PODPORA NULOVÉ HYPOTÉZY A JEJÍ MISKONCEPCE V PSYCHOLOGII: TEORETICKÉ PŘEDSTAVENÍ TESTOVÁNÍ EKVIVALENCE

DAVID LACKO[1], TOMÁŠ PROŠEK[1]
[1] Psychologický ústav, Filozofická fakulta, Masarykova univerzita

***Abstrakt**: Tento teoretický článek představuje způsoby, kterými lze statisticky argumentovat ve prospěch nulové hypotézy. Představuje čtyři způsoby, které lze využít k testování ekvivalence: dva jednostranné testy (TOST), p-hodnotu druhé generace (SGPV), Bayesův faktor (BF) a oblast praktické ekvivalence (ROPE). Článek je doplněn o praktické ukázky možných výsledků TOST. Součástí článku je také nezbytné objasnění logiky testování hypotéz a p-hodnoty a kritická analýza výhod a nevýhod popsaných postupů.*

***Klíčová slova:** P-hodnota; Testování ekvivalence; Nulová hypotéza; Testování hypotéz, TOST*

Meritem tohoto článku je teoretické představení statistických postupů, které umožňují vyjadřovat podporu nulovým hypotézám. Ačkoliv je znalost a chápání logiky testování hypotéz důležitou součástí vědomostí psychologa, ne vždy je tento postup užíván a interpretován korektně. Mnohé relevantní statistické údaje nejsou reportovány a interpretovány přesně (Hoekstra a kol., 2014), nebo dokonce nejsou reportovány vůbec (Fritz a kol., 2013). Mechanické užívání statistických postupů společně s jejich nepřesnou interpretací mnohdy nabádá výzkumníky k neadekvátním a chybným závěrům. Typickým příkladem je interpretace *p*-hodnoty přesahující stanovenou hladinu významnosti (α) jako podporu pro nulovou hypotézu (Goodman, 2008; Lakens, 2017). Tato nepřesná interpretace se v různých (většinou méně explicitních) podobách objevuje i v učebnicích (Cassidy a kol., 2019) a ve velké míře i v odborných článcích v prestižních vědeckých časopisech (Aczel a kol., 2018).

[1] Psychologický ústav, Filozofická fakulta, Masarykova univerzita, Arne Nováka 1, 602 00 Brno, Česká republika

Korespondenční autor: David Lacko, e-mail: david.lacko@mail.muni.cz

Doručeno do redakce 4. 8. 2020

S přihlédnutím k výše zmíněnému je na místě uvést, že psychologové mají k dispozici metody, s jejichž pomocí lze testovat a případně vyjádřit podporu pro nulovou hypotézu a které mají potenciál v mnoha ohledech obohatit nejen psychologický výzkum. V tomto článku podrobně představujeme soubor metod, který lze souhrnně nazvat jako testování ekvivalence. Testování ekvivalence nespočívá v ničem jiném než v ověřování toho, zda je efekt tak malý, že jej lze z praktického hlediska považovat za zanedbatelný (Lakens a kol., 2018b; Linde a kol., 2020; Meyners, 2012). Tento postup je vhodný hned v celé řadě situací. Kupříkladu když chce výzkumník dokázat, že méně nákladná intervence dosahuje stejného výsledku jako intervence dražší, nebo když usiluje o nalezení, identifikování a následné falzifikování takových aspektů teorie, které se nedaří potvrdit, kvůli následnému redefinování teorie (Linde a kol., 2020).

Mezi jedny z nejčastěji užívaných přístupů testování ekvivalence lze zařadit dva jednostranné testy, *p*-hodnotu druhé generace, oblast praktické ekvivalence a Bayesův faktor. Tyto níže představené postupy mimo jiné umožňují výzkumníkům překonávat některé nedostatky tradičního testování hypotéz a vyjadřovat se k důležitým, byť stále opomíjeným statisticky nevýznamným zjištěním. Pro pochopení testování ekvivalence je však potřeba nejprve připomenout principy klasického testování hypotéz a význam *p*-hodnoty.

KLASICKÉ TESTOVÁNÍ HYPOTÉZ

Tradiční postupy pro testování hypotéz rozlišují nulovou (představující absenci efektu nebo rozdílu) a alternativní (obvykle výzkumníkem predikovanou) hypotézu, přičemž výzkumníci zpravidla usilují o vyvrácení nulové hypotézy, což interpretují jako důkaz hypotézy alternativní. Tento přístup se nazývá Null Hypothesis Significance Testing (NHST; pro přehled např. Christensen, 2005; Nickerson, 2000; Perezgonzalez, 2015) a vznikl kombinací metodologického paradigmatu Fischerova testování signifikance založeného na snaze zamítnout nulovou hypotézu a Neyman-Pearsonově testování akceptace založeného na principu přijetí alternativní hypotézy. Přestože je NHST z mnoha důvodů často kritizováno, zůstává v psychologii obecně nejpoužívanějším přístupem (Nickerson, 2000).

Výzkumníci operující v rámci tohoto paradigmatu usilují o zamítnutí nulových hypotéz (tzv. *reject*), čímž získávají podporu pro své alternativní hypotézy. Tato situace nastává v případě, že je *p*-hodnota menší než stanovená hladina významnosti (tzn. $p < \alpha$), což indikuje nesoulad mezi daty a statistickým modelem (např. s nulovou hypotézou). V některých případech se výzkumníkovi nulovou hypotézu zamítnout nepodaří (tzv. *fail to reject*), neboť *p*-hodnota je naopak větší než stanovená hladina významnosti (tzn. $p > \alpha$). Tato situace poukazuje na to, že pozorovaná data nejsou za předpokladu platnosti nulové hypotézy nijak překvapivá, a nelze proto činit závěry ani o alternativní, ani o nulové hypotéze (tzn. data jsou neprůkazná; Wasserstein & Lazar, 2016).

Situace statisticky nevýznamného rozdílu však bývá, jak již bylo řečeno, často nepřesně interpretována jako důkaz neexistence zkoumaného efektu (Gagnier & Morgenstern, 2017). K takovému závěru výzkumník ale nemůže dojít, neboť v tomto konkrétním případě prizmatem tradičního NHST platí, že absence důkazu není důkazem absence efektu. Podobné interpretace jsou proto nekorektní, zkreslující a *de facto* chybné (pro přehled miskoncepcí viz Gagnier & Morgenstern, 2017; Goodman, 2008; Greenland a kol., 2016; Nickerson, 2000), neboť výsledek nemusel být způsoben výhradně reálnou absencí zkoumaného efektu, ale např. nízkou silou testu nebo nevhodným výběrem zkoumaného vzorku.

P-HODNOTA

Jak již bylo naznačeno, fundamentálním aspektem tradičního testování hypotéz je *p*-hodnota a její interpretace. Ačkoliv byla *p*-hodnota do vědy poprvé uvedena už na začátku 20. století a stala se ústředním pojmem frekvencionistického přístupu (tj. klasické statistiky), její využívání, chápání a interpretování stále nebývá vždy korektní. I proto vydala Americká Statistická Asociace (ASA) stanovisko k adekvátnímu využívání a správnému interpretování *p*-hodnoty, kterou zároveň definují jako "*pravděpodobnost, že data v rámci specifického statistického modelu (např. rozdíl mezi výběrovými průměry dvou skupin) budou stejná, nebo extrémnější než její pozorovaná hodnota, platí-li daný statistický model*" (Wasserstein & Lazar, 2016, str. 131). Součástí stanoviska je i šest principů, dle kterých *p*-hodnota:

1) může naznačovat rozsáhlost nekompatibility dat se specifickým statistickým modelem (nejčastěji nulovou hypotézou);
2) neměří pravděpodobnost, že daná hypotéza je pravdivá, ani že byla získána pouze náhodou;
3) nižší, než stanovená hladina významnosti by neměla nikdy sloužit jako jediné kritérium pro tvorbu vědeckých závěrů;
4) musí být vždy reportována transparentně a v plném rozsahu (ve snaze vyhnout se cherry-pickingu a *p*-hackingu[2] způsobeným selektivním reportováním výsledků);
5) neměří velikost efektu ani důležitost výsledku;
6) sama o sobě bez uvedení do kontextu neposkytuje dostatečné důkazy týkající se modelu či hypotézy (Wasserstein & Lazar, 2016).

[2] *Cherry-picking* představuje vybrání nejlepšího a nejpříznivějšího výsledku (většinou právě na základě *p*-hodnoty) z několika provedených statistických postupů, které všechny testují stejnou hypotézu (Murphy & Aguinis, 2019). *P-hacking* spočívá v řadě technik, jejichž smyslem je přinést statisticky významný výsledek bez ohledu na nejadekvátnější možný postup. Výzkumník tedy např. testuje hypotézu a pokud je výsledek nesignifikantní, tak pokračuje dále ve sběru dat, rekóduje používané proměnné, zkouší postupně vyřazovat outliery nebo zahrnuje do analýzy analýz další proměnné, nebo je z ní vyřazuje (Head a kol., 2015). K těmto neetickým a metodologicky neadekvátní postupů lze zařadit také např. *data dredging* (tzn. provedení velkého množství statistických testů bez korekce pro mnohonásobné testování, a následné selektivní reportování pouze žádoucích výsledků; Mayo, 2020), *HARKing* (*post hoc* tvorba hypotéz až poté, co známe výsledky), *Sharking* (*post hoc* tvorba teorie až poté, co známe výsledky; Hollenbeck & Wright, 2017), a mnohé další nekorektní postupy.

Zkoumaný efekt se považuje za statisticky signifikantní, pokud je nalezená *p*-hodnota menší než stanovená hladina významnosti. Obecně je v psychologické výzkumné praxi nejčastěji využívána hladina významnosti 0,05 (Goodman, 2008). V rámci Fisherova paradigmatu je sice možné signifikanci *p*-hodnoty interpretovat až *a posteriori* po analýze dat, v rámci Neyman-Pearsonova paradigmatu stejně jako v dominujícím NHST je nutné hranici významnosti stanovit *a priori* před analýzou dat. Toto nastavení se však nemusí striktně a mechanicky držet běžné hodnoty hladiny významnosti, ale může být nastaveno variabilně, podle cílů studie, předchozích výzkumů i teoretického ukotvení (Christensen, 2005; Nickerson, 2000; Perezgonzalez, 2015).

Zmíněná hodnota hladiny významnosti 0,05 je však kvůli mnoha důvodům často kritizována. Mnozí výzkumníci ji považují za příliš benevolentní a navrhují všeobecné snížení hladiny významnosti na $\alpha = 0{,}005$, což by mělo zlepšit replikovatelnost a možnosti generalizace výsledků (např. Benjamin a kol., 2018; Ruiter, 2019). Jiní naopak doporučují zcela upustit od rozšířeného termínu "statistické signifikance" stejně jako od snahy ustanovit všeobecné kritérium pro její hladinu. Místo toho navrhují stanovovat tuto hladinu individuálně, vždy podle specifických okolností daného výzkumu (např. Lakens a kol., 2018a; Miller & Ulrich, 2019). Kupříkladu ve výzkumech užívajících velké výzkumné vzorky nebo v situacích, kdy falešně pozitivní výsledek (tzn. chyba 1. typu) může poškodit participanta, je vhodné užívat nižší hladinu významnosti. A naopak u menších vzorků nebo třeba u explorativních výzkumů je možné užít benevolentnější hladinu významnosti. Další odborníci pak *p*-hodnotu přímo nahrazují jinými statistickými indikátory (pro přehled Wasserstein a kol., 2019) nebo se odklánějí od tradičního NHST směrem k bayesovské statistice (Wagenmakers a kol., 2018). Někteří autoři rozšiřují redukcionistický a dichotomický přístup testování hypotéz o "novou statistiku," která je založena výhradně na odhadech hodnot parametrů, intervalů spolehlivosti a velikostí efektů (Cumming, 2012).

Ačkoliv *p*-hodnota čelí dlouhodobě kritice, přičemž tato kritika v poslední době nabývá na intenzitě, její podstatná část pramení z nepochopení významu *p*-hodnoty a z jejího nekorektního užívání spíše než z její podstaty (Greenland, 2019). Navíc je nepravděpodobné, že by se v blízké budoucnosti od užívání *p*-hodnot upustilo (Ruiter, 2019), a proto považujeme její pochopení, adekvátní užívání a interpretování ve vědeckém výzkumu za zásadní.

KLASICKÉ TESTY EKVIVALENCE

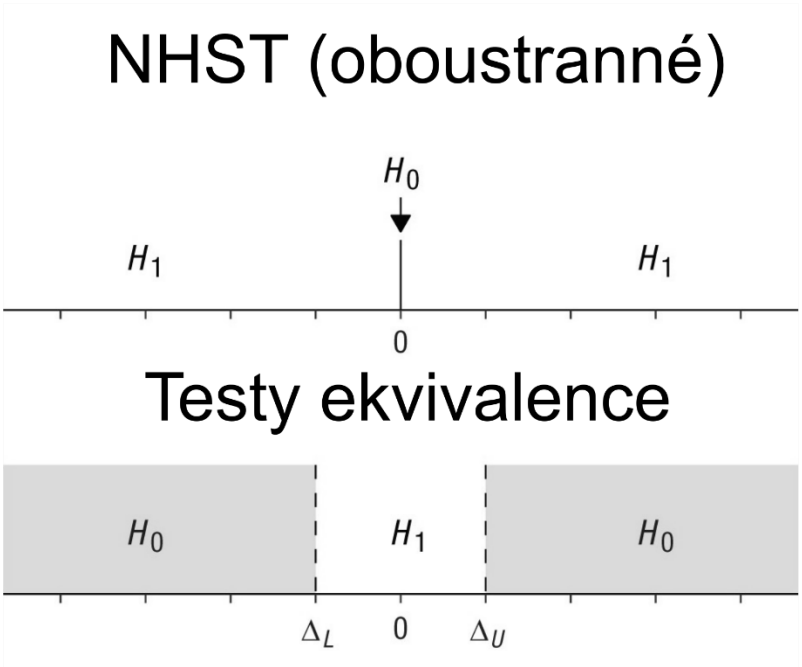
Navzdory tomu, že tradiční NHST neumožňuje dokazovat absenci efektu, jsou navrženy statistické metody, které jsou schopny doložit důkazy pro platnost nulové hypotézy. V rámci klasického přístupu ke statistice se jako možnost nabízí využít klasické testy ekvivalence (Lakens, 2017; Walker & Nowacki, 2011), které zahrnují řadu procedur, jimiž lze podpořit nulovou hypotézu či absenci prakticky významného efektu (tzn. příliš malého efektu, který je v praxi z hlediska teorie zanedbatelný; Lakens a kol., 2018b; Linde a kol., 2020; Meyners, 2012).

Klasické testy ekvivalence zahrnují více postupů, mezi které lze zařadit např. "power" přístup, metodu Andersenové a Haucka, metodu Daniela a Johna Ennisových založenou na otevřených symetrických intervalech, nebo třeba postup LEAD (Least Equivalent Allowable Difference; pro přehled Meyners, 2012). Nejfrekventovanějším z nich je metoda dvou jednostranných testů (Two One Sided Tests, TOST), která byla poprvé představena před více než třiceti lety (Schuirmann, 1987). Ačkoliv se TOST v porovnání s některými jinými klasickými testy ekvivalence vyznačuje nižší silou, je lehce srozumitelný i interpretovatelný, není tak technicky náročný a ve většině případů je podobně efektivní (Meyners, 2012). Navzdory mnohým dřívějším (např. Rogers a kol., 1993; Stegner a kol., 1996) i nedávným (např. Lakens, 2017) snahám o představení TOST v kontextu psychologie se tomuto postupu však stále nedostává dostatečné pozornosti.

V jádru tohoto přístupu leží nutnost stanovit největší možný efekt, který ještě lze považovat za prakticky ekvivalentní nulovému efektu, resp. nejmenší možný efekt, který již je prakticky významný. Tím vznikne rozsah ekvivalence (viz Obr. 1), přičemž krajní (horní/upper = Δ_U a spodní/lower = Δ_L) body tohoto rozpětí představují nejnižší (pozitivní i negativní) hodnoty efektu stojícího v zájmu výzkumníka (tzv. Smallest Effect Size of Interest, SESOI; Lakens, 2017). SESOI je nutné definovat před zahájením výzkumu, díky čemuž se zvedá potenciální informativnost studie (Lakens a kol., 2018b), a to buď pomocí hrubých skórů, nebo prostřednictvím standardizované velikosti efektu (Lakens, 2017). Výzkumník posléze může hodnoty uvnitř tohoto rozpětí považovat za prakticky nevýznamné a tedy srovnatelné (ekvivalentní) s nulovou hypotézou (Lakens a kol., 2018c).

Obrázek 1

Logika NHST a TOST (upraveno podle Lakens a kol., 2018c)



Poznámka. V rámci oboustranného NHST usilujeme o zamítnutí (rejekci) nulové hypotézy (H0) signalizující přítomnost nulový efekt. V rámci TOST usilujeme o zamítnutí jednak nulové hypotézy (H0) definované jako efekt vyšší než spodní hranice (ΔL) ekvivalenčního rozpětí a zároveň o zamítnutí druhé nulové hypotézy (H0) definované efekt nižší než horní hranice (ΔU) ekvivalenčního rozpětí.

Důkazy o ekvivalenci lze získat inspekcí intervalu spolehlivosti, jehož rozpětí je z důvodu provedení dvou testů stanoveno jako $1 - 2 \times \alpha$ (tradiční α = 0,05 tedy odpovídá 90% úrovni spolehlivosti; Rogers a kol., 1993; Lakens, 2017). Pokud je získaný interval spolehlivosti zcela obsažen ve stanoveném rozpětí ekvivalence, můžeme deklarovat statistickou ekvivalenci. Ekvivalenci můžeme také posoudit na základě výsledků dvou jednostranných *t*-testů, pomocí kterých srovnáváme odhadnuté hodnoty parametrů s krajními hodnotami ekvivalenčního rozpětí. Pomocí prvního testu se zjišťuje, zda lze zamítnout hodnoty vyšší, než je horní hranice ekvivalenčního rozsahu, zatímco druhý test usiluje o zamítnutí hodnot menších, než je spodní hranice ekvivalenčního rozpětí (Rogers a kol., 1993). Vyjdou-li oba testy statisticky signifikantně stran zvolené hladiny významnosti, lze dospět k závěru, že daná hodnota parametru (např. průměrný rozdíl mezi dvěma skupinami) spadá do ekvivalenčního rozpětí, a je prakticky srovnatelná s nulovým efektem (Lakens, 2017; Schuirmann, 1987), což ovšem *a priori* neznamená, že efekt je nulový či absentuje (Campbell & Gustafson 2018; Harms & Lakens, 2018). Jelikož se provádí dva jednostranné testy a oba musí vyjít statisticky významně, aby mohla být podpořena ekvivalence, není třeba korigovat hladinu významnosti pro mnohonásobné testování (Meyners, 2012).

TVORBA SESOI

Kruciálním a nejspíš i nejnáročnějším prvkem TOST je apriorní nastavení rozsahu SESOI. Jelikož je tento proces principiálně společný pro všechny postupy testování ekvivalence (byť s unikátní terminologií a drobnými rozdíly), které článek popisuje, rozhodli jsme se v samostatné podkapitole detailně rozebrat několik možných postupů stanovení SESOI. Podobně jako i u jiných statistických postupů je nejméně vhodným způsobem užití obecných orientačních vodítek bez návaznosti na teorii, jakou jsou například obecné klasifikace (tzv. rules of thumb) velikosti efektu na malý, střední a velký (Correll a kol., 2020; Lakens, 2017; Schäfer, & Schwarz, 2019). Naopak informačně přínosnějším způsobem je stanovení SESOI na základě předchozích studií, kdy lze spodní a horní hranici ekvivalenčního rozpětí odvodit na základě nejmenší hodnoty velikosti efektu, která by byla v originální výzkumu odhadnuta jako stále statisticky signifikantní. I tento způsob však má svá negativa, protože velikosti efektů zjištěné v předchozích studiích mohou být výrazně zkreslené díky tzv. publikačnímu zkreslení (v literatuře se totiž objevují spíše nadhodnocené velikosti efektu; viz např. Correll a kol., 2020; Schäfer, & Schwarz, 2019; Scheel a kol., 2020).

Doporučeným způsobem je stanovení SESOI na základě použité teorie, jejích postulátech a toho, co teorie predikuje, případně na základě praxe a zkušenosti (Lakens a kol., 2018c). Tento doporučovaný způsob však není vždy možné aplikovat, např. z důvodu limitace

výzkumného vzorku nebo neexistujícího či nejednoznačného teoretického pozadí. V takovém případě je možné využít pro stanovení SESOI velikost efektu získaného pomocí analýzy síly testu (power analysis), u které se nastaví realistický odhad vzorku a přiměřená hladina významnosti a síla testu (tj. $1 - \beta$, běžně 80 %). Tuto velikost efektu pak lze využít jako SESOI (Lakens, 2017). Simonsohn (2015) argumentuje, že studie, jejichž statistická síle se nachází pod hladinou 33 % ($\beta < 0{,}33$), mají nedostatečnou statistickou sílu (jsou tzv. underpowered). Jako referenční bod proto používá právě hodnotu velikosti efektu, která by v originální studii odpovídala statistické síle o hodnotě 33 % a která dle něj představuje nejmenší smysluplný efekt.

Další možné stanovení SESOI spočívá v kvantifikaci nejmenšího rozdílu, který je schopen participant individuálně a subjektivně zaznamenat (Minimal Detectable Differences, MDD). K odhadu MDD lze využívat tzv. Global Rating of Change (GRoC; Anvari & Lakens, 2019). Tato metoda spočívá v opakované administraci vybrané metody, přičemž druhé měření je doplněno o položku zjišťující změnu (tzv. GRoC položka), jakou participant pocítil v úrovni měřené proměnné (nejčastěji na pětibodové škále, např. velká pozitivní změna – malá pozitivní změna – beze změny – malá negativní změna – velká negativní změna). Hodnotu MDD lze odvozovat od průměrného rozdílu mezi prvním a druhým měřením těch participantů, kteří uvedli, že zaznamenali malou pozitivní nebo negativní změnu. Odpovědi ostatních participantů nejsou pro stanovení MDD v tomto případě podstatné, protože cílem je zjistit pouze nejmenší zaznamenatelný rozdíl (MDD).

Statistickou analýzu TOST lze provádět v programu *R* s balíčkem "TOSTER" (Lakens, 2018) či v programu *jamovi* s modulem "TOSTER" (Lakens, 2017). Oba balíčky umožňují testovat ekvivalenci z výsledků jednovýběrového *t*-testu, *t*-testu pro nezávislé výběry či korelace, a také odhadovat adekvátní velikosti vzorku vzhledem k žádoucí síle testu a stanovenému ekvivalenčnímu rozsahu. Výpočet TOST pomocí programu *R* je demonstrován v příloze ke článku.

P-HODNOTA DRUHÉ GENERACE

K podpoře nulové, ale také alternativní hypotézy lze využít nově představenou *p*-hodnotu druhé generace (Second Generation *P*-Value, SGPV; Blume a kol., 2018). I v rámci tohoto přístupu je nutné stanovit ekvivalenční rozpětí, které koresponduje s nulovou hypotézou (Lakens & Delacre, 2020).

Vzhledem ke své deskriptivní povaze udává SGPV poměr překrytí ekvivalenčního rozpětí a rozsahu 95% intervalu spolehlivosti odhadovaného parametru (Blume a kol., 2018; Blume a kol., 2019). 95% Interval spolehlivosti je založen na výzkumné tradici a výzkumník si jej i zde může definovat sám. Pokud interval spolehlivosti zcela spadá mezi krajní hodnoty ekvivalenčního rozpětí, nabývá SGPV hodnotu 1, která vyjadřuje podporu dat pro nulovou hypotézu. Nachází-li se ovšem interval spolehlivosti zcela mimo ekvivalenční hodnoty, poté se hodnota SGPV rovná 0, což reflektuje podporu dat pro hypotézu alternativní. Hodnoty SGPV mezi těmito krajními body ($0 < p < 1$) vypovídají o

nejednoznačném důkazu dat pro nulovou či alternativní hypotézu, a nelze tak učinit závěr v žádném směru. Tato hodnota *de facto* ukazuje, kolik % intervalu spolehlivosti spadá do stanoveného ekvivalenčního rozpětí. SGPV stejně jako TOST nevyžaduje, aby ekvivalenční rozpětí a 95% interval spolehlivosti odhadovaného parametru (resp. SESOI v případě TOST) dodržoval symetrickou strukturu (lze tedy zvolit např. –0,2 až +0,4 místo symetrického –0,4 až +0,4; Blume a kol., 2018; Lakens, 2017). SGPV lze vypočítat pomocí následujícího vzorce:

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times max\left\{\frac{|I|}{2|H_0|}, 1\right\}$$

kde *I* prezentuje interval hodnoty parametru podporovaný daty, zatímco H_0 pak interval nulové hypotézy. $I \cap H_0$ znamená průnik obou intervalů. Druhá část vzorce pak reprezentuje korekci pro případy, kdy je interval spolehlivosti odhadnutého parametru alespoň dvakrát širší než SESOI a zároveň se aspoň částečně se SESOI překrývá (Blume a kol., 2018). Děje se tak zejména v situacích, kdy je ekvivalenční interval příliš úzce definován či v případech, v nichž výzkumník disponuje malým výzkumným vzorkem (Lakens & DeLacre, 2020). δ představuje polovinu délky intervalu nulové hypotézy. A právě tento ukazatel lze využít při srovnání dvou studií, které měly hodnotu $p_\delta = 0$. Vzdálenost mezi intervalem nulové hypotézy a intervalem získaného parametru v jednotkách delty reprezentuje podle Blume a kol. (2018) tzv. "delta gap". Pokud je hodnota delta gap první studie vyšší než druhé studie, signalizuje tato skutečnost silnější statistický důkaz o přítomnosti efektu.

Kromě podpory nulové hypotézy umožňuje SGPV také kontrolovat chybu 1. typu. Autoři dále upozorňují, že *p*-hodnotu není třeba upravovat při provádění mnohonásobných testů, neboť, jak již bylo zmíněno, SGPV je deskriptivní ani nikoliv inferenční povahy (Blume a kol., 2018; Blume a kol., 2019). SGPV je schopno navíc pracovat také s malým výzkumným souborem i s úzkým rozpětím intervalu ekvivalence (jak již vyplývá z představeného vzorce), neboť koriguje svou hodnotu ve chvíli, kdy se oba intervaly překrývají a zároveň je interval spolehlivosti získaného parametru dvakrát širší, než je rozsah hodnot signalizujících ekvivalenci (Blume a kol., 2018). V těchto krajních případech odpovídá hodnota SGPV vždy 0,5, což naznačuje nejednoznačnost podpory pro jakoukoliv ze dvou alternativ.

SGPV lze podobně jako TOST počítat v programu *R*, konkrétně v aktuálně vyvíjeném balíčku "sgpv" (Welty a kol., 2018).

PŘÍSTUPY V RÁMCI BAYESOVSKÉHO PARADIGMATU

Vzhledem k rostoucímu zájmu o Bayesovské paradigma a jeho aplikaci v psychologickém výzkumu (van de Schoot a kol., 2017) budou v následujících odstavcích stručně představeny dvě metody – oblast praktické ekvivalence a Bayesův faktor – pomocí nichž lze taktéž vyjádřit podporu pro nulovou hypotézu a demonstrovat absenci praktického

efektu. Analogicky k úvodním kapitolám o testování hypotéz a *p*-hodnotě je pro jejich pochopení však nutné nejprve stručně přiblížit hlavní principy bayesovské statistiky (více detailněji např. Wagenmakers a kol., 2018; Kruschke & Lidell, 2018a; 2018b; Lambert, 2018).

V jádru bayesovské statistiky stojí tři stěžejní složky: apriorní rozdělení, věrohodnostní funkce (likelihood) a aposteriorní rozdělení (van de Schoot a kol., 2014). Apriorní rozdělení umožňuje specifikovat a vyjádřit znalosti výzkumníka o parametru jeho zájmu, kterými disponuje před zahájením sběru dat či před analýzou. Toto vyjádření znalostí je stanoveno v podobě pravděpodobnostního rozložení (Wagenmakers a kol., 2018). Apriorní rozdělení lze vyjádřit pomocí celé řady distribucí, obecně jej však můžeme zařadit do tří kategorií – a to podle množství informací, které se v něm odráží: neinformativní, lehce informativní a informativní (van de Schoot & Deapoli, 2014). Apriorní rozdělení odráží výzkumníkův stupeň důvěry (degree of belief) ohledně hodnot parametru jeho zájmu před analýzou dat (Harms & Lakens, 2018). Podstata apriorního rozdělení tedy spočívá v přidělení pravděpodobnosti všem možným hodnotám parametru tak, aby výsledné aposteriorní pravděpodobnostní rozdělení odpovídalo předběžným znalostem výzkumníka, resp. apriornímu přesvědčení o relativní plauzibilitě různých hodnot parametru (Kruschke & Lidell, 2018a; 2018b).

Poté, co je přiřazeno apriorní rozdělení každému parametru v modelu, sehrává důležitou roli druhá složka bayesovské statistiky, kterou jsou získaná data (van de Schoot a kol., 2014). Tato komponenta je pak vyjádřena pomocí věrohodnostní funkce (likelihood) vyjadřující pravděpodobnost získaných dat vzhledem k daným parametrům užitého modelu (Lambert, 2018). Prostřednictvím bayesova teorému dochází k aktualizaci apriorního rozdělení daty, kdy pod vlivem těchto nových informací dochází k přesunu kredibility mezi hodnotami a vzniká tak výsledné aposteriorní rozdělení (Kruschke & Lidell, 2018a; van de Schoot a kol., 2014; Wagenmaker a kol., 2018).

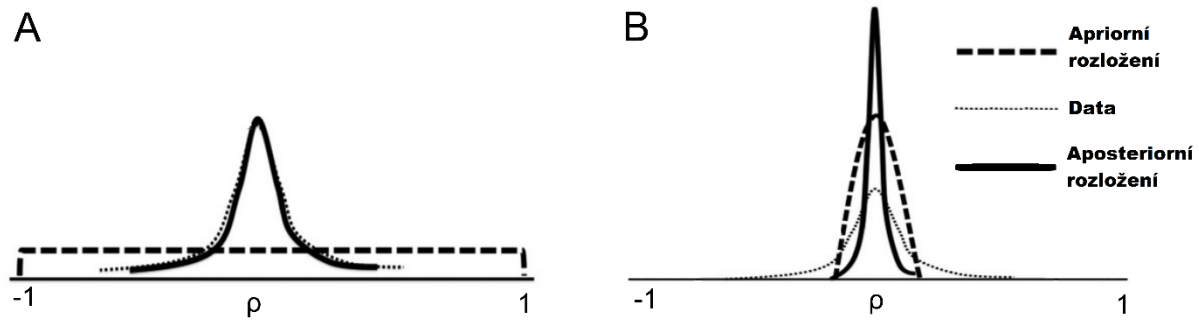
Aposteriorní rozdělení, třetí složka bayesovské statistiky, reprezentuje výzkumníkovu nejistotu či stupeň důvěry ohledně hodnot zkoumaného parametru po zahrnutí dat a apriorního rozdělení (Harms & Lakens, 2018; Wagenmaker a kol., 2018). Aposteriorní distribuci lze charakterizovat např. prostřednictvím HDI (High Density Interval; Kruschke, 2018). V souladu s reportováním intervalů spolehlivosti se obecně uvádí hodnota 95 % HDI, která obsahuje 95 procent nejpravděpodobnějších hodnot, kterých může zkoumaný parametr nabývat. Širší HDI pak odráží znatelnější nejistotu ohledně odhadu hodnoty zkoumaného parametru oproti užšímu 95% HDI (Kruschke & Lidell, 2018b).

Vztah všech tří komponent je demonstrován na Obr. 2. Na příkladu *A* můžeme vidět, že v případě použití rovnoměrného apriorního rozdělení, kdy všechny hodnoty Pearsonova korelačního koeficientu mají stejnou kredibilitu, hrají zásadní roli při tvorbě aposteriorního rozdělení data. Příklad *B* pak zachycuje situaci, v níž volba informativního prioru hrubě odpovídá informacím z dat, což následně vede k užšímu aposteriornímu

rozdělení, a tedy k redukci nejistoty výzkumníka ohledně potenciálních hodnot parametru po zahrnutí dat.

Obrázek 2

Apriorní rozdělení, věrohodnostní funkce (data) a aposteriorní rozdělení (upraveno podle van de Schoot a kol., 2014)



Poznámka. Plná čára představuje aposteriorní rozdělení, přerušovaná čára představuje apriorní rozdělení a tečkovaná čára představuje věrohodnostní funkci (tedy naměřená data).

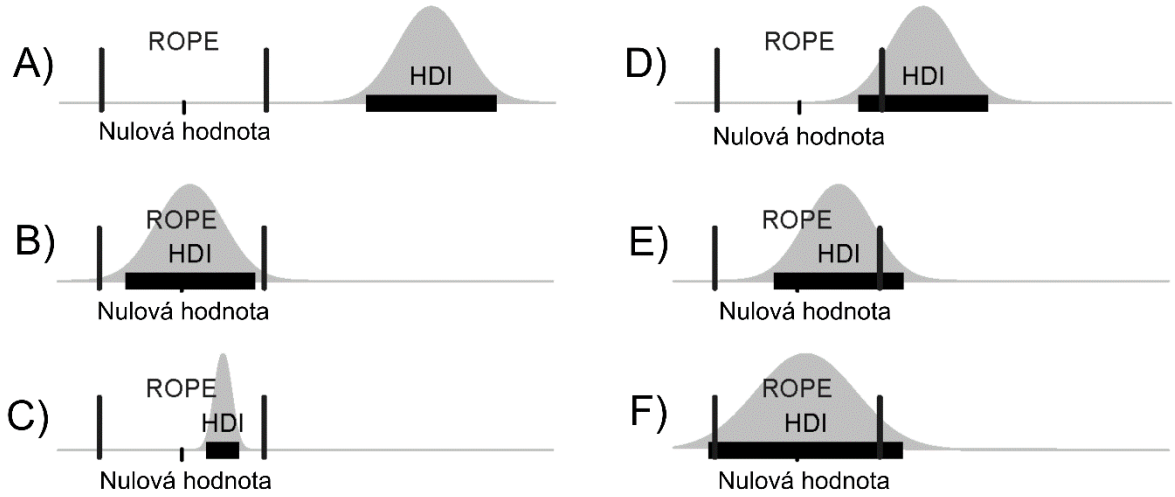
OBLAST PRAKTICKÉ EKVIVALENCE

Prvním popsaným postupem, kterým lze v rámci bayesovského odhadu hodnot parametrů vyjádřit podporu pro nulovou hypotézu, je tzv. oblast praktické ekvivalence (Region Of Practical Equivalence, ROPE). ROPE si lze představit jako ekvivalent k TOST interpretovaný v rámci bayesovského paradigmatu. ROPE lze připodobnit k ekvivalenčnímu rozsahu v tom smyslu, že specifikuje hodnoty, které lze považovat z praktického hlediska za nulové (Kruschke, 2018). Krajní body ROPE je nezbytné stanovit v souladu s konkrétní teorií a tím, co tato teorie predikuje (tedy podobně jako u TOST). Volba krajních bodů by proto opět neměla být založena na pomocných obecných vodítkách (např. slabá, střední či velká velikost efektu; Kruschke & Lidell, 2018a).

Užitím metody ROPE v kombinaci s 95% HDI lze dospět ke třem závěrům (Kruschke, 2011; 2018). *1)* Pokud se 95% HDI nachází zcela mimo vymezenou oblast praktické ekvivalence, zamítá se nulový efekt (viz obr. 3A). *2)* Nachází-li se však 95% HDI mezi vytyčeným rozsahem ROPE, pak můžeme tyto hodnoty označit za prakticky ekvivalentní (Kruschke, 2018, str. 272; viz obr. 3B a 3C). *3)* V situacích, kdy část 95% HDI spadá do ROPE a část se nachází mimo vymezené hranice ROPE, nelze učinit jasný závěr (viz Obr. 3D, 3E a 3F). Statistická analýza je možná taktéž v programu *R* s balíčkem "bayestestR" (Makowski a kol., 2019).

Obrázek 3

Možné výsledky ROPE (upraveno podle Kruschke, 2018)



Poznámka. 95% HDI je zobrazeno tučnou vodorovnou linií, nulový efekt je zobrazen tenkou krátkou svislou linií a oblast praktické ekvivalence je stanovena dvěma svislými delšími liniemi. Obr. A zobrazuje 95% HDI nacházející se zcela mimo oblast praktické ekvivalence, což umožňuje zamítnout zanedbatelný efekt. Obr. B zobrazuje 95% HDI zcela uvnitř ekvivalenčního rozpětí, proto se nulový efekt naopak přijímá. Obr. C zobrazuje 95% HDI, který sice neobsahuje nulu, ale nachází se zcela v oblasti ekvivalence, což taktéž umožňuje označit hodnotu parametru za prakticky ekvivalentní nulovému efektu. Obr. D zobrazuje situaci, kdy sice 95% HDI neobsahuje nulu, avšak část tohoto intervalu spadá do ROPE, a proto nelze učinit jednoznačné rozhodnutí. Obr. E zobrazuje 95 % HDI, který obsahuje nulový efekt a zároveň se část tohoto intervalu nachází mimo ROPE, a proto opět nelze činit jednoznačné rozhodnutí. Obr. F demonstruje případ, v němž se oblast ROPE překrývá s 95% HDI, ale jeho konce navíc zasahují i mimo oblast praktické ekvivalence ROPE. Ani v tomto případě proto nelze dospět k definitivnímu verdiktu.

BAYESŮV FAKTOR

Výše popsaná metoda ROPE slouží k přijímání a odmítání specifikovaných hodnot parametru (v našem případě nulového efektu), avšak (podobně jako SGPV) není procedurou k testování hypotéz. K testování hypotéz, respektive k relativnímu srovnání dvou soupeřících modelů (hypotéz), se v rámci bayesovského přístupu používá tzv. Bayesův faktor (BF), jenž představuje "*formu statistické inference, ve které je jeden model, např. alternativní hypotéza, postaven proti druhému modelu, např. nulové hypotéze*" (Dienes, 2016, str. 78). Testování hypotéz pomocí BF vyžaduje jak specifikaci nulové hypotézy, tak i stanovení apriorního rozložení alternativní hypotézy (Kruschke & Liddell, 2018a). V nejobecnějším případě je při stanovení nulové hypotézy veškerá kredibilita přiřazena nulovému efektu (či hodnotě signalizující absenci efektu, rozdílu atd.; Kruschke & Liddell, 2018a), nicméně lze stanovit také intervalovou nulovou hypotézu (Morey & Rouder, 2011). Roli BF si můžeme ukázat na zjednodušené formě Bayesova teorému (Wagenmakers a kol., 2018):

$$\frac{p(H1|data)}{p(H0|data)} = \frac{p(H1)}{p(H0)} \times \frac{p(data|H1)}{p(data|H0)}$$

Vztah "$\frac{p(H1)}{p(H0)}$" vyjadřuje tzv. apriorní šanci neboli "*relativní pravděpodobnost alternativní (H_1) oproti nulové (H_0) hypotéze před zvážením dat*", zatímco vztah "$\frac{p(H1|data)}{p(H0|data)}$" reprezentuje aposteriorní šanci, která "*kvantifikuje relativní pravděpodobnosti H_1 oproti H_0 po zahrnutí dat*" (Ly, 2017, str. 15). Z uvedené rovnice pak vyplývá, že BF představuje "*poměr pravděpodobností dat podmíněných dvěma hypotézami, které jsou srovnávány*" (Morey a kol., 2016, str. 9). Jak z rovnice vyplývá, BF $\frac{p(data|H1)}{p(data|H0)}$ je tvořen poměrem pravděpodobností dat vzhledem k jednotlivým modelům či hypotézám (Kruschke, 2014). BF ve své podstatě říká, kolikanásobně vyšší je (aposteriorní) šance alternativní hypotézy po zahrnutí dat ve srovnání s původní (apriorní) šancí (Kass & Raftery, 1995; Wagenmakers a kol., 2018). Je-li apriorní šance rovna 1 (stává se tak v případě, kdy před analýzou dat oběma teoriím či modelům přisuzujeme stejnou míru pravděpodobnosti) hodnota BF odpovídá hodnotě aposteriorní šance (Kass & Raftery, 1995).

Samotná hodnota BF pak udává pravděpodobnost dat pod jednou hypotézou oproti druhé, čili kolikanásobně více data podporují jednu hypotézu oproti druhé (Dienes, 2014). Většinou se označuje jako BF_{10} nebo BF_{01} podle toho, jestli srovnáváme alternativní (H_1) s nulovou (H_0) hypotézou a vice versa. Např. výrok "$BF_{10} = 4$" svědčí o tom, že data jsou čtyřikrát pravděpodobnější pod hypotézou alternativní než pod hypotézou nulovou. Bayesův faktor tak podává relativní důkaz vyplývající z dat pro jednu hypotézu vůči druhé (Rouder & Morey, 2011). BF může nabývat hodnot od nuly do nekonečna (Dienes, 2014). Na základě hodnot BF můžeme činit rozhodnutí ve třech směrech (Dienes, 2014). Hodnota BF_{10} vyšší než jedna vyjadřuje podporu pro alternativní hypotézu, hodnota nižší než jedna pro nulovou a hodnoty BF kolem 1 pak naznačují, že data nejsou dostatečně průkazná, aby podporovala jednu, či druhou hypotézu (Dienes, 2011).

Přestože BF je spojitou veličinou, pro snadnější interpretaci ji lze kategorizovat (Lee & Wagenmakers, 2013; Dienes, 2016). Hodnoty BF, které jsou vyšší než 3, lze orientačně interpretovat jako opodstatněný důkaz o podpoře pro alternativní hypotézu. Hodnoty BF nižší než ⅓, lze naopak vnímat jako podporu pro nulovou hypotézu. Hodnoty BF v rozmezí 20-150 signalizují silný důkaz pro jednu z hypotéz a hodnota BF od 150 výše pak reflektují velmi silný důkaz ve prospěch jedné z hypotéz (Kass & Raftery, 1995). I když takových interpretačních schémat pro hodnoty BF existuje celá řada, podobně jako u *p*-hodnot je i u nich kritizováno mechanické a nekritické užívání předem stanovené hranice „významnosti" (např. Simonsohn, 2019). V konečném důsledku by se proto na otázku týkající se toho, co je podpora pro nulovou či alternativní hypotézu, mělo odpovídat na základě vlastních a obhajitelných norem (viz kapitola věnující se SESOI). Zároveň je třeba upozornit, že hodnota BF podobně jako *p*-hodnota nikterak nevypovídá o velikosti efektu (Wagenmakers a kol., 2018).

Bayesovskou statistiku dnes umí mnohé softwary, statistickou analýzu lze provádět např. v programu JAGS, WINBUGS, STAN, JASP, *jamovi*, IBM SPSS Statistics nebo v programu *R*, kde existuje velké množství obecných i specializovaných balíčků.

SROVNÁNÍ A KRITIKA JEDNOTLIVÝCH METOD

Ačkoliv se popsané metody částečně překrývají stran nejen jejich účelu, ale i metodologických a statistických argumentů, v určitých aspektech se odlišují, a proto u nich lze pozorovat rozdílné výhody a nevýhody. Jelikož není v možnostech tohoto článku detailně srovnat jednotlivé postupy, budou zde představeny pouze stručně zásadní rozdílnosti a podobnosti.

Všechny metody sdílí, jak už ostatně bylo zmíněno v kapitole věnující se SESOI, nutnost apriorního stanovení si rozsahu praktické ekvivalence. Ačkoliv tento postup vysoce zvyšuje informativní přínos studie, samotné stanovení je náročné a často nerealizovatelné; a bohužel ani sami autoři metod neposkytují dostatečné množství komplexních a složitějších příkladů a návodů, jak si rozsah praktické ekvivalence vlastně stanovit. Lze sice považovat za vysoce žádoucí skutečnost, že výzkumník musí specifikovat rozpětí efektu, který nemá žádný praktický význam, ještě před sběrem dat, v běžné výzkumné praxi však bývá stanovení takového rozpětí problematické a komplikované. Jak Simonsohn (2015) dodává, psychologické teorie mají převážně kvalitativní charakter, který dovoluje spíše stanovit pouze směr vztahu mezi proměnnými, než určit velikosti nejmenšího efektu s praktickým a teoretickým významem. Nejspíše i proto autoři velmi často popisují doporučené stanovení SESOI na základě teorie velice vágně, čímž nicméně nechávají výzkumníka docela napospas.

Dalším nedostatkem, který metody sdílejí, je nutnost velkého výzkumného vzorku, a to zejména v případech, kdy je ekvivalenční rozpětí úzké (tzn. SESOI je blízké nule; Simonsohn, 2015; Linde a kol., 2020). Pokud totiž výzkumník disponuje například vzorkem o velikosti 100 participantů či méně, což je docela běžná velikost v experimentálním psychologickém výzkumu, metody TOST a ROPE vykazují minimální diskriminační schopnost, zatímco BF takové situace zvládá relativně dobře (Linde a kol., 2020).

Co se týče srovnání TOST a SGPV, Lakens a Delacre (2020) shrnují několik zásadních rozdílů. Zatímco SGPV je deskriptivní statistika, *p*-hodnota v rámci TOST je svou povahou inferenční. Další odlišnost pak spočívá ve skutečnosti, že u SGPV představuje rozsah ekvivalence nulovou hypotézu, zatímco při provádění TOST pod nulovou hypotézu spadají hodnoty, které se naopak nachází mimo toto rozpětí. Obě zmíněné metody navíc nevyžadují, aby ekvivalenční rozpětí a 95% interval spolehlivosti odhadovaného parametru dodržoval symetrickou strukturu (tzn., že lze zvolit např. –0,2 až +0,4 místo symetrického –0,4 až +0,4; Blume a kol., 2018; Lakens, 2017). Dále lze ještě zmínit, že na základě SGPV lze učinit tři závěry: *a)* podpořit nulovou hypotézu, *b)* podpořit alternativní hypotézu a *c)* označit data za neprůkazná; zatímco u TOST v kombinaci s NHST čtyři: *a)*

efekt je statisticky signifikantní a není ekvivalentní, *b)* výsledek je statisticky signifikantní a spadá do rozsahu ekvivalence, *c)* efekt není statisticky signifikantní, ale je ekvivalentní a *d)* data jsou neprůkazná.

Na tomto místě je potřeba podotknout, že TOST i SGVP a vlastně i ROPE spolu také sdílejí jeden potencionální nedostatek. Ačkoliv autoři metod často kritizují mechanické nastavování hladin významnosti, sami ve svých příkladech paradoxně užívají arbitrární hodnoty pro hladiny významnosti a intervaly spolehlivosti (resp. HDI v případě ROPE) bez hlubšího odůvodnění, což může vést k tomu, že se tyto hodnoty začnou užívat mechanicky i při testování ekvivalence. Na druhou stranu i v těchto případech mohou výzkumníci samozřejmě tyto hodnoty při plánování výzkumu upravovat dle kritérií zmíněných výše.

Ačkoliv by se mohlo zdát, že BF poskytuje více výhod ve srovnání s TOST a SGP, ani on se nevyhnul kritice. Kritizována je skutečnost, že finální hodnota BF se může výrazně měnit v závislosti na volbě apriorního rozdělení (Kruschke & Liddell, 2018a), zatímco v případě odhadů hodnot parametrů vliv apriorního rozdělení obecně klesá s rostoucím množstvím dat (Harms & Lakens, 2018). Poukazováno je i na nadužívání přednastavených apriorních rozdělení, které přirozeně nemohou být adekvátní pro všechny teorie (Kruschke & Liddell, 2018a), i další mechanické užívání bayesovských procedur (Lakens, 2021). Kupříkladu ve 33 % publikovaných článků, ve kterých byla data analyzována v rámci bayesovského přístupu, nebylo specifikováno apriorní rozdělení a bylo pouze použito výchozí nastavení v daném statistickém softwaru (van de Schoot a kol., 2017). Je však potřeba zdůraznit, že mechanické užívání je typické taktéž pro NHST a lze jej paradoxně identifikovat, jak bylo demonstrováno výše, i v ostatních postupech.

Další diferenci lze spatřit mezi ROPE a TOST. Nejenom že se liší ve filozofii statistické inference, ze které vychází, ale ROPE v kombinaci s bayesovským odhadem parametrů navíc prostřednictvím aposteriorní distribuce umožňují posoudit kredibilitu různých hodnot parametru a stanovit interval, v němž se stanovenou pravděpodobnostní nachází hodnota odhadovaného parametru. TOST naopak umožňuje činit jen a pouze dichotomická rozhodnutí (tzn. ekvivalence ne/dosažena; viz Harms & Lakens, 2018).

Nelze nezmínit také skutečnost, že frekventistické metody TOST a SGPV nejsou výpočetně tak náročné jako bayesovské metody ROPE a BF. Na druhou stranu, ROPE i BF jsou implementovány do celé řady statistických softwarů, které lze označit za uživatelsky přívětivé, i TOST je dostupný mimo prostředí *R* také (např. v *jamovi*), zatímco SGPV lze aktuálně provádět pouze v programu *R*.

Kterou metodu by si měl výzkumník tedy zvolit, pokud chce testovat ekvivalenci? Ačkoliv se BF ukazuje jako spolehlivější v případě menších výzkumných vzorků (Linde a kol., 2020), preference konkrétní metody v konečném důsledku závisí na samotném výzkumníkovi, neboť za běžných situací dosahují představené metody podobných výsledků (Lakens a kol., 2018b), přičemž výzkumníkům nic nebrání v reportování více procedur.

ZÁVĚR

Zejména v posledních letech se často mluví o tom, že psychologii sužuje replikační krize, projevující se v neschopnosti replikovat statisticky významné výsledky původních studií. Tato „krize důvěry" má velké množství příčin a její dopady pro psychologii jako takovou mohou být dalekosáhlé (Pashler & Wagenmakers, 2012). Ačkoliv se neúspěšným replikacím a dalším negativním zjištěním obecně dostává ve vědě stále více prostoru,[3] a to zejména s ohledem na praktiky otevřené vědy a pre-registrované výzkumné designy, je tato oblast stále relativně opomíjená (Schimmack, 2020).

Jednou z cest, jak lze replikační krizi překonat, je právě i užívání výše představených postupů za účelem získávání důkazů pro podporu nulových hypotéz (Anderson & Maxwell, 2016; Amrhein a kol., 2019; Colling & Szűcs, 2018; Maxwell a kol., 2015; Schimmack, 2020). Psychologický výzkum by se totiž neměl zaměřovat pouze na hledání existence efektu, ale měl by také zkoumat jeho absenci a s ní spjaté hodnoty, které z praktického hlediska nemají žádný význam. Statistická signifikance by měla být obohacena o signifikanci praktickou. A proto je přínosné zahrnout do výzkumných studií jednu či více metod sloužících k podpoře nulové hypotézy. Někteří autoři dokonce začínají mluvit o dvoustupňovém, resp. podmíněném testování ekvivalence (tzv. Conditional Equivalence Testing, CET; Campbell & Gustafson, 2018), což je přístup, podle kterého by se měly frekventistické metody testování ekvivalence aplikovat automaticky pokaždé, když se nepodaří zamítnout nulovou hypotézu v rámci NHST.

Zmíněné metody testování ekvivalence tedy adresují a do jisté míry řeší nejen častou a nepřesnou interpretaci nesignifikantního výsledku coby důkazu absence efektu, ale také umožňují výzkumníkům lépe pochopit a interpretovat tolik důležitá negativní zjištění. Znalost a korektní užívání výše představeného testování ekvivalence má proto potenciál zvýšit validitu psychologických výzkumů a chápání důsledků z nich plynoucích. Závěrem lze říci, že ačkoliv aplikace těchto metod s sebou přináší celou řadu překážek, jejich benefity rozhodně převažují nad potenciálními negativy.

Poděkování

Článek byl podpořen Filozofickou fakultou Masarykovy univerzity v rámci projektu specifického výzkumu MUNI/A/1323/2020. Rádi bychom poděkovali PhDr. Martinovi Jelínkovi, Ph.D. a Mgr. Vojtěchovi Juříkovi, Ph.D. za podnětné komentáře a připomínky k textu. Dále bychom chtěli poděkovat recenzentům za jejich přínosné poznámky, které napomohly udělat text srozumitelnější a přesnější.

[3] Dnes se negativní zjištěním věnují pravidelné speciální vydání významných světových časopisů, některé mezioborové žurnály studie s nesignifikantními výsledky zveřejňují zcela běžně již několik let (např. *PLoS ONE*, *Frontiers in Psychology*) a vznikají dokonce časopisy, které negativní zjištění vítají nebo se na ně zcela zaměřují (v psychologii např. *Journal of Articles in Support of the Null Hypothesis* nebo *Meta-Psychology*).

REFERENCE

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., ... , & Wagenmakers, E. J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation. *Advances in Methods and Practices in Psychological Science, 1*(3), 257-366. https://doi.org/10.1177/2515245918773742

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*(1), 1–12. https://doi.org/10.1037/met0000051

Anvari, F., & Lakens, D. (2019, February 1). Using Anchor-Based Methods to Determine the Smallest Effect Size of Interest. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/syp5a

Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential Statistics as Descriptive Statistics: There Is No Replication Crisis if We Don't Expect Replication. *The American Statistician, 73*(sup1), 262–270. https://doi.rog/10.1080/00031305.2018.1543137

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E. J., Berk, R., ... Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6-10. https://doi.org/10.1038/s41562-017-0189-z

Blume, J. D., D'Agostino McGowan, L., Dupont, W. D., & Greevy, R. A. (2018). Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLOS ONE, 13*(3), e0188299. https://doi.org/10.1371/journal.pone.0188299

Blume, J. D., Greevy, R. A., Welty, V. F., Smith, J. R., & Dupont, W. D. (2019) An Introduction to Second-Generation p-Values. *The American Statistician, 73*(sup1), 157-167. https://doi.org/10.1080/00031305.2018.1537893

Campbell, H., & Gustafson, P. (2018). Conditional equivalence testing: An alternative remedy for publication bias. *PLoS ONE 13*(4), e0195145. https://doi.org/10.1371/journal.pone.0195145

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing Grade: 89% of Introduction-to-Psychology Textbooks That Define or Explain Statistical Significance Do So Incorrectly. *Advances in Methods and Practices in Psychological Science, 2*(3), 233–239. https://doi.org/10.1177/2515245919858072

Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *The American Statistician, 59*(2), 121-126. https://doi.org/10.1198/000313005X20871

Colling, L. J., & Szűcs, D. (2018). Statistical Inference and the Replication Crisis. *Review of Philosophy and Psychology*. https://doi.org/10.1007/s13164-018-0421-4

Correll, J., Mellinger, Ch., McClelland, G. H., & Judd, Ch. M. (2020). Avoid Cohen's 'Small', 'Medium', and 'Large' for Power Analysis. *Trends in Cognitive Sciences*, https://doi.org/10.1016/j.tics.2019.12.009

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

Demidenko, E. (2016). The p-Value You Can't Buy. *The American Statistician, 70*(1), 33-38. https://doi.org/10.1080/00031305.2015.1069760

Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science, 63*(3), 274-290. https://doi.org/10.1177/1745691611406920

Dienes, Z. (2014). Using Bayes to get the most out of non-significant results. *Frontiers in Psychology, 5*, 781. https://doi.org/10.3389/fpsyg.2014.00781

Dienes, Z. (2016). How Bayes factors change scientific practice. *Journal of Mathematical Psychology, 72*, 78-89. https://doi.org/10.1016/j.jmp.2015.10.003

Fritz, A., Scherndl, T., & Kuhberger, A. (2013). A comprehensive review of reporting practices in psychological journals: Are effect sizes really enough?. *Theory & Psychology, 23*(1), 98-122. https://doi.org/10.1177/0959354312436870

Gagnier, J. J, & Morgenstern, H. (2017). Misconception, misuses, and misinterpretation of P values and significance testing. *Journal of Bone and Joint Surgery, 99*(18), 1598-1603. https://doi.org/10.2106/JBJS.16.01314

Goodman, S. N. (2008). A dirty dozen: Twelve P-value misconceptions. *Seminars in Hematology, 45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Greenland, S. (2019). Valid P-Values Behave Exactly as They Should: Some Misleading Criticisms of P-Values and Their Resolution With S-Values. *The American Statistician, 73*(1), 106-114. https://doi.org/10.1080/00031305.2018.1529625

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. https://doi.org/10.1007/s10654-016-0149-3

Harms, C., & Lakens, D. (2018). Making 'Null Effects' Informative: Statistical Techniques and Inferential Frameworks. *Journal of Clinical and Translational Research, 3*(2), 382–393.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology, 13*(3), e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review, 21*(5), 1157–1164. https://doi.org/10.3758/s13423-013-0572-3

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, Sharking, and Tharking: Making the case for post hoc analysis of scientific data [Editorial]. *Journal of Management, 43*(1), 5–18. https://doi.org/10.1177/0149206316679487

Kass, R., & Raftery, A. (1995) Bayes Factors. *Journal of the American Statistical Association, 90*(430), 773-795. http://dx.doi.org/10.2307/2291091

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299-312. https://doi.org/10.1177/1745691611406925

Kruschke, J. K. (2014). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Boston: Academic Press.

Kruschke, J. K. (2018). Rejecting or accepting parameter values in Bayesian estimation. *Advances in Methods and Practices in Psychological Science, 1*(2), 270-280. https://doi.org/10.1177/2515245918771304

Kruschke, J. K., & Liddell, T. M. (2018a). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review, 25*(1), 155–177. https://doi.org/10.3758/s13423-017-1272-1

Kruschke, J. K., & Liddell, T. M. (2018b). The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review, 25*(1), 178-206. https://doi.org/10.3758/s13423-016-1221-4

Lakens, D. (2017). Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological & Personality Science, 8*(4), 355–362. https://doi.org/10.1177/1948550617697177.

Lakens, D. (2018). *Two One-Sided Tests (TOST) Equivalence Testing*. R package version 0.3.4. https://cran.r-project.org/web/packages/TOSTER/

Lakens D. (2021). The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspectives on psychological science: a journal of the Association for Psychological Science*, 1745691620958012. Advance online publication. https://doi.org/10.1177/1745691620958012.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., ... Zwaan, R. A. (2018a). Justify your alpha. *Nature Human Behaviour, 2*(3), 168-171. https://doi.org/10.1038/s41562-018-0311-x

Lakens, D., & Delacre, M. (2020). Equivalence Testing and the Second Generation P-Value. *Meta-Psychology, 4*, 1-11. https://doi.org/10.15626/MP.2018.933

Lakens, D., McLatchie, N., Isager, P. M., Scheel, A. M., & Dienes, Z. (2018b). Improving Inferences about Null Effects with Bayes Factors and Equivalence Tests. *The Journals of Gerontology: Series B. Psychological Science and Social Sciences 75*(1): 45-57. https://doi.org/10.1093/geronb/gby065

Lakens, D., Scheel, A. M., & Isager, P. M. (2018c). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. London: SAGE publications.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Linde, M., Tendeiro, J., Selker, R., Wagenmakers, E., & van Ravenzwaaij, D. (2020, November 10). Decisions About Equivalence: A Comparison of TOST, HDI-ROPE, and the Bayes Factor. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/bh8vu

Ly, A. (2017). *Bayes Factors for Research Workers* (Doctoral dissertation). Retrieved from: https://hdl.handle.net/11245.1/e601b852-1b29-407b-a276-1ccd2a2ed37b

Makowski, D., Ben-Shachar M. S. & Lüdecke, D. (2019). *Understand and Describe Bayesian Models and Posterior Distributions using bayestestR*. R package version 0.2.5. https://cran.r-project.org/web/packages/bayestestR

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. *The American psychologist, 70*(6), 487–498. https://doi.org/10.1037/a0039400

Mayo, D. G. (2020). Significance Tests: Vitiated or Vindicated by the Replication Crisis in Psychology? *Review of Philosophy and Psychology.* https://doi.org/10.1007/s13164-020-00501-w

Meyners, M. (2012). Equivalence tests — A review. *Food Quality and Preference, 26*(2), 231-245. https://doi.org/10.1016/j.foodqual.2012.05.003

Miller, J., & Ulrich, R. (2019). The quest for an optimal alpha. *PLoS ONE, 14*(1): e0208631. https://doi.org/10.1371/journal.pone.0208631

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419. https://doi.org/10.1037/a0024377

Morey, R. D., Romeijn, J.-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology, 72*, 6–18. https://doi.org/10.1016/j.jmp.2015.11.001

Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology, 34*(1), 1–17. https://doi.org/10.1007/s10869-017-9524-7

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301. https://doi.org/10.1037/1082-989x.5.2.241

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? [Editorial]. *Perspectives on Psychological Science, 7*(6), 528–530. https://doi.org/10.1177/1745691612465253

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in psychology, 6,* Article ID 223. https://dx.doi.org/10.3389/fpsyg.2015.00223

Rogers, J. L., Howard, K. I., & Vessey, J. T. (1993). Using significance tests to evaluate equivalence between two experimental groups. *Psychological Bulletin, 113*(3), 553-565. https://doi.org/10.1037/0033-2909.113.3.553

Rouder, J. N., & Morey, R. D. (2011). A Bayes-factor meta-analysis of Bem's ESP claim. *Psychonomic Bulletin & Review, 18*(4), 682–689. https://doi.org/10.3758/s13423-011-0088-7

Ruiter, J. P. (2019). Redefine or justify? Comments on the alpha debate. *Psychonomic Bulletin & Review, 26*(2), 430-433. https://doi.org/10.3758/s13423-018-1523-9

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology, 10*. https://doi.org/10.3389/fpsyg.2019.00813

Scheel, A. M., Schijen, M., & Lakens, D. (2020, February 5). An excess of positive results: Comparing the standard Psychology literature with Registered Reports. *PsyArCiv Preprints.* https://doi.org/10.31234/osf.io/p6e9c

Schimmack, U. (2020). A meta-psychological perspective on the decade of replication failures in social psychology. *Canadian Psychology/Psychologie canadienne, 61*(4), 364–376. https://doi.org/10.1037/cap0000246

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics, 15*(6), 657–680. https://doi.org/10.1007/BF01068419

Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science, 26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Simonsohn, U. (2019). [78c] Bayes Factors in Ten Recent Psych Science Papers. *Data Colada*, http://datacolada.org/78c

Stegner, B. L., Bostrom, A. G., & Greenfield, T. K. (1996). Equivalence testing for use in psychosocial and services research: An introduction with examples. *Evaluation and Program Planning, 19*(3), 193-198. https://doi.org/10.1016/0149-7189(96)00011-0

van de Schoot, R., & Depaoli, S. (2014). Bayesian analyses: Where to start and what to report. *European Health Psychologist, 16*(2), 75–84.

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. https://doi.org/10.1111/cdev.12169

van de Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods, 22*(2), 217–239. https://doi.org/10.1037/met0000100

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., . . . Morey, R. D. (2018). Bayesian statistical inference for psychological science. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review, 25*(1), 35-57. https://doi.org/10.3758/s13423-017-1343-3

Walker, E., & Nowacki, A. S. (2011). Understanding equivalence and noninferiority testing. *Journal of general internal medicine, 26*(2), 192–196. https://dx.doi.org/10.1007/s11606-010-1513-8

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA Statement on p-Values: Context, Process, and Purpose. *The American Statistician, 70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05." *The American Statistician, 73*(Suppl. 1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Welty, V., Stewart, T., Greevy, R., D'Agostino McGowan, L., & Blume, J. (2018). *R package for calculating second-generation p-values and associated measures*. R package version 0.0.1. https://github.com/weltybiostat/sgpv

David Lacko, Tomáš Prošek (2021): Null hypothesis support and its misconception in psychology: A theoretical introduction to equivalence testing

***Abstract**: This theoretical paper introduces approaches that can be used to support a null hypothesis. We introduce the four most promising techniques of equivalence testing: two one-sided tests (TOST), second-generation p-value (SPGV), Bayes Factor (BF), and region of practical equivalence (ROPE). The following paper clarifies the logic of the null hypothesis testing and the p-value, critically revises the benefits and drawbacks of presented techniques and also provides practical examples of TOST procedures.*

***Keywords**: P-value; Equivalence testing; Null hypothesis; Hypothesis testing, TOST*

PŘÍLOHA KE ČLÁNKU: PRAKTICKÁ UKÁZKA TOST

Vzhledem ke skutečnosti, že klasická statistika zůstává i nadále dominujícím směrem pro analýzy dat, následující příklady proto prakticky demonstrují proceduru TOST. Data užitá v příkladech byla vygenerována a pro lepší názornost uměle zasazena do psychologického kontextu. Statistické analýzy byly spočítány v programu *R* (v3.6.2) pomocí package "TOSTER" (Lakens, 2018).

Vygenerované grafy záměrně nejsou překládány ani nijak upraveny, aby čtenáři poskytly stejný output, jaký získá při vlastních analýzách. Tučná linie v grafech vždy zobrazuje 90% *CI* užívaný v TOST a tenká světlejší linie zobrazuje 95% *CI* používaný v rámci NHST. Svislé linie zobrazují nulovou hodnotu a krajní hodnoty ekvivalenčního rozpětí. Osa X označuje rozdíl průměrů (př. 1, 2 a 4), nebo korelační koeficient (př. 3).

Př. 1: NHST nesignifikantní, TOST signifikantní

Cílem této hypotetické studie bylo zjistit, zdali se muži a ženy liší v míře prožívané naděje měřené metodou Perceived Hope Scale (PHS). Před analýzou dat jsme určili SESOI, a to na základě stanovení nejvyšší hodnoty velikosti efektu, která by již nebyla statisticky signifikantní v naší prováděné studii. SESOI se tak nacházel v rozmezí od -0,23 až 0,23. Zatímco muži (n = 270) v průměru dosáhli hodnoty 20,52 (SD = 5,42) na škále PHS, ženy (n = 242) v průměru skórovaly o něco výše 20,94 (SD = 240). T-test pro nezávislé soubory na 5% hladině významnosti neodhalil mezi skupinami statisticky signifikantní rozdíl, $t(508) = -0,89$, $p = 0,374$, $d = -0,07$.

Tento výsledek neumožňuje dělání interpretací o neexistenci rozdílu mezi pohlavími. Pro podporu tvrzení, že průměry obou skupin se v zásadě neliší, byl proto použit TOST. Oba jednostranné testy vyšly statisticky signifikantně. První test indikuje, že lze zamítnout hodnoty menší než -0,23, $t(508) = 1,70$, $p = 0,045$, a druhý jednostranný test dovoluje zamítnout hodnoty větší než 0,23, $t(508) = -3,48$, $p < 0,001$. Z Obrázku A je patrné, že 95% interval pro NHST zakreslený světlou linií v sobě zahrnuje nulu, a proto nebylo možné zamítnout nulovou hypotézu. Naopak TOST ukázal, že tučná linie signalizující 90% *CI*, které se používají při testování ekvivalence, se zcela nachází mezi

oběma hraničními body ekvivalenčního rozsahu. Díky tomu lze dospět k závěru, že ačkoliv testovaný efekt není statisticky odlišný od nulové hodnoty, je statisticky ekvivalentní s nulou a výsledky proto lze interpretovat tak, že muži a ženy se neliší v míře prožívané naděje.

--- Obrázek A ---

Př. 2: NHST signifikantní, TOST signifikantní

Další analýza ověřuje účinnost programu zaměřeného na rozvoj pozitivních emocí, kterému bylo vystaveno 146 respondentů. Na začátku programu dosahovali respondenti průměrného skóru v subškále pozitivních emocí dotazníku Positive and Negative Affect Schedule (PANAS) 32,14 (SD = 7,42) a po ukončení programu se průměrné skóre navýšilo na 34,17 (SD = 6,56). Provedený párový t-test naznačuje, že rozdíl je statisticky signifikantní, $t(145) = -3{,}49$, $p < 0{,}001$, $d = -0{,}29$. Výzkumník si dopředu stanovil, že investice do aplikace tohoto programu by se vyplatila ve chvíli, kdyby vedl ke změnám větším, než je hodnota efektu Cohenova d 0,57, ta totiž odpovídá minimálně detekovatelnému rozdílu pro pozitivní emoci zjištěnému pomocí global rating of change (Anvari & Lakens, 2019).

Výsledky z TOST procedury signalizují, že je možné zamítnout přítomnost velikosti efektů, které jsou extrémnější než oba konce ekvivalenčního rozpětí, $t(145) = 3{,}40$, $p < 0{,}001$ (spodní hranice) a $t(199) = -10{,}38$, $p < 0{,}001$ (horní hranice). Účinnost programu tak lze z praktického hlediska považovat za ekvivalentní nulové hypotéze. Tento závěr je zřetelný i z Obrázku B, kde 95% CI (tenká linie) neobsahuje nulovou hodnotu, díky čemuž vyšla p-hodnota menší než byla stanovená alfa. Inspekce 90% CI (tučná linie) poukazuje na fakt, že tento interval neobsahuje žádnou z krajních hodnot ekvivalenčního rozpětí a nachází se tak zcela uvnitř něj. Ačkoliv se tedy participantům zvedla po intervenci úroveň pozitivních emocí, a to dokonce statisticky významně, výsledný efekt lze považovat za prakticky zanedbatelný.

--- Obrázek B ---

Př. 3: NHST nesignifikantní, TOST nesignifikantní

V následujícím příkladu je aplikován TOST pro korelační analýzu mezi škálou prožívané naděje Perceived Hope Scale (PHS) a soucitem měřeným metodou Santa Clara Brief Compassion Scale (SCBCS), kterou jsme aplikovali v zájmu ověření diskriminační validity PHS. V souladu s návrhem Simonsohna (2015) jsme určili, že 33% síla v naší původní studii o 77 respondentech by tvořila ekvivalenční rozpětí r = -0,24 až r = 0,24. Současná data jsme sesbírali od 140 jedinců. Korelační analýza neodhalila statisticky signifikantní vztah mezi dispoziční nadějí a negativními emocemi, r = 0,15 [95% *CI:* -0,02; 0,32], p = 0,077.

Na předem určené 5% hladině významnosti tak na základě obdržené p-hodnoty nemůžeme zamítnout nulovou hypotézu. Za této situace lze podle NHST tvrdit, že data byla neprůkazná. Následná analýza prostřednictvím TOST procedury měla za úkol ověřit, jestli je výsledek ekvivalentní s nulovou hypotézou. TOST však vyšel také statisticky nesignifikantně, p = 0,137 pro horní hranici SESOI a p < 0,001 pro spodní hranici. Nepodařilo se nám získat důkaz, pomocí kterého bychom mohli dospět k závěru, že hodnota korelace spadá do ekvivalenčního rozpětí. V tomto případě nelze zamítnout hodnoty extrémnější, než jsou krajní body ekvivalenčního rozpětí a je možné pouze konstatovat, že v dané situaci nejsou data průkazná (viz Obr. C).

--- Obrázek C ---

Př. 4: NHST signifikantní, TOST nesignifikantní

Poslední možný výstup kombinace NHST a TOST je demonstrován pomocí jednovýběrového t-testu. Výzkumník předpokládal, že lidé navštěvující univerzitu třetího věku budou inteligentnější než průměrná populace s IQ 100. Ke stanovení ekvivalenčního

rozpětí jsme využili Wechslerovy klasifikace IQ. Spodní bod ekvivalenčního rozpětí ve hrubých jednotkách jsme stanovili na -10 a horní hranici na 9, což odpovídá rozpětí průměrné inteligence vyjádřené v hodnotách od 90 po 109. U studentů univerzity třetího věku jsme naměřili inteligenčním testem průměrnou hodnotu 107 (*SD* = 14,82). Rozdíl mezi testovanou populací a celorepublikovým průměrem byl statisticky významný, $t(99) = 4.72$, $p < 0{,}001$, $d = 0{,}47$.

Následná TOST analýza nevyšla statisticky signifikantně pro horní hranici, $t(99) = -1{,}35$, $p = 0{,}090$, nicméně hodnota naměřeného průměru se lišila od spodní hranice $t(99) = 11{,}47$, $p < 0{,}001$, což lze vidět také na Obrázku D, kde se 90% konfidenční interval nachází zcela mimo ekvivalenční rozpětí. Na základě kombinace NHST a TOST tak lze dospět k závěru, že zjištěný průměrný rozdíl je statisticky odlišný od nuly a zároveň není statisticky ekvivalentní s nulovou hodnotou.

--- Obrázek D ---

R syntax

Pro výpočet byl užit následující *R* syntax:

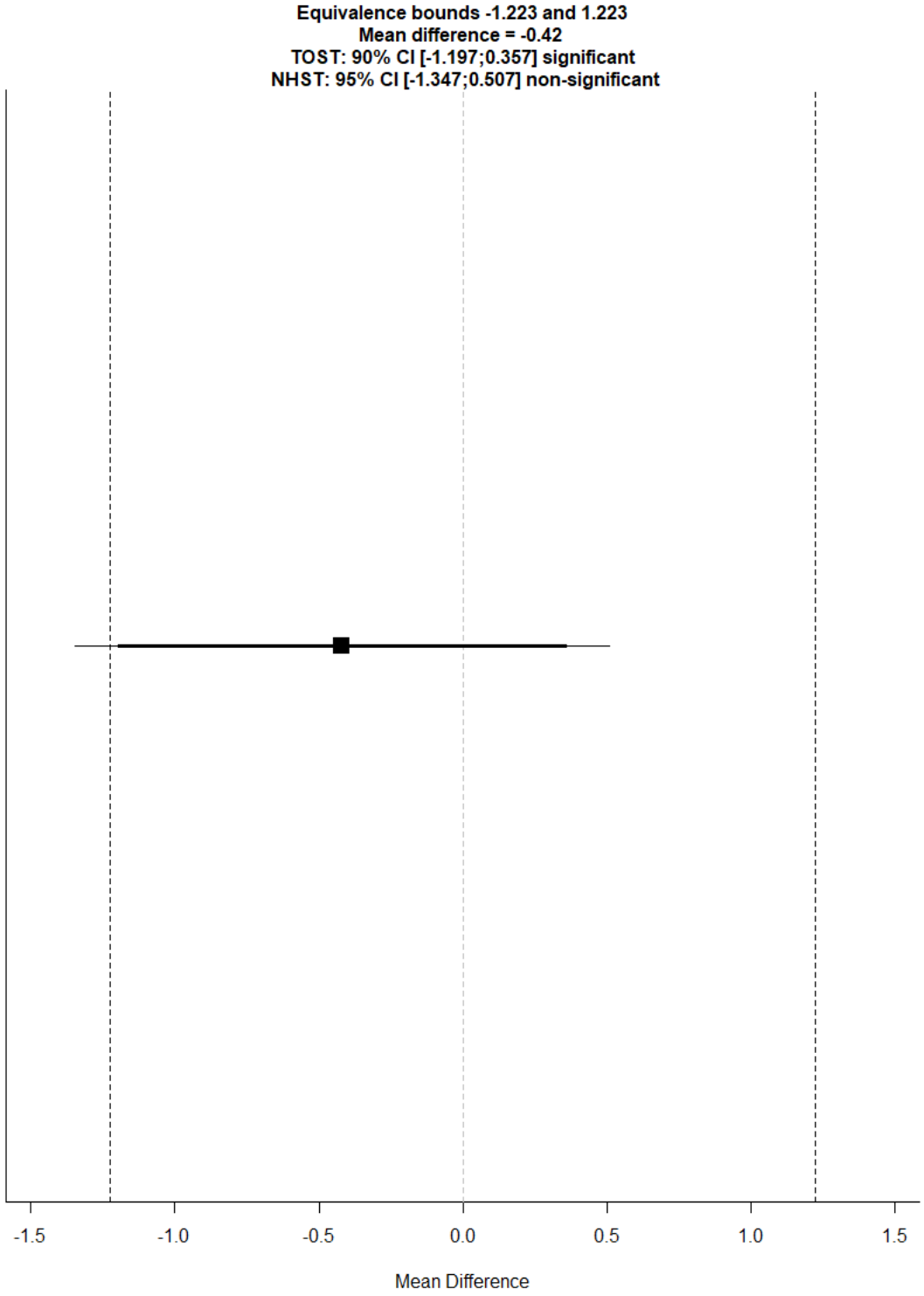
```
> # Instalace R balíčku
> install.packages("TOSTER")
> # Nahrání R balíčku
> library(TOSTER)
> # Pokud je porušen předpoklad homogenity rozptylu, lze nastavit var.equal na FALSE, což provede Welchův t-test
> # Příklad 1
> TOSTtwo(m1 = 20.52, m2 = 20.94, sd1 = 5.42, sd2 = 5.2,  n1 = 270, n2 = 240, low_eqbound_d = -0.23, high_eqbound_d = 0.23, alpha = 0.05, var.equal = TRUE, plot = TRUE, verbose = TRUE)
> # Příklad 2
> TOSTpaired(146, m1 = 32.14, m2 = 34.17 , sd1 = 7.42, sd2 = 6.56, r12 = 0.5, low_eqbound_dz = -0.57, high_eqbound_dz = 0.57, alpha = 0.05, plot = TRUE, verbose = TRUE)
> # Příklad 3
> powerTOSTr(alpha = 0.05, statistical_power = 0.33, N=77) # síla testu
```

> TOSTr(140, 0.15, low_eqbound_r = -0.24, high_eqbound_r = 0.24, alpha = 0.05, plot = TRUE, verbose = TRUE)

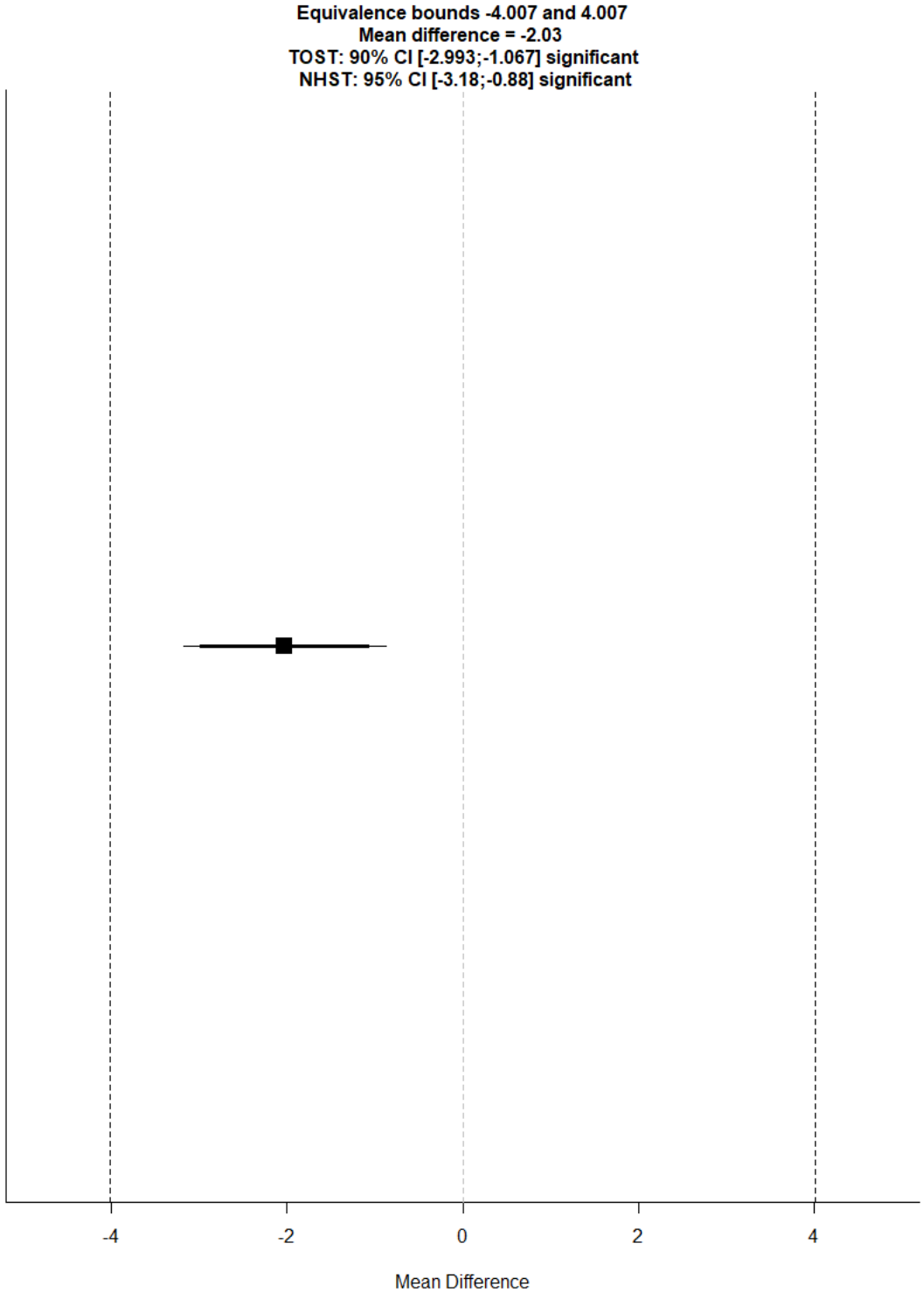
> # Příklad 4

> TOSTone.raw(107, 100, sd = 14.82, 100, low_eqbound = -10, high_eqbound = 9, alpha = 0.05, plot = TRUE, verbose = TRUE)

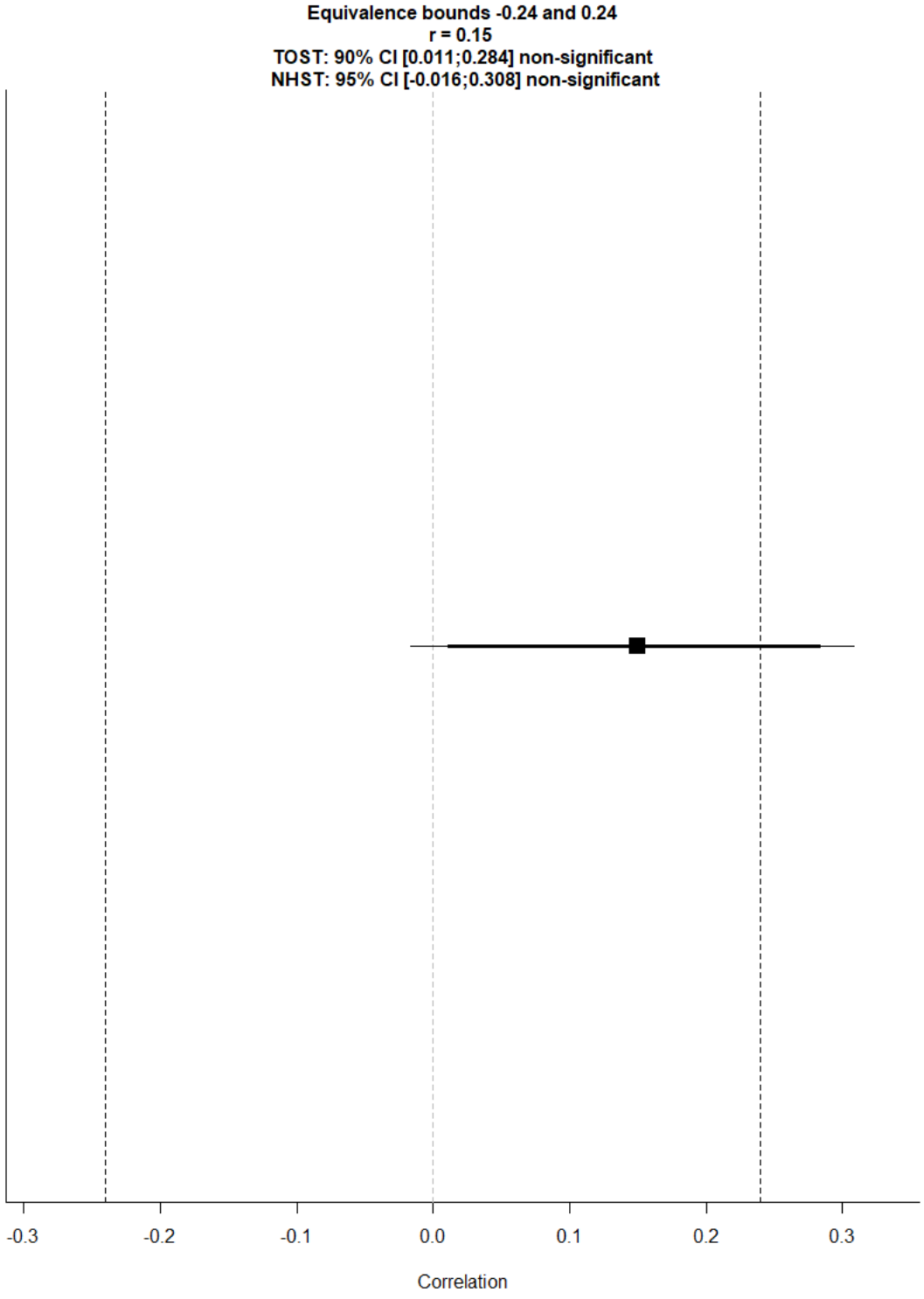
Obrázek A: NHST je nesignifikantní, ale TOST je signifikantní



Obrázek B: NHST i TOST jsou signifikantní



Obrázek C: Ani NHST ani TOST nejsou signifikantní



Obrázek D: NHST je signifikantní, ale TOST je nesignifikantní