smith, C., Schreyer, K., J., Trickett, K., P., (2014)	porovnanie demografic-kých kategórií detí so skúse-nosťou zlého zaobchádzania	longitudinál-ny dizajn, dáta od súdu pre mladis-tvých – posielanie listov – 303 opatrovateľov dalo súhlas	rozdiely v percentách sexuálneho násilia medzi chlapcami a dievčatami	...väčšina z mládeže so skúsenosťou sexuálneho násilia sa stala obeťou mužského páchateľa (91,7%) dievčatá boli štatisticky významne menej často zneužívané ženami v porovnaní s chlapcami $\chi2\ (1) = 9,62$; $p = 0,02$	$\chi2\ (1) = 9,62$ $p = 0,02$
3.	Sobsey, D., Randall, W., & Parrila, R. K. (1997)	...dievčatá, v dvoch skupinác h – so zdravot-ným znevý-hodnením a bez neho budú častejšie vystavené týraniu v porovnaní s chlapcami...	reprezenta-tívna vzorka z hlásení o zneužívaní detí v USA (1249 spisov)	rozdiely v percentuálno m zastúpení chlapcov a dievčat v dvoch skupi-nách – s a bez zdravotného znevýhodnenia	...prevalencia sexuálneho zneužívania medzi chlap-cami a dievčatami (18%/82%) v skupine bez zdravotných znevýhodnení $\chi2\ (1) = 89,91$; $p < 0,001$ a so zdravotnými znevý-hodneniami (38%-62%), $\chi2\ (1) = 10,01$, $p < 0,05$	$\chi2\ (1) = 89,91$ $p < 0,001$ $\chi2\ (1) = 10,01$ $p < 0,05$
4.	Karkošková, S., Ropovík, I., (2018)	...existujú rozdiely v konfrontácii sexuálneho násilia medzi pohlaviami?	observačný, 2186 študen-tov stredných škôl s náhodným výberom klastrov	sig. rozdiely v percentuálno m zastúpení u chlapcov a dievčat v konfrontácii so sexuálnym násilím	...dievčatá čelili častejšie-mu vystaveniu sexuálnemu násiliu v porovnaní s chlapcami (kontaktnému aj nekontaktnému).. $\chi2\ (1) = 144,471$; $p = 0,01$	$\chi2\ (1) = 144.471$ $p = 0.01$
5.	IVPR, Fico, M., (2017)	...existujú rozdiely v konfrontácii sexuálneho násilia medzi pohlaviami?	observačný, reprezenta-tívna vzorka (pohlavie, región) 2856 detí 8. a 9 ročníkov, náhodný výber základ-ných škôl v SR	sig. rozdiely v percentuálno m zastúpení u chlapcov a dievčat v konfrontácii so sexuálnym násilím	..dievčatá čelili častejšie vystaveniu sexuálnemu násiliu**..** dievčatá: 29.8% (27,4 – 32,2) chlapci: 19.9% (17,7 - 22.2) $\chi2\ (1) = 37,049$; $p = 0,01$	$\chi2\ (1) = 37,049$ $p = 0,01$

Zahrnuté texty musia spĺňať tri podmienky: vzťah ku skúmanej hypotéze, uniformnú distribúciu *p* hodnôt za predpokladu *H0* a ich vzájomnú nezávislosť. Dodržanie prvej podmienky možno zistiť zo sprievodnej tabuľky. Druhá podmienka je implicitne prítomná. Z hľadiska praktického postupu je najdôležitejšia nezávislosť *p*-hodnôt. Tu je nutné nadviazať na inštruktážne postupy, podľa ktorých sa možno riadiť (napr. podľa výsledkov testov v interakciách (obrátenie/zmiernene vzťahu, sa vyberajú príslušné p hodnoty, bližšie viď: inštrukčná tabuľka č. 5, s. 541-542, Simonsohn et al. 2014a). Niekedy sa môže stať, že štúdie reportujú viac analýz pre rovnaký zámer, v rôznych podmienkach, použitím parametrických a neparametrických testov (napr. rozdiely v rozsahu násilia páchanom na deťoch v závislosti od typu prostredia – v rodine, v škole). V takom prípade nemožno v *p*-krivkách reportovať výsledky obidvoch analýz, pretože je porušená podmienka nezávislosti *p*-hodnôt.

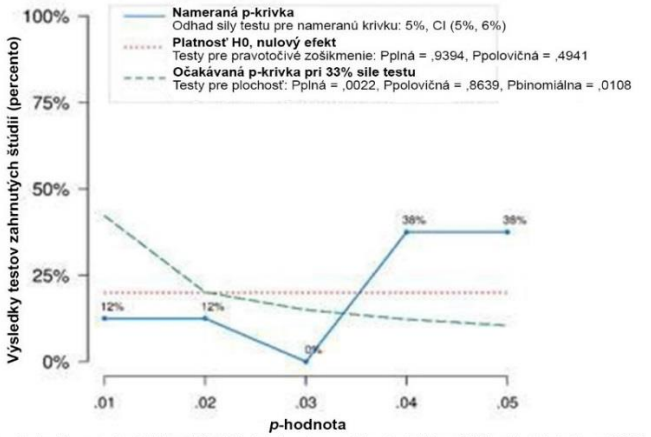
Využitie *p*-kriviek, v preskúmaní dôkaznej hodnoty vplyvu pohlavia detí na sexuálne zneužívanie (CSA), so zoznamom zahrnutých textov obsahuje tabuľka č.3. Analýza, obsahuje päť štúdií – tri zahraničné a dve zo Slovenska. Podľa inštrukcií, každý z textov sprevádzajú dodatočné charakteristiky: autor/ka, hypotéza, dizajn a hlavné zistenia štatistických testov. Vyhodnotenie obsahuje graf č. 1, ktorý poukazuje na pravotočivú distribúciu s dôkaznou hodnotou (nameraná *p*-krivka). Graf č. 2 sme prebrali z práce (Lakens 2017c) a slúži na ilustratívne porovnanie. Autor vyhodnocoval štúdie, venujúce sa vonkajším podnetom, ovplyvňujúcim konanie a myslenie (priming) u starších ľudí (viď tiež Kačmár 2019). Tu má nameraná *p*-krivka skôr ľavotočivý tvar s absenciou dôkaznej hodnoty a prítomnosťou selektívneho vykazovania.

Graf č. 1: ***p*-krivka: dôkazná hodnota** (pohlavie a CSA)



Poznámka: Nameraná p-krivka zahŕňa 7 štatisticky významných p hodnôt (p < 0.05), z ktorých je 7: p < 0.025.
Nesignifikantné výsledky neboli zahrnuté.

Graf č. 2: ***p*-krivka: selektívne vykazovanie** (priming u starších ľudí)



Poznámka: Nameraná p-krivka zahŕňa 7 štatisticky významných p hodnôt (p < 0.05), z ktorých je 7: p < 0.025.
10 výsledkov nebolo zahrnutých, lebo dosahovali p > 0.05.

Tabuľka č. 4: **Výsledky *p*-kriviek** (hlavný výstup)

	p-krivka 1 záver: dôkazná hodnota vplyv pohlavia detí na zlé zaobchádzanie (CSA – sexuálne násilie na deťoch)			*p*-krivka 2 záver: selektívne vykazovanie priming u starších ľudí (Lakens, 2017c)		
	BINOMICKÝ TEST *Binomický test*	SPOJITÝ TEST agregácia Stoufferovou metódou		BINOMICKÝ TEST	SPOJITÝ TEST agregácia Stoufferovou metódou	
	podiel výsledkov $p < 0,025$	plná *p*-krivka $p < 0,05$	polovičná *p*-krivka $p < 0,025$	podiel výsledkov $p < 0,025$	plná *p*-krivka $p < 0,05$	polovičná *p*-krivka $p < 0,025$
texty s dostatočne silnou dôkaznou hodnotou (nameraná *p*-krivka zošikmená doprava)	$p = 0,0078$	$Z = -10,71$ $p < 0,0001$	$Z = -10,1$ $p < 0,0001$	$p = 0,9648$	$Z = 1,55$ $p = 0,9394$	$Z = -0,01$ $p = 0,4941$
texty s nedostatočnou resp. žiadnou dôkaznou hodnotou (nameraná *p*-krivka plochejšia ako zodpovedá 33% sile testu)	$p > 0,999$	$Z = 9,18$ $p > 0,9999$	$Z = 9,95$ $p > 0,999$	$p = 0,0108$	$Z = -2,84$ $p = 0,0022$	$Z = 1,1$ $p = 0,8639$
sila testu pre *p*-krivku	Odhad: 99% 90% interval spoľahlivosti: (99%, 99%)			Odhad: 5% 90% interval spoľahlivosti: (5%, 6%)		

Poznámka: krátený prepis z online výstupu: www.p-curve.com (verzia 4.0, 2017) po zadaní výsledkov štatistických testov

Pri *textoch s dostatočne silnou dôkaznou hodnotou* (*p*-krivka dostatočne zošikmená doprava) nás zaujímajú dve alternatívy výsledkov. Pri prvej alternatíve musí byť ukazovateľ *Z* skóre záporný pre jednosmerný spojitý test aspoň u polovičnej *p*-krivky (t.j. odchýlka v prospech *H1*), kde $p < 0,05$. Druhá alternatíva spočíva v zápornom *Z* skóre súčasne pre polovičnú aj plnú *p*-krivku na 10 % hladine významnosti ($p < 0,1$). Ak je aspoň jedna alternatíva splnená, tak nameraná distribúcia vykazuje štatisticky významné pravé zošikmenie a umožňuje vylúčiť selektívne vykazovanie.

Výsledky textov, *vplyvu pohlavia detí na zlé zaobchádzanie (CSA*), poukazujú na záporné hodnoty *Z* skóre i štatistickú významnosť pre obidve *p*-krivky: polovičná: $Z = -10,1$, $p < 0,0001$; úplná: $Z = -10,71$, $p < 0,0001$. Binomický test má taktiež signifikantnú hodnotu $p < 0,0078$. Zahrnuté texty preto disponujú dôkaznou hodnotou. Pri štúdiách, zaoberajúcich sa *primingom starších ľudí* je v prvom odseku *Z* skóre nulové a *p*-hodnoty štatisticky nevýznamné (polovičná krivka: $Z = - 0,01$, $p = 0,4941$; úplná krivka: $Z = 1,55$, $p = 0,9394$). Binomický test je rovnako štatisticky nevýznamný ($p = 0,9468$), čo znamená, že selektívne vykazovanie nemožno vylúčiť.

Pre texty s *„nedostatočnou dôkaznou hodnotou“* je nameraná *p*-krivka testovaná na plochejší (menej strmý) pravotočivý sklon ako zodpovedá predpokladanej distribúcií pri 33 % sile testu. V analyzovaných textoch *zlého zaobchádzania s deťmi* je *Z* skóre pre polovičnú aj plnú *p*-krivku kladné a štatisticky nevýznamné (polovičná: $Z = 9,95$, $p = 0,999$, plná: $Z = 9,18$, $p = 0,999$). Rovnako to platí aj pre binomický test ($p = 0,999$). Nameraná distribúcia nie je štatisticky významne menej strmá, ako by bola distribúcia, zachycujúca efekt s pravdepodobnosťou 33 %. Selektívne vykazovanie preto možno vylúčiť.

Pri textoch *„primingu u starších ľudí s nedostatočnou dôkaznou hodnotou“* je binomický test štatisticky významný ($p = 0,0108$), podobne ako aj u zápornej *Z*-hodnoty *spojitého testu* pre plnú *p*-krivku ($Z = -2,84$, $p = 0,0022$). Takéto čísla poukazujú na nedostatok dôkaznej hodnoty v prospech priminigu a hrozbu selektívneho vykazovania.

K výstupu patrí aj interpretácia 90 % intervalu spoľahlivosti štatistickej sily nameranej *p*-krivky. Ide o údaj replikovateľnosti výsledkov v analyzovaných textoch. Ak by boli štúdie *zlého zaobchádzania s deťmi* realizované opäť, za rovnakých podmienok, potom najlepší odhad replikovateľnosti má hodnotu 99 %. Znamená to, že v priemere 99 % zahrnutých textov by bolo v replikácii úspešných. Pre *priming u starších ľudí* je sila testu *p*-krivky iba 5 %. Je preto iba málo pravdepodobné, že by úspešné replikácie boli v priemere zastúpené vo väčšom ako 6 % podiele, čo je horný interval 90 % intervalu spoľahlivosti v tab. č. 4 (Simmons – Simonsohn 2017).

Ak je účelom výskumného zámeru preukázanie súvislosti, potom v kombinácii s kladnými závermi *p*-kriviek získame silnú preferenciu v prospech

potenciálneho vzťahu. Cieľom testovania ale nemusí byť iba preukazovanie štatisticky významných rozdielov/efektu. V mene objektivity sa možno pýtať aj opačne: *„Možno zamietnuť skúmanú teóriu/hypotézu?“.* Použitie klasického prístupu *NHST* je nedostatočné, keďže vysoké *p*-hodnoty umožňujú jediné konštatovanie: *„Údaje by boli rovnaké aj vtedy, ak v „náhodnom hluku“ nie je „žiadny signál“*. Ak preto chceme hlbšie preskúmať štatisticky nevýznamné závery, zvýšiť informačnú hodnotu „triviálnych zistení“ a pýtať sa relevantnejšie otázky, nespočívajúcich iba v zamietaní nulových hypotéz, je potrebné siahnuť po inom vhodnom nástroji, ktorým je *ekvivalenčné testovanie* (*ET*) (Harms – Lakens 2018; Earp 2017; Lakens, 2017b; Lakens et al. 2018a).

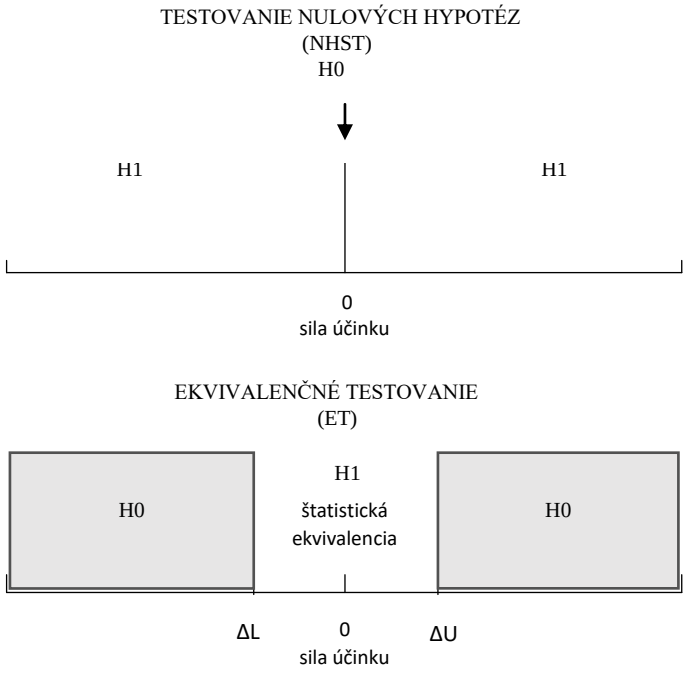
Ekvivalenčné testovanie – ako lepšie preskúmať triviálne výsledky

Ekvivalenčné testovanie (*ET*) je považované za variáciu *NHST* prístupu. Tvorí alternatívu klasického testovania a reaguje na potrebu lepšieho preskúmania „nulových“ záverov. Ide o test s vopred stanovenou mierou chybovosti, ktorý dokáže účinnejšie nahliadnuť do nesignifikantných výsledkov a urobiť ich viac informatívnymi (Harms – Lakens 2018). *ET* vylepšuje koherentnosť zistení a identifikuje skreslenia, spôsobené zverejňovaním falošne pozitívnych záverov. Jeho hlavnou výhodou je rozlíšenie nielen štatisticky nevýznamného výsledku, ale aj nepresvedčivého, štatisticky nevýznamného zistenia, ktoré je príliš malé na to, aby mohlo byť považované za zmysluplné (Harms – Lakens 2018). Dáva tak možnosť stanoviť rozdiel medzi triviálnym a zanedbateľným – štatisticky ekvivalentným zistením, ktoré je menšie ako vopred stanovená minimálna hranica záujmu *SESOI* (*Smallest Effect Size of Interest)* (Tryon 2001, In: Weber – Popov, 2012).

Hlavným cieľom *ET* je konštatovanie štatistickej ekvivalencie tzn. preukázanie neprítomnosti rozdielov, zmysluplných odlišností, absencie efektu alebo falzifikácia teórie. Pýtame sa: *„Je nameraný rozdiel natoľko malý, že sa môže považovať, vzhľadom na vopred stanovené minimálne hranice nášho záujmu, za triviálny?“* Alebo: *„Je možné odmietnuť rozdiely, rovnaké alebo väčšie ako je naša najmenšia hranica, ktorú považujeme za zmysluplnú?“.* Prípadne: *„Je efekt menší, ako najmenší možný, o ktorý sa zaujímame*?“. Zámerom nie je nájsť podporu v údajoch v prospech teórie alebo hypotézy (ako je tomu u *NHST* prístupu), ale preukázanie takej evidencie v dátach, ktorá na takúto podporu nestačí. Inými slovami: V *ET* prístupe sa nepýtame na očakávaný rozdiel/veľkosť účinku, ak je tento v skutočnosti rovný nule (čo je často nerealistické očakávanie NHST testovania), ale na najmenšiu veľkosť, ktorá by znamenala zamietnutie/podporu teórie, ak v skutočnosti rozdiel/efekt existuje (realistickejšie očakávanie *ET* prístupu). Za nulovú hypotézu sa tu nepovažuje predpoklad náhody/nezávislosti/nulových diferencií ale opak: populačné rozdiely minimálne rovnaké alebo väčšie, ako je najmenšia hranica vopred stano-

veného záujmu *SESOI* (detailné porovnanie *NHST* a *ET*, obrázok č. 1 a tabuľka č. 5). Vzhľadom na požadovanú minimálnu a vopred stanovenú veľkosť *SESOI*, konštatujeme štatistickú ekvivalenciu t.j. pozorovaný účinok, ktorý je v kontexte predpokladu existencie rozdielu nášho záujmu *SESOI* menší, a preto môže byť považovaný za zanedbateľný (Harms – Lakens 2018). *ET* testovaním možno urobiť relevantnejší záver, na základe ktorého možno teóriu zamietnuť alebo falzifikovať pomocou jasne špecifikovaných kritérií, čo urobí štúdiu viac informatívnou. Ak teória napriek tomu obstojí, je postavená na silných základoch pretože je vopred určené, ako minimálne veľký by rozdiel mal byť, aby viedol k relevantnej predikcií, prípadne k jej zamietnutiu.

Obrázok č. 1: **Rozdiel medzi testovaním nulových hypotéz** (*NHST*) **a ekvivalenčným testom** *(ET)*



Zdroj: Lakens et al. 2018a, krátené

Na správne zamietnutie *H0* je potrebné definovať ekvivalenčný priestor. Ide o interval vymedzený dolnou (ΔL) a hornou hranicou (ΔU), v ktorom platí štatistická ekvivalencia. Interval môže byť symetrický $<\Delta L = -0,3, \Delta U = 0,3>$ alebo asymetrický $<-0,2, 0,3>$ (Lakens et al. 2018b). Pre stanovenie ekviva-

lenčného výsledku musí byť pozorovaný efekt štatisticky významne väčší ako spodná (ΔL) a zároveň aj štatisticky významne menší ako horná ekvivalenčná hranica (ΔU). Na určenie štatistickej významnosti u obidvoch hraníc sa používa Welchowov *T* – test (*TOST – Two One Side Test*). Na rozdiel od klasického Studentovho *T*-testu je robustnejší, tzn. je odolný voči málo zastúpeným odľahlým prípadom, ktoré môžu dominovať výsledku a vychýliť hodnoty testovacej štatistiky. Lepšie tiež kontroluje chybu prvého druhu a je presnejší, ak nie sú splnené predpoklady homogenity rozptylov v skupinách. Ak je Welchowov *T*-test signifikantný pre dolnú aj hornú hranicu, potom sa *H0* zamieta a v záveroch postačuje interpretovať iba jednu – štatisticky významne väčšiu – *p*-hodnotu. Žiadny test ale neumožňuje potvrdiť rozdiely presne rovné nule. Zamietnutie *H0* a konštatovanie štatistickej ekvivalencie preto môže znamenať aj existenciu odlišnosti. Vieme ale povedať, že vo vymedzenom a vopred stanovenom intervale ide o príliš malý (respektíve nedostatočne veľký) výsledok (rozdiel) na to, aby bol relevantný a malo sa ním zmysel zaoberať.

Používanie *ET* by dávalo iba malý zmysel bez zdôvodnenia rozsahu ekvivalenčného priestoru, ktorý sa určuje nastavením šírky hraníc s dolnou a hornou hranicou (ΔL)/ΔU). Neexistuje tu ale jednoznačný návod, iba odporučenia. Weber – Poppov (2012) zdôrazňujú nastavenie hodnôt pred realizáciou testu, podobne ako pri stanovení chybovosti u klasického *NHST* testovania (stanovenie sily testu, dostatočného počtu respondentov pre identifikáciu efektu, či chyby prvého a druhého druhu). Použitím relevantnej literatúry (referenčný súhrn ukazovateľov z predošlých zistení, výsledky metaanalýz) je potrebné transparentne vyargumentovať šírku hraníc. Lakens et al. (2018a,c) zdôrazňujú aj preregistráciu, v rámci ktorej by mala prebehnúť rozsiahla diskusia, pripomienkujúca zvolenú možnosť. Morey – Lakens (2016) pripomínajú dôležitosť správneho určenia obidvoch hraníc v kontexte overovania teórií. Ak sa presne nedefinuje, ktoré hodnoty sú príliš malé na to, aby ich zamietnutie bolo zmysluplné a správne sa nešpecifikuje najmenší efekt záujmu *SESOI*, tak nemôže dôjsť k falzifikácii teórie. Široký ekvivalenčný priestor je tvorený vysokými *SESOI* hodnotami a na správne zamietnutie je potrebný menší počet respondentov. U užšieho priestoru potrebujeme naopak väčšiu vzorku. Keď štúdia s malým počtom respondentov konštatuje štatistickú ekvivalenciu, potom je možné vždy tvrdiť, že skutočný účinok aj tak neexistuje, keďže ekvivalenčný interval mohol byť v skutočnosti užší a my sme ho nastavením malého počtu respondentov iba nezachytili. Ak pre špecifickú otázku neexistujú zdroje o dostatočnej veľkosti vzorky, pomocou ktorej by sme – pri danej sile testu – vedeli preukázať šírku ekvivalenčného priestoru, potom voľbou hraníc môžeme poukázať aj na potrebu väčšej vzorky (Lakens 2017b).

Doteraz existuje iba málo objektívnych štandardov *SESOI* nastavenia. Detailnejšie návody sú uvedené v textoch Weber – Popova (2012), Simonshon

(2015), Harms – Lakens (2018), Lakens et al. (2018a). O konečnom zvolení vhodného spôsobu rozhodujú naše ciele. Spomínaní autori považujú arbitrárne nastavenie za najmenej vhodné. Tzn. subjektívne určenie napr. Cohenovho *d* = 0,5, čo je hodnota, ktorá pri porovnaní priemerov dvoch nezávislých skupín poukazuje na stredne veľký efekt. Na konštatovanie štatistickej ekvivalencie potom zamietame pozorovanie menšie ako *d* = ± 0,5. Ak máme iný cieľ, napr. pokúsiť sa falzifikovať predchádzajúce výsledky, ďalšia možnosť spočíva v použití hodnoty štúdie, z ktorej vychádzame. Pri metóde „malého teleskopu" je ekvivalenčnou hranicou identifikované *„d"* – pri danej veľkosti vzorky – ktoré sme schopní zachytiť 33 % silou testu (Simonsohn 2015a). Existujú ale aj iné možnosti – použitie priemeru, vypočítaného z koeficientov všetkých predchádzajúcich replikácií, alebo stanovenie orientačného bodu v podobe najmenšej štatisticky významnej hodnoty v texte, z ktorého vychádzame pri nadväzovaní na realizovaný výskum.

Pre identifikáciu ekvivalenčných hraníc možno použiť objektívne alebo subjektívne ukazovatele (Lakens, et al. 2018a). Konkrétne rozhodnutie závisí od povahy otázok, ktoré si kladieme, ako aj použitých ukazovateľov. Pre dizajn dvoch nezávislých skupín je objektívnym ukazovateľom napr. štandardizované Cohenovo *d*. Na jeho výpočet sa používa štandardná odchýlka, ktorej hodnota sa môže meniť v závislosti od veľkosti vzorky, čo ovplyvňuje aj výsledok ekvivalenčného testu. Niekedy má zmysel použiť hrubé skóre (napr. priemerný rozdiel 0,5 bodu na 7 stupňovej škále), ktoré je nezávislé od veľkosti štandardnej odchýlky. Keď sa vychádza z predchádzajúcich výsledkov, pre stanovenie ekvivalenčných hraníc by mali byť použité identické ukazovatele ako v štúdiách, z ktorých vychádzame

Posledným problémom je určenie veľkosti vzorky. Ak sa nad počtom respondentov zamyslíme vopred a nastavíme mieru chybovosti (chybu prvého i druhého druhu), tak nám to umožní urobiť správne rozhodnutie o ekvivalenčnom výsledku a identifikovať požadovanú *SESOI*. Je to dôležitý krok, pretože veľkosť vzorky má vplyv na najmenší možný rozdiel, ktorý je *ET* schopné zamietnuť. S veľkosťou vzorky rastie možnosť identifikácie menších rozdielov a z nich vyplývajúcich užších hraníc ekvivalenčného priestoru. Ak preto chceme zamietnuť veľmi malý efekt, potom s minimálnym počtom respondentov sa nám to nepodarí. Nevýhodou malých vzoriek sú aj širšie intervaly spoľahlivosti, ktoré pretínajú ekvivalenčné hranice. To znižuje nielen presný odhad populačných parametrov, ale aj pravdepodobnosť dosiahnutia štatisticky ekvivalenčného výsledku. Aby sa bolo možné vyhnúť uvedeným skresleniam, je nutné správne nastavenie veľkosti vzorky[12]. Ak požadujeme *SESOI*: Cohenovo *d* =

[12] Pre výpočet sa najčastejšie používa voľne dostupný program GPOWER. Zadajú sa hodnoty apriori požadovaného effect size napr. (*d* = 0,4), štatistickej významnosti (0,01), sily testu (0,8) dizajn (between group) vyberie sa metóda (*t*-test). Na

± 0,5, pri stanovenej hodnote alfa (napr. 0,05) a sile testu (0,9), potom požadovaný počet respondentov, pre dve nezávislé skupiny a obojstranné testovanie je 172 (86 pre každú skupinu). V interpretácii to znamená, že ak v skutočnosti platí $d = 0,5$ a v každej skupine je 86 respondentov, tak existuje iba 10 % pravdepodobnosť nenájdenia ($p < \alpha$) uvedenej hodnoty v očakávanej distribúcii koeficientov mier asociácie. Hodnota d < 0,3 nebude nikdy s 86 respondentmi v každej z obidvoch skupín štatisticky významná. Pokiaľ neurčíme inak, *SESOI* o hodnote 0,3, ako aj akékoľvek iné, menšie účinky, sú príliš malé na to, aby boli zaujímavé[13] (Lakens 2017b). V dlhodobom horizonte (in the long run) potom existuje 90 % pravdepodobnosť, že rovnakú alebo väčšiu hodnotu Cohenovho *d* (ak reálne existuje) naozaj zachytíme, a iba 5 % pravdepodobnosť, že mylne identifikujeme štatistickú ekvivalenciu aj tam, kde nie je (obojstranné testovanie). Prísnejšie kritériá na silu testu, nižšia miera požadovanej štatistickej významnosti zvyšujú nároky na počet respondentov[14].

V ekvivalenčnom testovaní sa odporúča porovnať výsledky so zisteniami *NHST* prístupu. Závery je potom možné lepšie zasadiť do kontextu a hlbšie zhodnotiť podobnosť/odlišnosť. V tabuľke č. 5 uvádzame štyri možnosti, ktoré môžu nastať. Prvá možnosť ukazuje štatisticky nevýznamný výsledok u *NHST*, spoločne so štatisticky významným *ET*, nižším ako najmenší efekt nášho záujmu *SESOI*. Takáto kombinácia indikuje neprítomnosť podpory pre skúmanú teóriu u *NHST* a rovnako aj zanedbateľný efekt v *ET*. Pri druhej možnosti klasické testovanie nulovú hypotézu opäť nezamieta a *ET* konštatuje neprítomnosť ekvivalenčného výsledku. V dôsledku tohto rozporu ostávame nerozhodnutí a na vyvodenie definitívneho záveru je potrebných viac údajov. Tretia možnosť hovorí o štatisticky významnom rozdiele, ale tiež aj o ekvivalenčnom výsledku menšom, ako čokoľvek, na čom nám záleží. Výsledky *NHST* sú síce štatisticky významné, ale z hľadiska *ET* zanedbateľné a prakticky nepoužiteľné. Tretí variant je tak najväčším benefitom *ET* testovania - umožňuje urobiť záver o nerelevantnom rozdiele, ktorý je menší ako by sme vo výsledku očakávali aj keď je výsledok v klasickom testovaní signifikantný. A nakoniec, štvrtý variant ukazuje štatisticky významný výsledok *(NHST)* a v *ET* teste, je rozdiel väčší ako minimálna hodnota *SESOI*, ktorú požadujeme zamietnuť. Ide o presvedčivý dôkaz v prospech rozdielu, dosť veľkého na to, aby sme sa ním zaoberali. Štyri možnosti výsledkov platia nielen pre priemery dvoch nezávislých skupín.

základe vstupných parametrov sa potom vypočíta veľkosť vzorky. V balíku TOSTER, je rozhranie nastavené priamo na ET, ktorý používa Welchov *T*-test pre obidve ekvivalenčné hranice.

13 Ako sila testu determinuje hranice *SESOI*, viď tiež blog: The 20 % Statistician, dostupné na: http://daniellakens.blogspot.com/2017/05/how-power-analysis-implicitly-reveals.html

14 Ak chceme identifikovať hranice Cohenovho $d = \pm 0,5$ a požadujeme 90% silu testu na 1% hladine významnosti, potom potrebujeme ešte väčšiu vzorku: 254 respondentov (127 pre každú skupinu/dvojstranné testovanie, balík TOSTER), (Lakens, 2017a, Lakens et al. 2018a)

Rovnako tomu je aj pre iné typy dizajnov – korelácie, porovnanie percentuálnych rozdielov alebo vyhodnotenie metaanalýz (Lakens 2017b).

Tabuľka č. 5: **Alternatívy kombinovaného testovania ET a NHST**

	NHST testovanie nulových hypotéz	*ET* ekvivalenčné testovanie	ROZHODNUTIE *NHST*/*ET*
1.	$p > 0,05$ *H0* nezamietame	$p \leq 0,05$ *H0* zamietame ekvivalenčný výsledok, rozdiel < *SESOI*	nie je štatistická významnosť je štatistická ekvivalencia
2.	$p > 0,05$ *H0* nezamietame	$p > 0,05$ *H0* nezamietame nie je ekvivalenčný výsledok, rozdiel > *SESOI*	nie je štatistická významnosť ani štatistická ekvivalencia
3.	$p \leq 0,05$ *H0* zamietame	$p \leq 0,05$ *H0* zamietame ekvivalenčný výsledok, rozdiel < *SESOI*	je štatistická významnosť je štatistická ekvivalencia
4.	$p \leq 0,05$ H0 zamietame	$p > 0,05$ H0 nezamietame nie je ekvivalenčný výsledok, rozdiel >SESOI	je štatistická významnosť nie je štatistická ekvivalencia

Pozn**:** *H0* pre *NHST*: predpoklad nulového rozdielu, náhody, absencia súvislostí, neprítomnosť vzťahu
H0 pre *ET*: predpoklad rozdielov v populácii väčších ako *SESOI*
SESOI: *Smallest Sffect Size of Interest* – najmenšie (požadované) hranice nášho záujmu
Zdroj: upravené na základe Lakens(2017b) (podľa obr. č. 1)

Pre prvotnú predstavu ako *ET* pracuje s nulovými hypotézami, uvádzame dva príklady z Lakens et al. (2018a). *Budú sa líšiť reakcie mužov a žien pri pozeraní hororu?* Zamietame *H0* o rozdieloch v reakciách na hororové scény v závislosti od pohlavia. *Kedy dochádza k subjektívnemu zlepšeniu u ľudí pociťujúcich depresiu?* (Button et al. 2015). Norman et al. (2003) píšu o najmenšom klinickom efekte 0,5 štandardnej odchýlky, keď sa dá uvažovať o zlepšení.

Problematiku ilustrujú ďalšie dva praktické príklady, v ktorých sú použité reálne údaje z reprezentatívneho prieskumu násilia, páchaného na deťoch 8. a 9. ročníka IVPR (2017). Na vyhodnotenie je použitý balík *Toster* v programe R[15] (Lakens 2017a). Toster balík umožňuje grafické porovnanie výsledkov *NHST* i *ET*, spoločne so zodpovedajúcimi intervalmi spoľahlivosti. Príklad obsahuje všetky potrebné kroky: nastavenie veľkosti vzorky, ekvivalenčného priestoru, stanovenie jednotiek najmenšieho záujmu *SESOI*, interpretáciu vzhľadom na zvolené ekvivalenčné hranice.

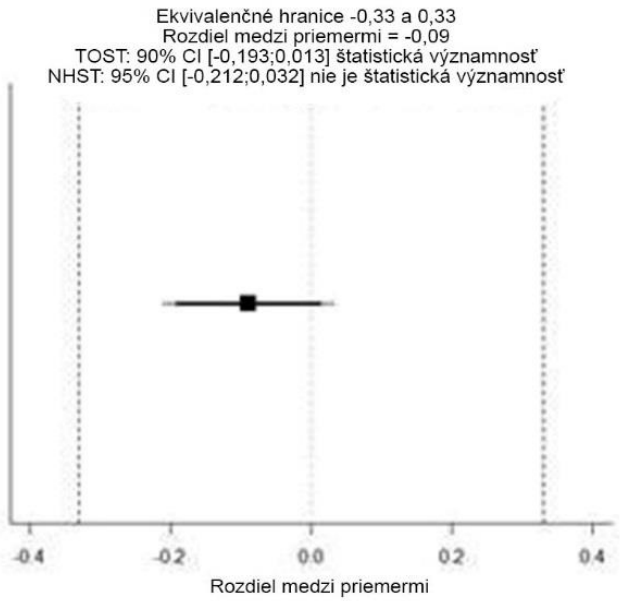
[15] *ET* možno vyvolať príkazom v programe R. Ide o balík TOSTER: >if(!require(TOSTER)){install.packages('TOSTER')} > library(TOSTER) alebo použiť užívateľsky kód v programe R dostupný na MOOCs (Lakens 2017a, Lakens et al. 2018a)

Ženy sa môžu stať ľahkým terčom domáceho násilia zo strany dospelého partnera (WHO 2017). Feministické sociálne teórie, zdôrazňujúce rodovo podmienené násilie, prichádzajú s jedným z vysvetlení: rodové stereotypy „predpisujú" mužom a ženám ako sa majú správať (Harrington a kol. 2006). Rodové stereotypy sú zjednodušené, nerealistické obrazy „mužskosti" a „ženskosti". U mužov môže byť ich súčasťou potreba nerovných mocenských vzťahov, prejavujúcich sa kontrolou ženy, vrátane používania násilia. Pomocou ekvivalenčného testovania sa dá preskúmať úloha rodových stereotypov aj v starostlivosti o deti. Dievčatá by mohli čeliť v rodine rozsiahlejším skúsenostiam (väčšiemu počtu rôznych prejavov násilia) v podobe fyzického trestania, represívnejšej výchovy, zlého zaobchádzania s cieľom zabezpečenia budúcej konformnejšej a pasívnej rodovej role. *Existuje v dátach evidencia na odmietnutie hypotézy násilnej socializácie v zhode s rodovými stereotypmi, v podobe väčšieho rozsahu konfrontácie s prejavmi násilia, v závislosti od pohlavia?* V IVPR prieskume prevalencie násilia, páchaného na deťoch 8. a 9. ročníkov, bolo použitých 11 rôznych prejavov fyzického trestania. Porovnali sme priemerné počty konfrontácie u chlapcov a dievčat (N = 2857). Aby bolo možné rozhodnúť, či sa v dátach nachádza ekvivalenčný výsledok, je najskôr nutné nastaviť hranice *SESOI*. Na základe expertnej diskusie a publikačného prehľadu (ilustratívne nastavenie v online rešerši nebolo k dispozícii), považujeme za rozumnú najmenšiu pozorovanú hodnotu záujmu: $d = \pm 0{,}2$ (malý efekt). Ekvivalenčný interval je pomerne úzky a na zamietnutie *H0,* v podobe rozdielov rovnakých alebo väčších ako je uvedený koeficient, bude treba prístup k väčšiemu počtu respondentov. Ak požadujeme užšie hranice, potom je nutné mať dostatočnú veľkosť vzorky. Inak hrozí, že údaje nie sú - vzhľadom na veľkosť hraníc, informatívne, pretože nedokážu zachytiť rozdiely. Keďže je vždy nejasné, aký efekt môžeme očakávať, musíme mať istotu, že sila testu je dostatočne veľká na odhalenie najmenšieho možného efektu *SESOI,* o ktorý sa zaujímame. Ak požadujeme 80 % silu testu s maximálnou chybou prvého druhu 5 %, na zamietnutie *SESOI*: $d = \pm 0{,}2$ potrebujeme 858 respondentov (429 v každej z dvoch skupín) (výpočet v balíku TOSTER). Počet respondentov na stanovený *SESOI* dosiahol a ďaleko ho prevyšuje: N = 2857. Existuje preto iba malá pravdepodobnosť, že by sme omylom nezamietli výsledok, ktorý by pre zvolenú hranicu mohol byť ekvivalenčný. Alternatívy výsledkov kombinovaného testovania *ET* a *NHST* zobrazuje graf č. 3 a č. 4.

Použitím Welchovho robustného *T*-testu získavame štatisticky nevýznamné rozdiely $t(2763{,}07) = -1{,}442$, $p = 0{,}149$. Aplikácia klasického *NHST* prístupu nezamieta predpoklad H0 o nulových rozdieloch v rozsahu fyzického násilia medzi chlapcami a dievčatami (chlapci: $M = 1{,}33$, $SE = 0{,}042$, dievčatá: $M = 1{,}42$, $SE = 0{,}046$, $p > 0{,}05$). Odlišnosti v priemeroch sú príliš malé (-0,09). (95 % *CI* pre *NHST* v grafe č. 3 prekrýva nulovú hodnotu priemerných rozdie-

lov). Dáta sú v zhode s výsledkom, ktorý by sme dostali aj vtedy, ak by platila náhoda. *Nenachádzame podporu pre hypotézu podporujúcu rodové stereotypy násilnejšou socializáciou dievčat, v podobe väčšej konfrontácie s rôznymi prejavmi násilia v porovnaní s chlapcami.*

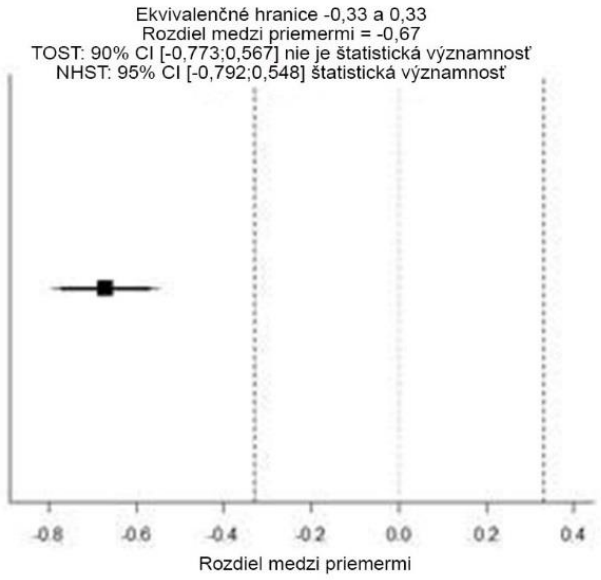
Graf č. 3: **Výsledky kombinovaného testovania ET A NHST** (alternatíva 1)



V ekvivalenčnom testovaní je – vzhľadom na stanovený najmenší efekt záujmu *SESOI*: ($d = \pm 0{,}2$), výsledok Welchovho *T*-testu štatisticky významný: $t(2763{,}07) = 3{,}844$, $p = 0{,}0000619$, na 5 % hladine významnosti pre dolnú aj hornú hranicu (SESOI zodpovedá v grafe č. 3 ekvivalenčný priestor hrubého skóre s hranicami ± 0.33). 90 % *CI* neprekračuje ekvivalenčné hranice ani v jednom smere. *H0* o rozdieloch v pohlaví rovných alebo väčších ako je minimálna hranica záujmu *SESOI* ($p < 0{,}05$) sa zamieta. Rozdiel síce nie je presne nulový, ale dostatočne malý na to, aby sme nulovú hypotézu pre *ET* odmietli. Ide o štatisticky ekvivalenčný výsledok. *Dôkazná hodnota pre podporu rodových stereotypov, meraná rozdielom v priemernom rozsahu fyzických trestov väčších, ako je nastavená SESOI (d = ± 0,2) v závislosti od pohlavia, nie je dostatočná.* V celkovom zhrnutí a porovnaní obidvoch alternatív konštatujeme štatisticky nevýznamný výsledok *NHST*, ale rovnako (a to je pridaná hodnota *ET*) aj nepresvedčivý, ekvivalenčný výsledok, príliš malý na to, aby zistené rozdiely mohli byť považované za zmysluplné, vzhľadom na vopred zvolenú

minimálnu hodnotu záujmu. Diferencie v prejavoch násilia medzi chlapcami a dievčatami existujú. Rozdiely sú však štatisticky nevýznamné (*NHST*), menšie ako čokoľvek, čo nás zaujíma (*ET*) a z praktického hľadiska irelevantné (v tab. č.4 je tento výsledok v zhode s prvou alternatívou výsledkov). Hypotéza o podpore rodových stereotypov rozsiahlejšou, represívnou výchovou dievčat (aj keď dievčenská skupina vykazovala v priemere zanedbateľne vyššie hodnoty fyzickej konfrontácie), zamieta *NHST* aj *ET*.

Graf č. 4: **Výsledky kombinovaného testovania ET A NHST** (alternatíva 2)



Výsledky alternatívneho spracovania údajov obsahuje graf č. 4. Sú to ilustratívne údaje, v ktorých vstupné parametre (počet respondentov, štandardná odchýlka, priemer konfrontácie s prejavmi násilia u chlapcov) zostali rovnaké, až na priemer počtu prejavov fyzického násilia u dievčat, ktorý bol (umelo) zvýšený na hodnotu $M = 2{,}00$[16]. Testovanie nulových hypotéz *NHST* ukazuje štatisticky významné rozdiely v konfrontácii s násilím v prospech dievčat (-0,67), $t(2763{,}07) = -10{,}738$, $p < 0{,}001$. V *ET* je Welchow *T*-test štatisticky nevýznamný pre dolnú ekvivalenčnú hranicu $t(2763{,}07) = -5{,}452$, $p = 1{,}00$. Preto nekonštatujeme štatistickú ekvivalenciu, ktorej podmienkou je štatistická významnosť pre obidve hranice (v tab. č. 4 je tento výsledok v zhode so štvrtou alternatívou výsledkov). Prístupom *NHST* sa nulová hypotéza zamieta, t.j. na-

[16] Ide o dodatočná úpravu, pred ktorou text varuje - zmena parametra uvedená z didaktických dôvodov.

chádzame v dátach podporu stereotypných rodových úloh s rozdielnym rozsahom fyzických trestov u chlapcov a dievčat. V ekvivalenčnom testovaní *ET*, naopak, neexistuje dostatočná dôkazná hodnota, aby sme mohli zamietnuť H0 o rozdieloch väčších, ako je minimálny rozsah nášho záujmu, definovanom ekvivalenčnými hranicami. Diferencie sú príliš veľké na to, aby sme ich mohli ignorovať. Inými slovami: aj *ET* podporuje predpoklad prísnejšej výchovy a starostlivosti u dievčat, čo je v zhode s podporou hypotézy o rodových stereotypoch. Zvýšením hodnoty priemeru konfrontácie v dievčenskej skupine bol v grafe č. 4 dosiahnutý odlišný výsledok, v porovnaní s predchádzajúcim grafom č. 3. Pre ľahšie zhrnutie porovnáva záverečná tabuľka č.6 základné princípy testovania nulových hypotéz *(NHST)* a ekvivalenčného testovania *(ET).*

Tabuľka č. 6: **Porovnanie princípov testovania nulových hypotéz** (*NHST)* **a ekvivalenčného testovania** *(ET)*

NHST vs. *ET*	*NHST*	*ET*
DEFINÍCIA *H0*	predpoklad neexistencie rozdielov/vzťahu/efektu	predpoklad o rozdiele rovnakom alebo väčšom, ako vopred stanovená minimálna hodnota záujmu *SESOI*
AKO SA ROZHODUJE O ZAMIETNUTÍ *H0*	*p*-hodnoty nižšie ako vopred nastavená hladina významnosti 0,05. $p <$ (0,05, 0,01, 0,001) alebo pomocou intervalov spoľahlivosti	na základe výsledkov intervalov spoľahlivosti, nepretínajúcich vopred stanovenú dolnú a hornú ekvivalenčnú hranicu
NASTAVENIE HRANÍC EKVIVALENCIE	nie sú, čím nižšia *p*-hodnota, tým vyššia štatistická významnosť a menšia pravdepodobnosť, že dáta pochádzajú zo základného súboru, kde platí *H0*	na základe požadovanej najmenšej/minimálnej hodnoty záujmu *SESOI* v podobe koeficientov mier asociácie (napr. Cohenovo *d*, veľkosť korelácie, alebo rozdielov v %)
CIEĽ	zamietnutie H0, preukázanie štatisticky významných rozdielov/vzťahu/efektu, kedy dáta odporujú *H0*	zamietnutie *H0*, preukázanie ekvivalencie/efektu menších ako je vopred stanovená veľkosť dolnej a hornej hranice ekvivalencie

Zdroj: spracované na základe dostupných textov, Lakens – Evers 2014; Lakens 2017b; Lakens, et al. 2018a,c

Obmedzenia a diskusia

Ako upozorňujú autori, aj pri používaní uvedených nástrojov existujú obmedzenia. P-krivky boli pôvodne málo citlivé na kombináciu vyhodnocovania textov s nedostatočnou dôkaznou hodnotou a intenzívnym *p*-hackingom. Novšie verzie tento nedostatok odstránili. *P*-krivky hodnotia iba publikované texty a závery z nich plynúce neznamenajú to isté, ako ne/správnosť teórie. Teória môže byť pravdivá, len ju nameraná distribúcia nemusí podporovať. Vyhodnotenie je založené na predpoklade, že veľkosť štatisticky významných *p*-hodnôt nemá vplyv na ich zverejnenie – na rozdiel od selektívneho výberu použitím *„p-hackingu“*, podľa toho čo „funguje“. *P*-krivky obsahujú hodnoty $p < 0{,}05$,

čo znamená, že vylučujú štatisticky nevýznamné výsledky, ktoré sú pri reálnom účinku/rozdiele zriedkavé. Nameraná distribúcia môže byť pravotočivá a vylučovať selektívne vykazovanie, ale s vecne zanedbateľným efektom. Obmedzenia sú aj pri niektorých typoch výskumného dizajnu. (bližšie sú inštrukcie rozpísané v tabuľke na s. 541-542, Simonsohn et al. 2014a alebo v Official User-Guide to the P-curve, dostupnom na www.p-curve.com) (Simonshon et al. 2014, 2015, 2016).

Ekvivalenčné testovanie je technikou, ktorú možno rovnako použiť pre rôzne dizajny – experimentálny, observačný. Aj v sociológii je jej hlavnou výhodou falzifikácia hypotéz, s posilnením dôveryhodnosti teórií, s dobrou predikciou, ktoré v *ET* obstoja. Štúdie získajú väčšiu informačnú hodnotu, a reálne základy – na rozdiel od testovania „ničotných nulových hypotéz". Ak výsledok padne do ekvivalenčnej hranice, už sa nemôže stať, že zistenie bude štatisticky významne ale prakticky irelevantné. Vieme tiež povedať, od akej hranice považujeme efekt za nehodný pozornosti.

ET testy môžu byť použité aj na postupné odmietnutie menších efektov, so zväčšujúcim sa počtom respondentov, až pokiaľ nebude nikto ochotný investovať čas a zdroje, potrebné na odmietnutie ešte menšieho efektu. Pre korektné rozhodnutie je rovnako nutné odolať pokušeniu stanoviť hypotézy a *SESOI* hranice až potom, ako sme prezreli výstupné údaje a zistili, čo (ne)funguje. Ekvivalenčné hranice by mali byť špecifikované pred zberom údajov, ideálne ako súčasť preregistrácie. V opačnom prípade sa vždy môžeme pozrieť na dáta a vybrať hranice vhodne široké pre potreby vykonania testu. Obmedzenia sa týkajú aj vysvetlení zvoleného rozsahu *SESOI*. Nemá zmysel reportovať text bez stručného výkladu, v akom kontexte je uvedená hranica zaujímavá, ak sú prechádzajúce – predbežné výsledky už k dispozícii (Morey – Lakens 2016; Lakens 2017b; Lakens et. al. 2018a,c).

Milan Fico *skončil štúdium sociológie na Filozofickej fakulte Univerzity Komenského v Bratislave. Pracuje v Inštitúte pre výskum práce a rodiny.*

LITERATÚRA

BELSKY, J., 1980: Child Maltreatment: An Ecological Integration, American Psychologist 35(4). Online cit. [4.3.2019] dostupné na: https://www.researchgate.net/publication/15812067_Child_Maltreatment_An_Ecological_Integration

BENJAMIN, D. J. et al., 2017: Redefine Statistical Significance. Nature Human Behaviour. Online, cit. [20.6.2019] dostupné na: doi:10.17605/OSF.IO/MKY9J

BUTTON, K. S. – KOUNALI, D. – THOMAS, I. – WILES, N. J. – PETERS, T. J. – WELTON, N. J. – LEWIS, G., 2015: Minimal clinically important difference on the Beck Depression Inventory – II according to patients persepctive. Psychological Medicine, 45, 3269-3279. doi:10.1017/S0033291715001270

CUMMING, G., 2012: Understanding the New Stastistics, Efect Sizes, Confidence Intervals, and Meta-Analysis, Routledge, ISBN:978-0-415-87967-5.
CUMMING, G. – JAGEMAN, C. R., 2017: Introduction to the New Statistics (Estimation, Open Science, and Beyond), Routledge. ISBN: 978-1-315-70860-7.
CUMMING, G., 2008: Replication and *p* Intervals: *p* Values Predict the Future Only Vaguely, But Confidence Intervals do Much Better, Perspectives on Psychological Science. 3:286. Online, cit. [10.4.2019] dostupné na: http://pps.sagepub.com/content/3/4/286
DE VAUS, D., 2002: Analyzing Social Science Data, 50 Key Problems in Data Analysis, Sage publication. ISBN 978-0-7619-5973-3.
DE VAUS, D., 2014: Surveys in Social Research, Routledge. ISBN: 978-0-415-53015-6.
EARP, D. B., 2017: The Need for Reporting Negative Results – a 90 Year Update. Journal of Clinical and Translational Research, 3(S2): 1-4. Online, cit. [14.9.2019] dostupné na doi:10.18053/jctres.03.2017S2.001
DIENES, Z., 2008: Uderstanding Psychology as a Science, Palgrave Macmillian. ISBN-13: 978-0-230-54231-0
DUCHESNE, G. A. – HÉBERT, M. – DASPÉ, ME., 2017: Gender as predictor of posttraumatic stress symptoms and externalizing behavior problems in sexually abused children, Child Abuse and Neglect, Vol.64, Online cit. [17.5.2020] dostupné na: doi.org/10.1016/j.chiabu.2016.12.008
FANELLI, D., 2012: Negative Results are Disappearing from Most Disciplines and Countries, Scientometrics 90: 981-904. Online, cit. [15.8.2019] dostupné na: doi: 10.1007/s11192-011-0494-7
FICO, M., 2017: Prevalencia násilia páchaného na deťoch 8. a 9. ročníkov. Inštitút pre výskum práce a rodiny (IVPR). Online cit. [31.12.2019], dostupné na: https://www.ceit.sk/IVPR/ images/IVPR/vyskum/2017/Fico/prevalencia_nasilia_pachaneho_na_detoch_2017.pdf
FRANCO, A. – MALHOTRA, N. – SIMONOVITS, G., 2014: Publication Bias in the Social Sciences: Unlocking the File Drawer. Science, 345(6203), 1502-1505. Online, cit.[15.4.2019] dostupné na: doi:10.1126/science.1255484
FIELD, A., 2018: Discovering Statistisc Using IBM SPSS tatistics. 5th. Edition, Sage. ISBN 978-1-5264-1951-4.
FIELD, A. – HOLE, G., 2003: How to Design and Report Experiments, Sage. ISBN, 9780-7-619-738-29.
GERBER, S. A. – MALHORTA, N., 2008: Publication Bias in Empirical Sociological Research. Do Arbitrary Significance Levevls Distort Published Results? Sociological Methods and Research, Volume 37/1, 3-30. Online, cit. [10.4.2019] dostupné na: doi:10.1177/0049124108318973
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. Online, cit. [20.5.2019] dostupné na: doi: 10.1007/s10654-016-0149-3

HARMS, CH., - LAKENS, D., 2018: Making „Null Effects“ Informative: Statistical Techniques and Inferential Framework. Journal of Clinical and Translational Research. Online, cit. [15.9.2019] DOI: 10.18053/jctres.03.2017S2.007
HARRINGTON, A. a kol., 2006: Moderní sociálni teorie (základní témata a myšlenkové proudy), Portál. ISBN 80-7367-093-3.
IOANNIDIS, P. A. J., 2005: Why Most Published Findings are False, PLOS, Medicine. Online, cit. [30.11.2019] dostupné na: https://doi.org/10.1371/journal.pmed.0020124
JACKSON, M. – COX, D. R., 2013: The Principles of Experimental Design and Their Application in Sociology. Annual Rreview of Sociology. Online, cit. [13.1.2020] dostupné na: https://doi.org/10.1146/annurev-soc-071811-145443
KANOVSKÝ, M., 2016: Robustné štatistické metódy v sociálnych vedách. Slovenská asociácia sociálnej antropológie. ISBN 978-80-970587-3-9.
KAČMÁR, P., 2019: Priming in the Context of Goal-Setting Theory (Review and *P*-Curve Analyses):The Good News, the Bad News and Some Recommendations. Online, cit. [1.12.2019], dostupné na: https://www.researchgate.net/publication/335422172
KARKOŠKOVÁ, S. – ROPOVÍK, I., 2018: The Prevalence of Child Abuse Among Slovak Late Adolescent. Researchgate. Online, cit. [5.9.2019] dostupné na: https://www.researchgate.net/publication/324006632
KERR, N. L., 1998: HARKing: Hypothesizing After the Results are Known. Personality and Social Psychology Review, 2.
KRUSHKE, K. J., 2015: Doing Bayesian Data Analysis, A Tutorial with R, JAGS, and Stan. Edition 2, Elsevier. ISBN 978-0-12-405888-0.
LAKENS, D. – EVERS, R., K., E., 2014: Sailing From the Seas of Chaos Into the Corridor of Stability: Practical Recommendations to Increase the Informational Value of Studies. Perspecitves on Psychological Science, Vol. Í(3) 27-292. Online, cit. [10.4.2019] dostupné na: doi: 10.1177/1745691614528520
LAKENS, D., 2017a: Improve your statistical Inference, MOOCs, dostupné na: www.coursera.org
LAKENS, D., 2017b: Equivalence Tests: A Practical Primer for *t*-Test, Correlations, and Meta-Analyses. Online, cit. [15.5.2019] dostupné na: https://doi.org/10.1177%2F1948550617697177
LAKENS, D., 2017c: Professors Are Not Elderly: Evaluating the Evidential Value of Two Social Priming Effects through *P*-Curve Analyses. PsyArXiv. Online, cit. [12.10.2019] dostupné na: https://doi.org/10.17605/OSF.IO/3M5Y9
LAKENS, D. – ETZ, A.J. 2017d: Too true to be bad: When sets of studies with significant and non-significant findings are probably true. Social Psychological and Personality Science. doi: 10.1177/1948550617693058
LAKENS, D. – SCHEEL, M. A., – ISAGER, P., 2018a: Equivalence Testing for Psychological Research: A Tutorial, Advances in Methods and Practices in Psychological Science. 1-11. Online, cit. [10.7.2019] dostupné na: doi: 10.1177/2515245918770963

LAKENS, D. – ADOLFI, F. G. – ALBERS, C. J. – ANVARI, F. – APPS, M. A. J. – ARGAMON, S. E. – ZWAAN, R. A., 2018b: Justify Your Alpha. Nature Human Behaviour. doi: 10.1038/s41562-018-0311-x

MASICAMPO, E. J. – LALANDE, D. R., 2012: A Peculiar Prevalence of *p* Values Just Below .05. The Quarterly Journal of Experimental Psychology, 65:11, 2271-2279 online, cit. [10.4.2019] dostupné na: http://dx.doi.org/10.1080/17470218.2012.711335

MEEHL, P. E., 1990: Why Summaries of Research on Psychological Theories are Often Uninterpretable. Psychological Reports, 66, 195-244.

MOREY, D. R. – LAKENS, D., 2016: Why Most Psychology is Statistically Unfalsifiable (draft). Online cit. [21-11.2019] dostupné na: https://raw.githubusercontent.com/richarddmorey/psychology_resolution/master/paper/response.pdf

MUNAFÓ, M. – NOSEK, B. – BISHOP, D., 2017: A Manifesto for Reproducible Science. Nat Hum Behav 1, 0021. Online cit [20-03.2020], dostupné na: https://doi.org/10.1038/s41562-016-0021

NEGRIFF, S. – SCHNEIDERMAN, J. U. – SMITH, C. – SCHREYER, J. K. – TRICKETT, P. K., 2014: Characterizing the Sexual Abuse Experiences of Young Adolescents. Child Abuse & Neglect, 38(2), 261-270. Online cit. [21-11.2019] dostupné na: doi:10.1016/j.chiabu.2013.08.021

NOSEK, B., et al. 2018: The Preregistration Revolution, Proceedings of the National Academy of Sciences of the United States of America (PNAS, online cit [20-03.2020] , dostupné na: https://doi.org/10.1073/pnas.1708274114

RABUŠIC, L. – SOUKUP. L. – MAREŠ, P., 2019: Statistická analýza sociálněvědných dat (prostředníctvim SPSS), Masarykova Univerzita, Munipress, ISBN 978-80-210-9248-8

ROPOVIK, I., 2018: On the Meaning of *p*-values: Criticism of Significance Tests Revisited [O význame *p*-hodnôt: reflexia na silnejúcu kritiku testov významnosti. Československá psychologie. 61. 502-516. online, cit. [14.7.2019] dostupné na researchgate.

SCHMIDT, S., 2009: Shall We Really Do It Again? The Powerful Concept of Replication is Neglected in the Social Science. Review of General Psychology, Vol. 13, No.2: 90-100. Online cit. [12.5.2020] dostupné na: https://doi.org/10.1037%2Fa0015108

SIMMONS, J. P. – NELSON, D. L. – SIMMONSOHN, U., 2011: False positive psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as significant, Sage Journals, Online cit. [12.5.2020] dostupné na: doi.org/10.1177%2F0956797611417632

SIMMONS, J. P. – SIMONSOHN, U., 2017: Power Posing: *P*-curving the Evidence, Psychological Science, 1-7. Online cit. [2.12.2019] dostupné na: doi: 10.1177/0956797616658563

SIMONSOHN, U., 2015a: Small Telescopes: Detectability and the Evaluation of Replication Result. https://doi.org/10.1177/0956797614567341

SIMONSOHN, U. – SIMMONS, P. J. – NELSON, D. L., 2014a: *P*-curve: A key to the File-Drawer. Journal of Experimental Psychology, Vol. 143, No. 2, 534-547. Online cit. [15.8.2019] dostupné na: http://pages.ucsd.edu/~cmckenzie/Simonsohnetal2014JEPGeneral.pdf

SIMONSOHN, U. – SIMMONS, P. J. – NELSON, D. L., 2014b: *P*-curve and Effect Size: Correcting for Publication Bias using only Significant Result. Online cit. [15.8.2019] dostupné na: doi: 10.1177/1745691614553988

SIMONSOHN, U. – SIMMONS, P. J. – NELSON, D. L., 2015: Better *P*-curves, 2015: Making *p*-Curve Analysis More Robust To Error, Fraud and Ambitious *P*-Hacking. A Reply to Ulrich Miller online cit. [15.8.2019] dostupné na: http://p-curve.com/paper/Better%20p-curves%202015%2011%2026.pdf, DOI: 10.1037/xge0000104

SIMONSOHN, U. – NELSON, D. – SIMMONS, P. J., 2016: *P*-curve Won´t Do your Laundry, But it will Distinguish Replicable fromn Non-Replicable Findings in Observational Research: Comment on Bruns and Ioannidis. Online cit. [15.8.2019] dostupné na: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3289096

SOBSEY, D. – RANDALL, W. – PARRILA, R. K., 1997: Gender Differences in Abused Children with and Without Disabilities. Child Abuse & Neglect, 21(8), 707-720. Online cit. [12.10.2019] dostupné na: doi:10.1016/s0145-2134(97)00033-1

SOUKUP, P., 2010: Nesprávná užívání statistické významnosti a jejich možná řešení. Institut sociologických studií Fakulta sociálních věd, Univerzita Karlova v Praze, online cit. [24.11.2019] dostupné na researchgate.

SOUKUP, P., 2019: *P* a *d* (Používaní statistické a věcné významnosti v českých sociálních vědách), Sociologický časopis, Vol. 55, No. 2: 215-253. Online cit. [24.11.2019] dostupné na: https://doi.org/10.13060/00380288.2019.55.2.459

SOUKUP, P. – RABUŠIC, L., 2007: Několik poznámek k jedne obsesi českých socialních věd – statistické vyznamnosti, Sociologický časopis, Vol. 47, No. 2: 379-395. Online cit. [24.11.2019] dostupné na: http://sreview.soc.cas.cz/cs/issue/15-sociologicky-casopis-czech-sociological-review-2-2007/201

SCANNAPIECO, M. – CARRICK, C. K., 2005: Understanding Child Maltreatment, An Ecological and Developmental Perspective. University Oxford Press. ISBN-13 978-0-19-515678-2.

VAN DE SCHOOT, R. – KAPLAN, D. – DENISSEN, J. – ASENDORPF, J. B. – NEYER, F. J. – AKEN, M. A., 2014: A Gentle Introduction to Bayesian Analysis: Applications to Developmental Research. Child Development 85(3), 842-860. Online cit. [25.11.2019] dostupné na: https://doi.org/10.1111/cdev.12169

WEBER, E. – POPOVA, L., 2012: Testing Equivalence in Communication Research: Theory and Application. Communication Methods and Measures, 6:3, 190-213. Online cit. [7.9.2019] https://doi.org/10.1080/19312458.2012.703834

ZILIAK, T. S. – MCCLOSKEY, D., 2011: The Cult of Statistical Significance, How the Standard Error Cuts Us Jobs. Justice and Lives, The University of Michigan. ISBN.-13: 978-0-472-07007-7.

SPRÁVY, DOKUMENTY A WWW

American Statistical Assocation releases statement on statistical significance and *p*-values, Provides Principles to Improve the Conduct and Interpretation of Quantitative Science (2016). Online, cit. [24.6.2019] dostupné na: https://www.amstat.org/asa/files/pdfs/P-ValueStatement.pdf
Center for Open Science, Online, cit. [10.5.2020] dostupné na: www.cos.io
European Report on Preventing Child Maltreatment, WHO, 2013: Online, cit. [24.11.2019] dostupné na: http://www.euro.who.int/__data/assets/pdf_file/0019/217018/European-Report-on-Preventing-Child-Maltreatment.pdf
Social science replication project, 2016, Online, cit. [10.5.2020] dostupné na: http://www.socialsciencesreplicationproject.com/
Violence against woman, WHO, 2017, Online, cit. [13.1.2020] dostupné na: https://www.who.int/news-room/fact-sheets/detail/violence-against-women
Making the p-curve: www.p-curve.com