P a d
(Používání statistické a věcné významnosti v českých sociálních vědách)*

PETR SOUKUP**
Fakulta sociálních věd, Univerzita Karlova, Praha

The Use of Statistical and Substantive Significance in the Czech Social Sciences

Abstract: The aim of the article is to quantify how often in leading Czech social-science journals (Československá psychologie / Czechoslovak Psychology, Pedagogika/Pedagogy, and Sociologický časopis / Czech Sociological Review) authors choose the wrong procedures to analyse quantitative data. In particular, attention is focused on the incorrect choice of statistical tests, their misinterpretation and mechanical application, and the use of effect sizes, that are so highly recommended nowadays. The basic research period was ten years, from 2005 to 2014, and for the Czech Sociological Review the period was extended back to 1995. The results of the content analysis of published articles (N=363) show that statistical tests are applied quite often to data that are not suitable for statistical tests: this is found in about one-fifth of cases in Czech Sociological Review, one-half in Pedagogy, and more than three-quarters in Czechoslovak Psychology. In addition, authors often make mechanical use of statistical methods or make incorrect interpretations (in over 40% of articles in the Czech Sociological Review over the last 10 years) and there are rarely any substantive interpretations of results (especially in Czechoslovak Psychology). Effect sizes are applied relatively often, but there are also gaps in their usage. It is clear from the results that changes are necessary both in the teaching of quantitative methodology and publishing practices in this subject area.

Keywords: statistical significance, substantive significance, effect sizes, content analysis

Sociologický časopis / Czech Sociological Review, 2019, Vol. 55, No. 2: 215–253
https://doi.org/10.13060/00380288.2019.55.2.459

* Vznik článku byl podpořen z prostředků Univerzity Karlovy PROGRES Q15 (Životní dráhy, životní styly a kvalita života z pohledu individuální adaptace a vztahu aktérů a institucí). Článek je mírně upravenou poslední částí dizertační práce autora.

** Veškerou korespondenci posílejte na adresu: PhDr. Ing. Petr Soukup, Ph.D., Institut sociologických studií, Fakulta sociálních věd Univerzity Karlovy, U Kříže 8, 158 00 Praha 5 – Jinonice, e-mail: petr.soukup@fsv.cuni.cz.

© Sociologický ústav AV ČR, v. v. i., Praha 2019

> *Sociologie by se možná posouvala vpřed mnohem rychleji, pokud bychom kladli větší důraz na používání věcných standardů a zkoumali, zda mají proměnné velký či malý vliv, namísto toho, abychom zkoumali, zda má koeficient obdržet jednu nebo dvě hvězdičky při publikování statistické významnosti.*
>
> [Bollen 1995]

Článek[1] navazuje na desítky zahraničních textů, které již více než tři desítky let poukazují na problematičnost při práci s daty z kvantitativních šetření (srov. stručný přehled diskuse v následující části). V Česku je tato diskuse spíše okrajová, v posledních patnácti letech byly publikovány texty Blahuše [2000], Soukupa a Rabušice [2007], Soukupa [2010], Cubereka a Frömela [2011] a Soukupa [2013]. Z výše uvedených textů je pouze jediný [Cuberek, Frömel, 2011] ryze empirickou analýzou časopisecké produkce, nicméně je poměrně úzce časově zaměřený (2008–2009) a i tematický záběr je poměrně úzký (zkoumá pouze, zda použitá data pocházela z náhodného výběru či randomizovaného experimentu).

Základní výzkumná otázka textu zní: *Jaká je praxe při používání statistické a věcné významnosti v českých sociálněvědních časopisech?*

S ohledem na šíři tématu navazují na tuto komplexní otázku tři dílčí otázky.

1. *Jak časté je nesprávné (nevhodné) využívání statistických testů? A je úroveň tohoto fenoménu v oblasti sociologie, pedagogiky a psychologie srovnatelná? Dochází v čase ke zlepšení nebo ke zhoršení?*
2. *Jak častá je nesprávná interpretace a mechanická práce s výsledky matematické statistiky? A je úroveň tohoto fenoménu v oblasti sociologie, pedagogiky a psychologie srovnatelná? Dochází v čase ke zlepšení nebo ke zhoršení?*
3. *Jak časté je využívání měr věcné významnosti, jejich interpretace a obecné věcné interpretace výsledků?*

U všech dílčích otázek je pozornost primárně zaměřená na produkci *Sociologického časopisu*,[2] nicméně pro zohlednění kontextu v ČR je provedeno též srovnání

[1] P a d v názvu článku odkazuje k tomu, že při statistickém vyhodnocení výsledků zpravidla hraje roli vypočtená pravděpodobnost chyby prvního druhu (softwarově často označovaná jako P), při věcném zhodnocení výsledků se pak užívá nejrůznějších měr věcné významnosti [Soukup 2013] a nejčastější z nich je Cohenovo d. Název článku se souřadným spojením snaží navodit představu, že pro správnou práci je třeba obojího (tj. zhodnocení statistické i věcné významnosti), a článek samotný se pak snaží ukázat, zda je tato představa v reálných textech publikovaných v ČR realizována.

[2] Pro zjednodušení užíváme v celém textu název *Sociologický časopis* s tím, že do analýzy byly zahrnuty obě jazykové varianty – před rokem 2002 šlo o dva samostatné časopisy, *Sociologický časopis* a *Czech Sociological Review*, které se v roce 2002 sloučily do jednoho časopisu *Sociologický časopis / Czech Sociological Review* a samostatný anglický časopis zanikl.

s oblastmi pedagogiky (časopis *Pedagogika*) a psychologie (*Československá psychologie*). Doplňkově je též monitorován vývoj všech tří fenoménů v čase, tj. je zjišťováno, zda dochází ke zlepšení nebo ke zhoršení.

V souladu s těmito třemi dílčími otázkami je členěna empirická část textu s tím, že srovnávací analýza se zaměřuje na články ve třech předních českých časopisech (Československá psychologie, Pedagogika a Sociologický časopis) v letech 2005–2014, detailnější vhled na produkci Sociologického časopisu pak zahrnuje dvacetileté období (1995–2014).

V příloze 1[3] k článku jsou pojednány další (návazné) otázky z oblasti sociologie vědy, které vyvstaly při základní analýze. Konkrétně je zkoumáno, jak jsou fenomény popsané třemi dílčími výše uvedenými otázkami ovlivněny institucionálním kontextem (změnami v systému financování vědy v ČR, případně publikačními doporučeními časopisů či odborných asociací). Dále je zkoumán vliv autorství, konkrétně je pozornost zaměřena na odlišnosti mezi juniorními a seniorními výzkumníky a na dopady vícečlenného autorského týmu ve srovnání s individuálním autorstvím.

Diskuse o statistické a věcné významnosti

Diskuse o statistické významnosti a omezení statistických testů je v zásadě tak stará jako testování samotné. Již ve dvacátých letech 20. století se objevují texty, které kritizují koncepci statistického testování. Výraznější posun zaznamenala tato diskuse zhruba v posledních třiceti letech, zejména díky knize Morrisona a Henkela [1970], která je souborem článků vydaných na konci padesátých let a v šedesátých letech sociology, psychology a statistiky upozorňujícími na problémy koncepce statistické významnosti a její nesprávné užívání. Lze konstatovat, že sociologické diskuse zejména v American Sociological Review byly v té době tím nejpodstatnějším. Diskuse se poté výrazně přenesla zejména na stránky psychologických a pedagogických časopisů a zhruba od druhé poloviny osmdesátých let minulého století pak začíná vznikat mnoho článků (a později též knih), které ukazují na omezení koncepce testování statistické významnosti (tuto diskusi zachycují v české produkci zejména články Blahuše [2000] a Soukupa [2010]). Není výjimkou, že autoři původních zahraničních textů navrhují, aby se statistické testy přestaly používat (doporučení opakovaně vyslovili zejména odborníci z oblasti metodologie psychologických a pedagogických výzkumů). Jako reprezentativní výběr textů lze v této oblasti využít knihu editovanou Lisou L. Harlow, Stanleym A. Mulaikem a Jamesem H. Steigerem [1997], která si klade provokativní otázku, co by se stalo, kdyby neexistovaly testy statistické významnosti. Kniha poměrně vyváženě podává názory zastánců a odpůrců testů. Výrazně jednostranně je naopak zaměřená sekce časopisu Psychological Science (1997, vol. 8) nazvaná „Zrušit

[3] Přílohy jsou dostupné online na https://doi.org/10.13060/00380288.2019.55.2.459.

testy statistické významnosti". Výsledkem těchto diskusí v psychologii a pedagogice bylo doporučení formulované v manuálech příslušných asociací APA [2010] a AERA [2006], které vyzývá autory, aby nepoužívali jen statistické testy (srov. popis dále) a aby se pokud možno vyhnuli mechanickým aplikacím statistických testů (např. užívání hvězdiček, zápisů $P < 0{,}05$ apod.). Souběžně s diskusí o zrušení statistického testování či alespoň o správné interpretaci výsledků testování se probouzí i diskuse o alternativách (nejčastěji se místo testů doporučuje užívat intervaly spolehlivosti, sílu testu či informační kritéria [srov. Soukup 2010]) a zejména pak diskuse užívání měr věcné významnosti (effect size) a nutnosti jejich interpretace. Základní přehledovou knihu v této oblasti napsal Kline [2004]. Kniha uvádí desítky těchto měr pro různé situace. V Česku lze čerpat z přehledového textu Soukupa [2013]. Praxe používání měr věcné významnosti a požadavek na jejich interpretaci byly ostatně zakotveny i do poslední verze manuálu Americké psychologické asociace [APA 2010: 33–34]. Nutno s lítostí dodat, že tato mezinárodní diskuse se do Česka dostává s výrazným zpožděním, pokud vůbec. Z poslední doby je nutné ještě zmínit stanovisko Americké statistické asociace k P-hodnotě [Wasserstein, Lazar 2016], formulované 26 předními světovými statistiky. Nutno konstatovat, že statistici stáli poměrně dlouhou dobu mimo diskusi o užívání statistických testů (sporadicky publikovali krátké texty upozorňující na chybné pochopení a chybné užívání). Nicméně poslední vývoj donutil reagovat prestižní Americkou statistickou asociaci k vydání poměrně ostrého stanoviska, které je plně v souladu s výše uvedenými závěry sociologů a psychologů. Stanovisko reagovalo zejména na dvě bezprostřední události: zákaz používání statistických testů v časopise Basic and Applied Social Psychology [Trafimow 2014; Trafimow, Marks 2015] a několik článků od statistiků, zejména pak text Reginy Nuzzo [2014], publikovaný v prestižním časopise Nature. Samotné stanovisko ASA [Wasserstein, Lazar 2016] pak poskytuje definici P-hodnoty (značně technicistní, byť bez uvádění vzorců) a dále šest doplňkových principů, které konstatují, že P-hodnota neměří pravděpodobnost pravdivosti námi zkoumaných hypotéz, neslouží k vyhodnocení věcné významnosti a pro vědecké závěry a ekonomická či politická rozhodnutí by nemělo být využito pouze skutečnosti, že P-hodnota dosahuje určité velikosti. Nádavkem stanovisko doporučuje, aby se namísto P-hodnoty užívalo alternativních postupů – zmíněny jsou zejména intervaly spolehlivosti, kredibilní a predikční intervaly, dále pak zejména bayesovská statistika a Bayesův faktor, respektive poměry věrohodností pro jednotlivé modely.

Pro úplnost závěrem dodejme, že diskuse o problematičnosti používání statistické významnosti a absence interpretace věcné významnosti výsledků se odehrála výrazněji i na poli ekonometrie – produktem této diskuse je slavná kniha Steva T. Ziliaka[4] a Deirdre N. McCloskey [2008].

[4] Pro úplnost doplňme, že Ziliak byl jedním z 26 statistiků, kteří se účastnili formulace stanoviska ASA [Wasserstein, Lazar 2016].

Obdobné studie v ČR a zahraničí

Na diskuse o problematičnosti statistických testů a požadavku na uvádění měr věcné významnosti navázaly i mnohé empirické studie, které se snaží ukázat, zda v jednotlivých časopisech dochází k nesprávným používáním statistických testů, jejich nesprávné interpretaci a zda autoři užívají míry věcné významnosti. Pro ilustraci zaměření a výsledků zde na počátku popíšeme dvě zřejmě neslavnější studie, jednu zaměřenou na American Sociological Review [Morrison, Henkel 1969] a další zaměřenou na American Economic Review [Zliak, McCloskey 2008]. Slouží spíše jako typická ukázka směřování těchto studií, velice orientačně mohou posloužit též pro pochopení výsledků prezentovaných v tomto článku. Dále jsou uvedeny studie, které je možné přímo používat pro srovnání s výsledky prezentovanými v článku.

Morrison a Henkel [1969] se zaměřili na všechny kvantitativně orientované články publikované v American Sociological Review v letech 1947–1967 a položili si otázku, jak často je v těchto textech užíváno koncepce statistické významnosti pro data, která nepochází z náhodných výběrů. Doplňkově se snažili zjistit, zda text Selvina [1957], který na tuto problematiku výrazně upozornil, měl vliv na výzkumnickou praxi. Na základě obsahové analýzy Morrison a Henkel odhadují, že podíl textů, které nesprávně užívají statistické testy (tj. pracují s nimi, i když data mají charakter nenáhodného výběru), se pohybuje okolo 40 % s tím, že po vydání textu Selvina nelze shledat žádné vylepšení. Studie Morrisona a Henkela je nejspíše první studií tohoto typu a nadto ukázkou monotematicky zaměřené studie, tj. zkoumá jen užívání statistické významnosti s ohledem na použitá výzkumná data. Může také sloužit jako pomyslný „benchmark" pro srovnání současných českých výsledků s výsledky nejlepší světové sociologie před cca padesáti lety.

Zliak a McCloskey ve své knize [2008] přetiskují mj. články, které se zaměřily na výzkum textů publikovaných v American Economic Review, nejdříve v letech 1980–1989 a posléze 1990–1999. Na rozdíl od výše popisovaného výzkumu Morrisona a Henkela hodnotili Zliak a McCloskey více (19) kritérií, mj. mechanickou práci s výsledky statistických testů (např. řazení koeficientů dle velikosti statistické významnosti), publikaci síly testu, slovní záměnu statistické a věcné významnosti a věcnou interpretaci výsledků [detaily viz Ziliak a McCloskey 2008: 67–73]. S ohledem na detailnost jejich zkoumání (i mírně jiné oborové zaměření) zde zmíníme jen některé výsledky. První obecný závěr Zliaka a McCloskey byl, že ekonometrická praxe se ve dvou analyzovaných desetiletích sice místy zlepšuje (například častěji je užíváno síly testu, rozlišováno mezi statistickou a ekonomickou významností), ale v mnohých ohledech dochází ke zhoršení (zejména v případě mechanických aplikací, tj. řazení koeficientů dle statistické významnosti či výběru proměnných do modelu jen dle statistické významnosti). Za pozornost stojí zjištění, že kvalita článků od více autorů je horší než u článků s jedním autorem (tento fenomén bude doplňkově zhodnocen na české produkci v příloze 1 článku).

Situaci v oblasti používání statistických testů v novější světové sociologii (v časopisech American Sociological Review a Journal of American Sociology) přehledně zmapoval Leahey [2005]. Na stratifikovaném výběru článků dvou předních časopisů z let 1935–2000 ukázal vývojové trendy. Podíl textů, kde jsou užívány statistické testy, 5% hladina statistické významnosti a posléze hvězdičky k označení statisticky významných výsledků má výrazně rostoucí tendenci (v roce 1995 bylo 90 % textů se statistickými testy, téměř 80 % užívalo 5% konvenci a více než 40 % užívalo hvězdičky). Leahey dále konstatuje, že 54 % z článků obsahujících statistické testy, používá tyto testy pro data, která nepocházejí z pravděpodobnostních výběrů a autoři těchto článků toto někdy sami přiznávají jako nedostatek textu.

Zcela nedávno byla publikována studie Fabrizia Bernarda, Lely Chakhaiy a Liliy Leopold [2017], která se zaměřuje na texty publikované v European Sociological Review (mezi lety 2000–2004, respektive 2010–2014) využívající regresní metodologii. Cílem textu je zmapovat užívání statistické významnosti v těchto textech (celkem 356 článků) za pomoci pěti z 19 kritérií použitých ve studii Ziliaka a McCloskey [2008]. Autoři dospěli k závěru, že v polovině analyzovaných textů je chybná interpretace statistické významnosti (typicky je statisticky nevýznamný výsledek regresního koeficientu považován za důkaz nulového vlivu příslušné nezávisle proměnné) a obdobně zhruba v polovině textů chybí věcná interpretace (autoři nastavili poměrně přísná kritéria pro tuto klasifikaci, srov. jejich kódovací schéma). Při srovnání dvou pětiletých období (2000–2004, respektive 2010–2014) dospívají autoři k závěru (obdobně viz výše studie Ziliaka a McCloskey), že v čase dochází ke zhoršování, tj. počet nesprávných užití statistické významnosti v čase narůstá. Jediné zlepšení nacházejí autoři studie u uvádění měr věcné významnosti, kde podíl 46 % (2000–2004) byl navýšen na 59 % (2010–2014). Zajímavostí je, že proporce článků využívajících regresní metodologii je v ESR opravdu vysoká (cca 80–90 % produkce). Bude ukázáno, že situace v ČR je výrazně odlišná.

Na okraj je vhodné také zmínit některé výzkumy, které se věnují statistické a věcné významnosti z jiných úhlů pohledu, než je ten prezentovaný v tomto článku. Za pozornost stojí práce Oakese [1986] a Hallera s Kraussem [2002], kteří stejným testem otestovali výzkumníky i studenty z oblasti psychologie a dospěli k závěru, že v obou skupinách panují většinově chybná přesvědčení o koncepci statistické významnosti. Na okraj lze poznamenat, že autor tohoto článku použil test vyvinutý Oakesem několikrát v českém prostředí a výsledky jsou srovnatelné (jejich publikace, respektive publikace po rozšíření výzkumu, bude námětem samostatného článku).

Další zajímavou studií byl výzkum členů AERA prezentovaný v článku Kethleen C. Mittag a Bruce Thompsona [2000]. Výzkumníci zaměření na oblast pedagogiky většinově souhlasí s návrhy zakotvenými do publikačních manuálů a jednotlivých článků, zejména že se má užívat intervalů spolehlivosti, plného sousloví „statisticky významné“. Nicméně zhruba polovina výzkumníků se po-

stavila za užívání mechanických postupů statistického testování reprezentovaného stepwise postupem výběru proměnných. Poslední zde uvedenou studií je článek DeVaneyho [2001], který je založen na výzkumu šéfredaktorů časopisů z oblasti pedagogiky, psychologie a sociologie. Výzkum byl proveden na širokém souboru 311 časopisů. Dle zjištění DeVaneyho většina zkoumaných časopisů nemá dokument, který by určoval, jak je vhodné nakládat se statistickou a věcnou významností. Nicméně většina šéfredaktorů se vyjádřila, že takový dokument by byl vhodný, a také se vyslovila pro přijímání statisticky nevýznamných výsledků k publikaci a také k uvádění měr věcné významnosti. Poslední uvedené studie jsou inspirací pro další empirický výzkum v oblasti užívání statistické a věcné významnosti v ČR.

V Česku obdobná empirická reflexe téměř neprobíhá, autor zaznamenal jediný empirický text [Cuberek, Frömel 2011]. Cuberek s Frömelem v návaznosti na text Soukupa s Rabušicem [2007] zmapovali používání statistických testů pro data, která to neumožňují, v Československé psychologii a pěti kinantropologických časopisech v letech 2009–2010. Jejich závěr pro Československou psychologii byl, že 90 % kvantitativní produkce používá statistické testy nesprávně, protože data nepocházejí z pravděpodobnostního výběru ani randomizovaného experimentu. Cílem předloženého článku je navázat na uvedené snahy a podat komplexní obraz tří předních českých časopisů (Československá psychologie, Pedagogika a Sociologický časopis) v posledním desetiletí.

Dva exkurzy k problematice užívání statistických testů pro populační data a nenáhodné výběry

Exkurz o nadpopulacích[5] *(Je možné užít statistické testy pro celopopulační data?)*

Pro korektnost popisu možných přístupů k datům, která jsou získána za celou populaci, je namístě popsat alespoň ve zkratce koncept nadpopulací (superpopulation) a nabídnout úvahy o možnosti využívat statistické testy (případně alternativně bayesovskou statistiku) i pro tato data. Tento exkurz vychází z diskuse, která proběhla na stránkách časopisu Sociological Methodology mezi excelentními statistiky a metodology (pro úplnost dodejme, že většinově preferujícími spíše bayesovské přístupy před klasickými statistickými testy). Diskusi zahájili Berk, Western a Weiss [1995a] názorem, že pro data, která jsou celopopulační, lze samozřejmě používat popisnou statistiku, pokud je cílem popsat situaci „tady a teď“. Nicméně dodávají, že je také možné uvažovat o jakési nadpopulaci a že

[5] Tento i následující exkurz vznikly na základě podnětu jednoho z anonymních recenzentů, za což mu patří dík. Za případné nedostatky obou exkurzů odpovídá samozřejmě autor článku.

data za populaci lze pak považovat za výběr z této nadpopulace vzniklý náhodnými procesy ve společnosti.[6] Pokud je tento předpoklad přijat, lze uvažovat o zobecňování skrze statistické testy (či bayesovské popstupy) a není ideální používat jen popisnou statistiku (i když ji autoři nevylučují). Při srovnání klasických testů a bayesovské statistiky dospívají autoři textu k závěru (opírají ho zejména o nereálný požadavek nekonečné replikace výběrů zakotvený v klasickém testování), že pokud má býti zobecňováno z celopopulačních dat, má býti využito bayesovské statistiky, která je pro tato data z hlediska své filozofické koncepce vhodná (stačí jí jediný datový soubor a nežádá, byť hypotetickou, možnost generování dalších datových souborů). I přes preferenci bayesovského přístupu autoři v diskusi dále korektně zmiňují problémy, které taková aplikace bayesovské statistiky přinese, a upozorňují, že žádný ze tří popsaných přístupů (popisná statistika, klasické testy, bayesovský přístup) není pro celopopulační data bezproblémový, nadto varují před mechanickou aplikací bayesovského přístupu. V návazné diskusi vyjádřili mnohé pochybnosti nad uvedenými závěry Bollen [1995], Rubin [1995] a Firebaugh [1995]. Zejména Bollen a Firebaugh se postavili proti názoru, že užití bayesovského přístupu má být preferováno před klasickými testy pro celopopulační data – dle nich přináší minimálně srovnatelně problémů (Firebaugh explicite zmiňuje problematičnost volby apriorních rozdělení). Relativně nejpozitivněji se k tématu stavěl Rubin, který v zásadě konstatuje srovnatelnost klasického testování a bayesovského přístupu pro celopopulační data. Bollen [1995: 462–463] nadto nabízí další tři přístupy: bootstarap, metaanalýzu a srovnání s výsledky za stejnou populaci v jiném časovém okamžiku pro zajištění robustnosti výsledků. Bollen dále upozorňuje, že typická mezinárodní data (běžný příklad celopopulačních dat) nebývají úplná, protože údaje za některé země chybí. Tato data však nechybí náhodně, a to může výsledky poškodit daleko více než užití klasických statistických testů. Bollen svůj text uzavírá mj. konstatováním, že popisně statistický přístup je použitelný, a zejména zmiňuje, že diskuse o technikáliích zpracování nesmí zakrýt, že klíčové pro interpretaci výsledků (a pro posun sociologického vědění) jsou věcné výsledky, ne až tak jejich statistická významnost či jejich označení hvězdičkami [Bollen 1995: 468]. Ostatně toto demonstruje i citát z Bollenova textu uvozující tento článek. Berk, Western a Weiss [1995b] v odpovědi na tři diskusní příspěvky jednak poukazují na problémy při užívání klasických statistických testů a poukazují na problematičnost Bollenových návrhů. V závěru opět autoři doporučují pro celopopulační data použít bayesovské přístupy. Diskuse ukazuje poměrně výrazný nesoulad mezi jednotlivými názory na práci s celopopulačními daty. S ohledem na skutečnost, že používání popisné statistiky pro tato data žádný z výše uvedených příspěvků zcela nezpochybnil a že některými názory preferována bayesovská statistika není (bohužel?) v českých sociálněvědních časopisech užívána, bude analýza článků mj. sledovat, zda

[6] Tato úvaha se blíží teoriím alternativních historií, z nichž my pozorujeme pouze jedinou, a je pro mnoho vědců zcela nepřijatelná.

autoři užívají pro celopopulační data klasické statistické testy (tedy postup bude souladný s textem Soukupa a Rabušice [2007]). Pro úplnost dodejme, že oblíbené víceúrovňové modely, které jako druhou úroveň využívají země, vycházejí právě z představy jisté „nadpopulace zemí". V souladu s výše uvedenou polemikou probíhá i v této oblasti polemika o možnosti používání statistických testů [srov. Hox, Schoot, Matthijsse 2012; Stegmueller 2013; Bryan, Jenkins 2015]. I zde, pokud autoři souhlasí s použitím testových přístupů, je spíše doporučováno užití alternativních přístupů, zejména bayesovské statistiky a bootstrapu.

Exkurz o designově a modelově orientovaném přístupu při využívání statistických testů (Je možné užít statistické testy pro nenáhodné výběry?)

Druhý a poslední exkurz v tomto článku krátce popíše dvě možnosti při práci s výběrovými daty (designově a modelově orientovaný přístup) a poukáže na možnosti použití statistických testů pro data z nenáhodných výběrů (v případě sociálních věd však možnosti spíše hypotetické). Pro snadné přiblížení lze konstatovat, že designový přístup vychází z náhodných výběrů a pro zobecnění výsledků na zkoumanou populaci zpravidla používá váhy, které upravují nestejné pravděpodobnosti jednotek vybírání [např. Kalton 1983; Kish 2014; Levy, Lemeshow 2008]. Tento přístup by se dal označit jako klasický či tradiční. Zhruba od sedmdesátých let však někteří statistici (zejména s ohledem na klesající návratnost klasických výběrových šetření, dále s ohledem na rozvoj složitějších statistických modelů a zavedení latentních proměnných umožňujících kvantifikaci chyby měření jako doplňku klasické výběrové chyby) přicházejí s tím, že statistické testy bude možné použít i pro nenáhodné výběry (případně populační data) skrze složitější statistické modely, které mj. zohlední způsob získání dat, konkrétně zejména stratifikaci, shlukování případů a nestejné pravděpodobnosti výběru (prvním přehledovým textem srovnávajícím oba přístupy je zřejmě studie Särndala [1978]). Zatímco pro designový přístup je problémem možnost získat náhodný výběr s velkou návratností, v modelovém přístupu je problémem zohlednit v rámci modelu všechny skutečnosti, které ovlivňují generování dat (jednodušší ukázky lze nalézt např. v Little [2004]).

Modelový přístup pak umožňuje pracovat s celopopulačními daty či daty z nenáhodných výběrů a zobecnění provádět na hypotetickou nekonečně velkou populaci (opět je zde tedy nadpopulace), která sestává ze všech jevů, které může model generovat. Dodejme, že používání modelových přístupů je poměrně slibné v ekonometrii (zejména v analýze časových řad), kde modely zohledňují autokorelace, endogenitu a další fenomény. V oblasti sociálních věd se situace nadto komplikuje, protože autoři často užívají designový přístup (váží data) a modelový přístup (například víceúrovňové modely pro hierarchická data, speciální modely řešící mechanismus chybějících hodnot či modely pro latentní proměnné) a pak je možné korektně pracovat pouze s daty z náhodných výběrů.

S ohledem na výše popsané přístupy a skutečnost, že v sociálních vědách autoři používají většinově designově orientované přístupy, případně v kombinaci s modelovými přístupy, lze uzavřít, že pro použití klasických statistických testů pro průřezová data v sociálních vědách (v analyzovaných textech jsem nenalezl žádný příklad analýzy časových řad) je potřebné mít data pocházející pouze z náhodného výběru. U nenáhodných výběrů v rámci modelového přístupu by bylo třeba v rámci modelů striktně dodržovat jejich požadavky a mj. zajistit zohlednění stratifikace, shlukování případů a nestejných pravděpodobnostní vybrání [Sterba 2009: 715–716]. Proto bude užívání statistických testů pro nenáhodné výběry hodnoceno jako nevhodná praxe (tedy postup bude souladný s textem Soukupa a Rabušice [2007]).

Metodologie provedeného výzkumu

S ohledem na výše uvedené příklady empirických studií a dva exkurzy doplňující dříve publikované texty se předložený text snaží komplexněji zmapovat využívání koncepcí statistické a věcné významnosti (po vzoru Ziliaka a McCloskey). Nebylo možné automaticky přijmout kritéria užívaná Ziliakem a McCloskey [2008], byť sloužila pro inspiraci. Analytické kategorie navazují přímo na předchozí v Česku publikované texty Soukupa a Rabušice [2007] a Soukupa [2010, 2013]. Výsledky je tak možné částečně srovnat s výsledky Cubereka a Frömela [2011]. Autor článku provedl obsahovou analýzu produkce tří časopisů[7] dle dále uvedených kritérií. Konkrétně byla pozornost zaměřena na tři oblasti:

I. používání statistických testů pro data, kde tyto využívat nelze (první dílčí výzkumná otázka),
II. nesprávná užívání statistické významnosti, zejména interpretační pochybení a mechanická práce s daty a výsledky analýz (druhá dílčí výzkumná otázka),
III. věcná interpretace výsledků, používání měr věcné významnosti a jejich interpretace (třetí dílčí výzkumná otázka).

Pro jednoduchost popisu bylo při obsahové analýze využíváno kategorií, které se objevily v již publikovaných textech, a tyto kategorie byly ověřeny na předběžné analýze 20 náhodně vybraných textů.

[7] Všechny analyzované články pročetl sám autor, čímž je zajištěna větší shoda kódování. Kódovací schéma bylo vyzkoušeno na souboru dvaceti náhodně vybraných článků. Během tohoto čtení a kódování se ukázala nutnost doplnit některé kódy do seznamu použitých analytických technik, v ostatních částech nebylo potřebné a ke změnám nedošlo.

Ad I. V oblasti nesprávného používání statistických testů bylo sledováno, zda data nepocházejí:

a) z censu,[8]
b) ze záměrného výběru (kvótního, dostupného apod.),
c) z malého výběru,[9]
d) z výběru s extrémně velkým počtem vybraných jednotek či daty spojenými z různých datových souborů.[10]

V případech uvedených pod písmeny b) a c) je užívání statistických testů (respektive intervalů spolehlivosti) problematické, u případu pod písmenem d) zbytečné. Připomeňme, že postupy matematické statistiky (statistické inference) byly odvozeny původně pro náhodné (randomizované) experimenty a posléze pro náhodné (pravděpodobnostní) výběry. Pro jiné situace odvozené postupy automaticky neplatí a nelze je tedy korektně použít. Dodejme ovšem, že v této oblasti probíhají výrazné diskuse a o jejich stručné zachycení se pokousily dva výše uvedené exkurzy. V rámci oblasti popsané římskou I bylo tedy zachyceno, jaký je charakter výzkumných dat použitých v kvantitativní analýze prováděné v článku (jde tedy o charakteristiku, která přináleží jak článku, tak jednotlivým analýzám[11]).

Ad II. U nesprávného užívání statistické významnosti bylo v souladu s členěním v textu Soukupa [2010] sledováno, zda v jednotlivých analýzách v článcích nedochází k těmto problémům:

a) mechanické práce s klasickou 5% hladinou statistické významnosti (hvězdičky, stepwise, nejlepší modely apod.) a jiné mechanické aplikace (například skrývání malých hodnot faktorových zátěží),

---

[8] S ohledem na dříve uvedený exkurz o superpopulaci a diskusi o zpacování cenzových dat ve statistické literatuře se může jevit zařazení této kategorie jako problematické. Nicméně kategorie vychází z již publikovaného textu a nadto je v příslušné empirické části tohoto článku korektně upozorněno na skutečnost, že tento postup nelze pokládat automaticky za chybný.

[9] Pro různé testy je kritérium malého výběru obecně různé, nicméně platí, že výběry do cca 30 jednotek bývají považovány za malé a vyžadují pro případ, že rozdělení zkoumaných veličin není normální, užívání neparametrické statistiky místo parametrické.

[10] Přesně vymezit extrémní velikost (jedinou hodnotou), kdy již nemá smysl užívat statistické testy, nelze, nicméně již od řádu několika tisíc jednotek platí, že testy není třeba používat, protože výběrová chyba bude zanedbatelná a veškreré výsledky budou statisticky významné.

[11] Některé prezentované výsledky mají jako analytickou jednotku článek (platí pro oblast I), některé pak používají jako analytickou jednotku též jednotlivé analýzy v článcích (platí pro oblasti II a III). Toto členění umožňuje pracovat na úrovni jednotlivých statistických technik a konkretizovat problémy (obecné shrnující závěry toliko za články by nepřinášely kýžený analytický vhled).

b) záměna statistické a věcné významnosti (platí, že statisticky významné neznamená důležité),
c) slovní popis „významné, signifikantní" pro statisticky významné výsledky,
d) ignorování výsledku testů statistické významnosti, respektive interpretace v rozporu s těmito výsledky.

Pro obtížnost, respektive častou nemožnost, rozlišování případů pod písmeny b) a c) zejména v případě Československé psychologie[12] jsou výsledky analyzovány společně, tj. nerozlišuje se, zda autor zaměňuje věcnou a statistickou významnost, nebo „jen" užívá výrazů „významný", „signifikantní", „důležitý" pro výsledky, které jsou statisticky významné (věcně však často zcela nezajímavé). U jednotlivých analýz (statistických technik) je vždy zaznamenáno (ve formě mnohonásobné proměnné), jaké formy nesprávného užívání statistické významnosti se vyskytují.

Ad III. U měr věcné významnosti analýza navazuje na text Soukupa [2013] a literaturu tam uvedenou. Celkem bylo v analyzovaných textech objeveno 15 různých měr věcné významnosti použitých v článcích v Sociologickém časopise, v časopise Pedagogika a v Československé psychologii (jejich přehled viz část „Používání měr věcné významnosti a věcná interpretace výsledků" a přílohu článku). Obsahová analýza měr věcné významnosti a věcné interpretace je opět provedena na úrovni jednotlivých analýz (tj. různých statistických technik); vždy je zaznamenáno (ve formě mnohonásobné proměnné), jaké všechny míry věcné významnosti autor využil, a poté je zaznamenáno, zda je provedena jejich interpretace (tj. vyhodnocení její velikosti) a zda je případně provedena věcná interpretace výsledků (tj. není-li užito měr věcné významnosti a jejich interpretace, je sledováno, zda autor nějak věcně výsledky komentuje).[13]

Při obsahové analýze bylo též doplňkově sledováno, zda jsou používány doporučené alternativy ke klasickým testovacím postupům, konkrétně intervaly spolehlivosti a síla testu. Posledním tématem, na které byla zaměřena pozornost, bylo zjišťování, zda autoři využívají při výpočtech alternativní bayesovské přístupy, případně postupy resamplingu. Cílem tohoto zjišťování bylo zmapovat, jak rychle pronikají nové postupy do sociálněvědní analýzy v Česku. S ohledem na velice řídký výskyt však nejsou tyto výsledky v článku prezentovány.

Základním korpusem pro obsahovou analýzu byly všechny kvantitativní analýzy publikované v Sociologickém časopise v posledních dvaceti letech (2005–2014), celkem jich bylo 162. Pro doplnění obrazu je provedeno srovnání s výsledky analýzy hlavních časopisů příbuzných oborů, konkrétně časopisu Pedagogika a Československá psychologie (srovnání je založeno jen na posledních

[12] Texty v Československé psychologii jsou většinou poměrně krátké, přesto však obsahují zpravidla více analýz než texty v ostatních časopisech (srov. dále výsledky v analytické části).
[13] Detailní popis použitého kódování jednotlivých kategorií je v příloze 2 tohoto článku.

deseti letech, tj. 2005–2014[14]). Výběr těchto časopisů ke srovnání byl veden těmito kritérii:[15]

1. Mělo by jít o časopisy s podobnými metodologickými přístupy (tj. v rámci kvantitativní analýzy používají nejčastěji průřezová data, případně experimentální data a používají podobné statistické techniky k jejich zpracování (srov. výsledky dále uvedené). Lze shledat i podobnost či shodnost učebních a výzkumných textů, o které se výzkumníci v oblastech pedagogiky, psychologie a sociologie opírají.[16] Z uvedených důvodu není do srovnání zahrnut žádný ekonomický časopis (nejčastěji využívají jiné modely, zejména modely pro časové řady, nadto učební texty i studijní kurikula jsou výrazně odlišná).
2. Mělo by jít o časopisy dlouhodobě etablované v české (československé) vědní komunitě a reprezentativní pro produkci této komunity (všechny tři časopisy existují déle než 40 let a dva jsou nadto impaktované časopisy vedené ve Web of Science).
3. Mělo by jít o časopisy, které mají minimálně část produkce ve formě kvantitativních analýz, aby byl dostatek materiálu pro analýzu.
4. Mělo by jít o časopisy, kde publikují zejména čeští autoři (srov. dále), není důležité, zda jsou texty česky, nebo anglicky (srov. dále).

K výběru časopisů z oblasti sociologie, psychologie a pedagogiky došlo též v návaznosti na světový kontext. Byly to právě tyto tři disciplíny, kde se vedly (a dosud vedou) intenzivní diskuse nad používáním statistické významnosti (ve

[14] Zde autor pragmaticky vychází z poznatku, že před více než deseti lety byly kvantitativní analýzy zejména s využitím složitějších statistických nástrojů na stránkách časopisu Pedagogika zcela vzácné a srovnání by tedy nebylo možné. Nadto by analýza vyžadovala v případě Československé psychologie nezměrné množství práce, protože kvantitativně orientovaných textů zde vychází mnoho.

[15] Díky popsaným kritériím je zřejmé, že nebylo možné najít v zásadě jiné časopisy. Politologické časopisy kvantitativní analýzy téměř neobsahují, obdobné je to u časopisů z oblasti mediálních studií a teritoriálních studií. V oblasti pedagogiky a sociologie by bylo možné nalézt další časopisy (Data a výzkum, Naše společnost, Orbis Scholae, Pedagogická orientace aj.), nicméně tyto časopisy nejsou „vlajkové lodi" svých oborů, a tak lze čekat, že nejkvalitnější (česky psaná) produkce bude právě na stránkách analyzovaných časopisů.

[16] Jeden z anonymních recenzentů namítá, že srovnání sociologické produkce s pedagogickou či psychologickou, nadto jen na národní úrovni, není vhodné, protože metodika hodnocení vědy nás nutí srovnávat se s mezinárodní sociologickou komunitou. Není prostor zde detailně rozebírat metodiku hodnocení české vědy (částečně je jí věnována příloha 1), nicméně ambicí předloženého textu je podat obraz širší české kvantitativní společenskovědní produkce, srovnání se světem samozřejmě bude do budoucna provedeno též. Autor textu je přesvědčen, že příprava sociálních vědců v oblasti kvantitativní analýzy (s ohledem na její jednotný charakter) by měla být mnohem více integrována (například společnými semináři a konferencemi pro Ph.D. studenty a akademické či vědecké pracovníky).

světové sociologii zejména v šedesátých letech, v pedagogice a psychologii pak v posledních třiceti letech). Připomeňme, že studie DeVaneyho [2001], citovaná v přehledu obdobných studií, měla totožné zaměření (tj. časopisy z oblasti sociologie, psychologie a pedagogiky). Nezanedbatelná je oborová blízkost; připomeňme, že existují mnohé hraniční disciplíny, např. sociologie vzdělání, sociologie výchovy, školní psychologie.

Dodejme, že mezi vybranými časopisy existuje zcela minimální autorský překryv, tj. psychologové publikují v drtivé většině v Československé psychologii, pedagogové v Pedagogice a sociologové v Sociologickém časopise. Nebyl nalezen žádný autor, který by publikoval ve všech třech časopisech, pouze někteří sociologové či psychologové publikovali text v Pedagogice. Proto lze srovnání mezi časopisy přibližně chápat zároveň i jako srovnání tří odborných komunit (srov. ale diskusi o limitech publikovaného výzkumu v závěru článku).

Pro všechny články všech zmíněných časopisů (celkem 363 textů) byla provedena detailní obsahová analýza, která mj. mapovala, jaké statistické postupy pro kvantitativní analýzu dat byly v článcích použity a kolik různých statistických technik autor používá. Dále byla pozornost zaměřena na aktuálnost příspěvků, tj. bylo sledováno, jak stará data autor v analýze používá. Hlavní zaměření obsahové analýzy pak bylo směřováno ke zhodnocení korektně využívané statistické metodologie (viz výše uvedené body I–III). S ohledem na maximální rozsah jednoho článku není předložená analýza zcela jistě vyčerpávající, data a téma bude možné rozvíjet v dalších textech (získaná data budou veřejně dostupná k další analýze). Doplňkově jsou v příloze 1 provedeny analýzy zaměřené na institucionální kontext a vliv autorství na kvalitu kvantitativních analýz. Text nadto rozhodně není čítankou ze statistiky a neřeší, zda na daný výzkumný problém bylo možné (a lépe) použít jinou než autorem použitou statistickou techniku. Obdobně se text nezaměřuje na drobné nuance mezi parametrickými a neparametrickými přístupy.[17]

Dodejme, že pro potřeby orientačního sledování změn v čase u Sociologického časopisu bylo období 2005–2014 rozděleno na čtyři pětiletky,[18] u ostatních dvou časopisů nebylo toto členění využito (produkce byla sledována souhrnně v období 2005–2014). Kromě sledování výše uvedených dílčích kritérií jsou doplňkově uvedeny i některé další výsledky a v případě nesprávné interpretace a užívání měr věcné významnosti je pro produkci v Sociologickém časopise ještě doplňkově užito kritérium použité statistické techniky, tj. výsledky jsou členěny dle použité statistické techniky a umožňují tak konkrétní poučení, jaké problémy se u jednotlivých statistických technik objevují.

---

[17] Autor tímto nijak nezpochybňuje, že jde též o důležité téma, ale pro zvládnutí obsahové analýzy a její zpracování v rozumném čase, bylo téma nutné omezit.

[18] Toto členění poměrně dobře odděluje produkci z devadesátých let a zajišťuje srovnatelně dlouhá období s přibližně srovnatelným počtem článků (srov. dále). Jde o jednoduchou pomůcku. Doplňkově je ukázán i detailní vývoj ve sledovaném období a pro Sociologický časopis jsou zpravidla doplněny i trendové křivky.

Základní popis souboru analyzovaných textů: počty, autoři a doba mezi sběrem dat a publikací[19]

Během uplynulých dvaceti let bylo na stránkách Sociologického časopisu publikováno 162 textů, které využívají postupy kvantitativní analýzy.[20] V letech 1995–2004 bylo publikováno 72 takových textů, v posledních deseti letech pak 90 článků. Je tedy patrné, že produkce těchto textů mírně stoupá, i když je velice nerovnoměrná, jak bude ukázáno dále. Pro srovnání uveďme, že Pedagogika otiskla za posledních 10 let cca polovinu kvantitativně orientovaných textů (48), naopak v Československé psychologii bylo v posledních 10 letech otištěno těchto textů výrazně více než v Sociologickém časopise (155, 10 z nich byly psychometrické analýzy a ty jsou z následujících výsledků vyloučeny[21]).

Detailní přehled počtu článků zaměřených na kvantitativní analýzu v Sociologickém časopise uvádí Graf 1. Je patrné, že nejvíce článků obsahujících kvantitativní analýzu bylo publikováno v letech 2009 a 2013 (shodně 13), naopak nejméně v letech 2002 a 2014 (shodně dva texty). Část uvedených výkyvů má samozřejmě na svědomí náhoda (autor či recenzent mohou vydání textu zdržet či urychlit, některá monotématická čísla jsou záměrně zcela bez kvantitativních analýz), ale celkově při pohledu na graf jsou patrné jisté pravidelně se opakující výkyvy s periodou cca 6–7 let.

Při detailnější analýze se lze zaměřit na jednotlivá čísla a nalézt tak vydání Sociologického časopisu s nejvyšším počtem kvantitativně laděných textů. Vítězem této pomyslné soutěže bylo první číslo vydané v roce 2011, kde bylo námi sledovaných textů celkem devět (šlo o monotematické číslo zaměřené na výsledky studie EVS). Naopak v 64 (ze 120 analyzovaných čísel je to více než polovina) vydáních Sociologického časopisu se neobjevil ani jeden text obsahující kvantitativní analýzu dat. Pro srovnání uvádí Graf 2 vývoj počtu článků ve všech třech sledovaných časopisech za období 2005–2014. Za základ (100 %) je vzat průměrný roční počet kvantitativních článků v příslušném časopise ve sledovaném desetiletí. Graf naznačuje, že počty kvantitativně orientovaných článků neměly ani v jednom časopise tendenci narůstat či klesat, ale jejich počet v čase jen osciluje.

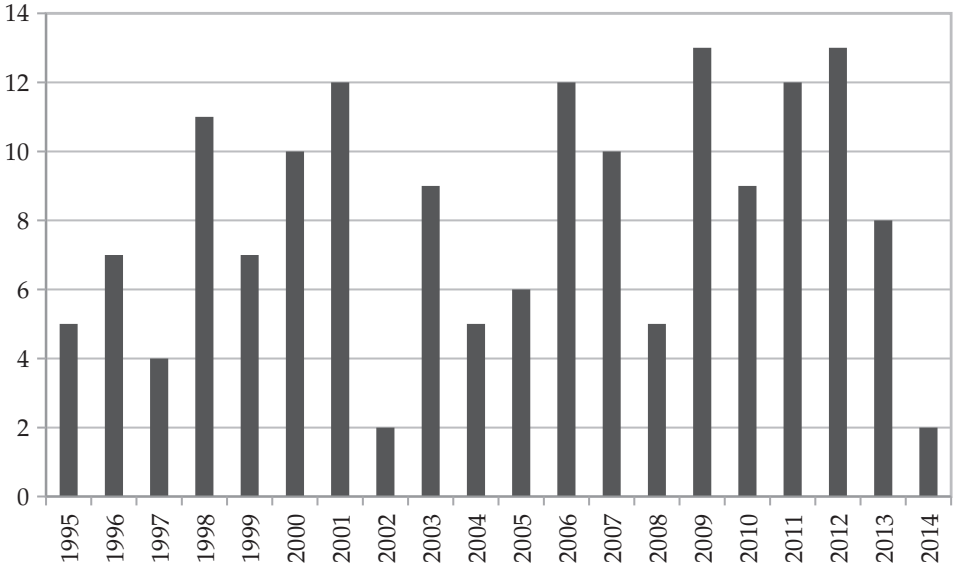
---

[19] Datový soubor je dostupný online na https://doi.org/10.13060/00380288.2019.55.2.459. Jde o anonymizovaný soubor, tj. nejsou zde uváděni konkrétní autoři a názvy článků.

[20] Autor zde přiznává, že i s ohledem na základní téma (statistická významnost a její nesprávná využití) upustil od původního záměru provést pouze náhodný výběr článků. Analýza je tak sice omezena na „pouhé“ popisné statistiky, ale díky tomu je každému zcela srozumitelná a není třeba používat konceptů statistické významnosti. Druhý důvod vedoucí k úplnému zjišťování je skutečnost, že některé okrajové fenomény by nebylo možné výběrem postihnout. Třetím důvodem pak byla vůle autora detailně se seznámit s kompletní kvantitativní produkcí všech tří časopisů.

[21] Psychometrické analýzy běžně ověřují fungování měřicích škál v českém kontextu. Typicky (a nejde zde o chybu) jsou prováděny na záměrných výběrech a jejich zařazení by tak neférově zhoršovalo výsledky zjištěné pro Československou psychologii. Dodejme, že některé zahraniční časopisy však vyžadují i pro psychometrické studie náhodné výběry.

Graf 1. Počty článků zaměřených na kvantitativní analýzu publikovaných v Sociologickém časopise (1995–2014)



Pro úplnost popisu sledovaného souboru článků se ještě zaměřme na autorství článků, respektive na institucionální příslušnost hlavního (prvního) autora (viz Graf 3).

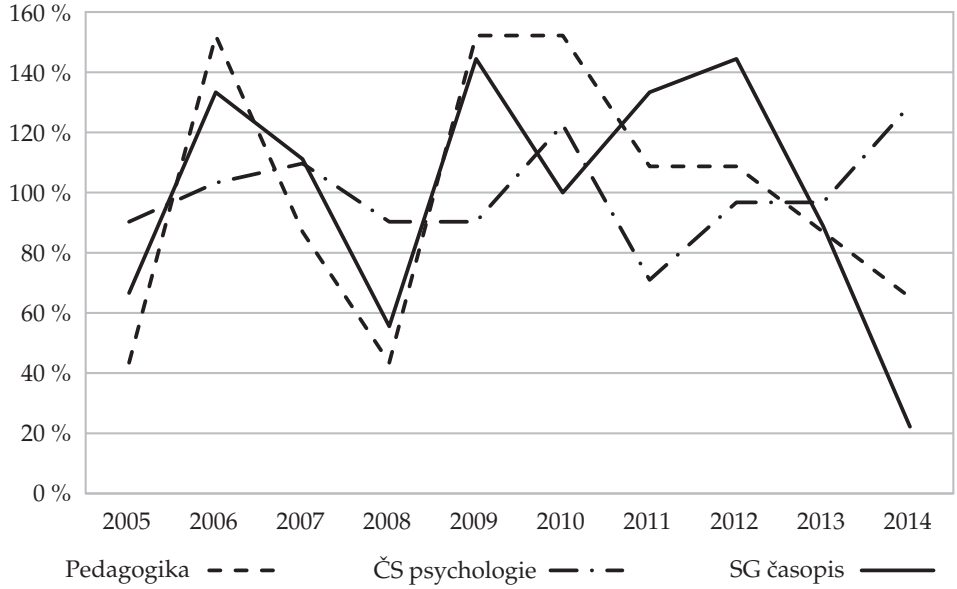
Téměř polovina textů v Sociologickém časopise vznikla na půdě Akademie věd (téměř všichni autoři byli ze Sociologického ústavu AV ČR), zbytek tvoří převážně texty autorů působících na českých vysokých školách, 4 % textů jsou od českých autorů mimo Akademii věd a české VŠ a 12 % textů (převážně anglicky psaných) pochází od autorů působících mimo ČR.[22]

Předposlední popisnou informací o analyzovaném souboru je podíl analyzovaných textů, které byly publikovány česky, respektive anglicky. Podíl anglicky psaných kvantitativně laděných textů v Sociologickém časopise je překvapivě poměrně vysoký (41 %)[23] a s ohledem na předchozí rozbor institucionálního zázemí je nutné konstatovat, že většinu těchto textů publikovali autoři působící v Česku

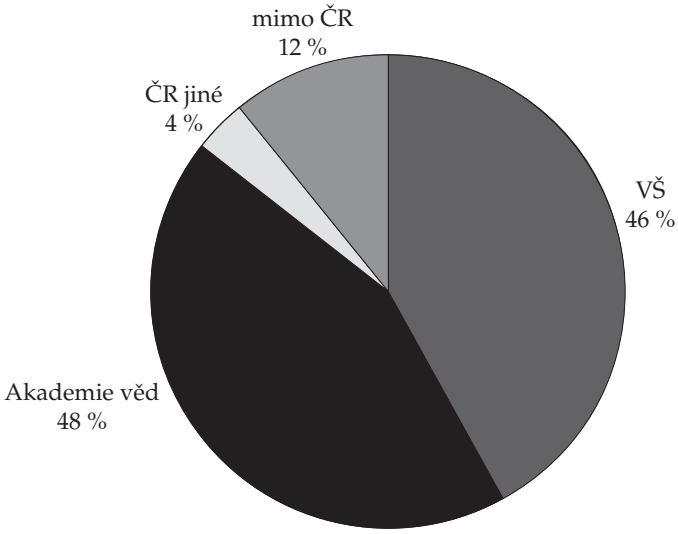
[22] Pro srovnání dodejme, že v Pedagogice i Československé psychologii pochází nejvíce kvantitativně laděných textů od autorů, kteří působí na českých vysokých školách (76 % v Pedagogice a 51 % v Československé psychologii). Je tedy zřejmé, že Sociologický ústav AV ČR má zcela specifické postavení a publikační aktivita jeho pracovníků na stránkách časopisu, který ústav sám vydává, je enormní.

[23] V Pedagogice nebyl za celé sledované období publikován anglicky jediný text, v Československé psychologii celkem osm článků (tj. 6 % kvantitativní produkce).

Graf 2. Vývoj počtu článků zaměřených na kvantitativní analýzu (100 % = průměr sledovaného období) publikovaných v letech 2005–2014 (dle časopisu)



Graf 3. Podíly článků zaměřených na kvantitativní analýzu v Sociologickém časopise (2005–2014) dle institucionální příslušnosti hlavního autora



(16 článků v angličtině napsali autoři působící mimo ČR, čeští autoři publikovali v angličtině v Sociologickém časopise za posledních 20 let celých 51 textů). Naopak minimum textů v českém jazyce má (logicky) autora působícího v zahraničí (byly nalezeny tři takové texty). Z pohledu jazyka článku a institucionální příslušnosti autora je zajímavé též zjištění, že zatímco u českých textů je podíl autorství mezi českými VŠ a Akademií věd téměř vyrovnaný (42, respektive 47 textů), u textů v angličtině je výrazný nepoměr (16 vs. 31 textů). Tyto rozdíly lze též interpretovat jako potvrzení většího tlaku na mezinárodní ohlas publikací na půdě Akademie věd, respektive Sociologického ústavu.

Poslední analyzovanou kategorií v rámci popisu souboru článků, je doba, která uplyne mezi sběrem dat a publikací textu, který tato data využívá. Obecně je dle názoru autora žádoucí, aby byl v článku uvden rok, kdy byla sbírána jimi analyzovaná data. To se na stránkách Sociologického časopisu bez výjimky daří (informace nikde nechyběla, oproti tomu v článcích publikovaných v Československé psychologii nebyla informace uvedena za posledních 10 let ve více než 40 % článků). Průměrná doba, která uplyne mezi sběrem dat a publikací v Sociologickém časopise je v posledních 20 letech 3,5 roku (směrodatná odchylka činila necelé dva roky).[24] Nutno zároveň dodat, že tato doba postupně narůstá, protože mezi lety 1995–1999 to byly průměrně jen dva roky a v poslední sledované pětiletce (2009–2014) činil průměr již 3,7 roku. Dílem lze toto prodloužení mezi sběrem dat a publikací připsat delšímu recenznímu řízení, ale bylo by možné též očekávat, že procesy sběru dat a jejich zpracování budou dnes rychlejší a k prodlužování nebude docházet. V každém případě je dle autora článku žádoucí, aby k publikaci docházelo co nejrychleji od získání výzkumných dat, aby byly publikované výsledky ještě platné pro zkoumanou společnost (v oblasti psychologického výzkumu může být ale situace odlišná). Toto platí, pokud chceme z našich dat zobecňovat na populaci, z níž byl proveden výběr jednotek. Pokud jsme zastánci nadpopulací (viz výše uvedený exkurz), pak je tento požadavek zbytný, nicméně ani zastánci nadpopulací se nevyhnou námitce o posunech například u postojů a chování zkoumané populace, které je nutno vzít v potaz. Modely pro tyto posuny sociálním vědám chybí. Tedy i zde je požadavek na rychlé zpracování získaných dat aktuální.

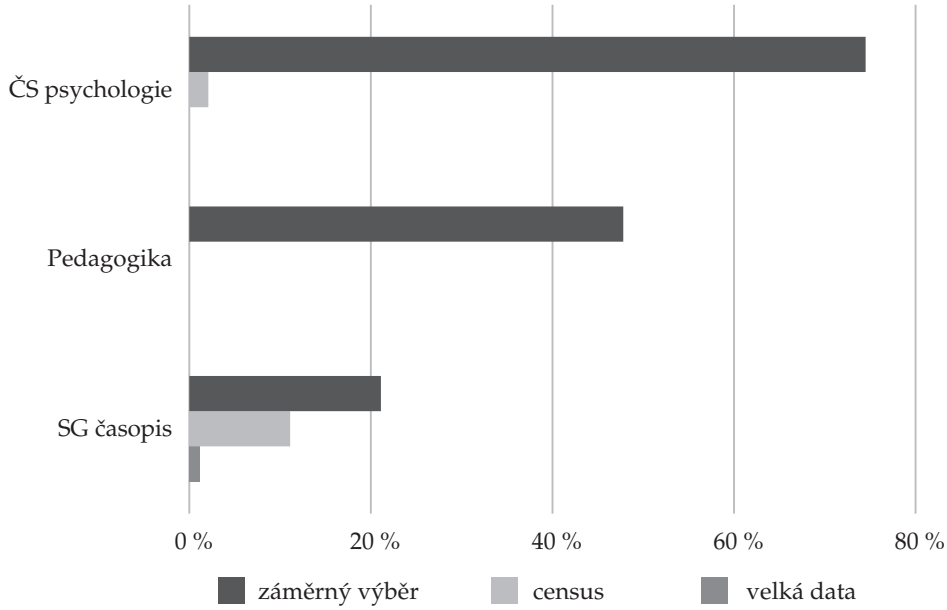
Nesprávné používání statistických testů

Srovnání časopisů (2005–2014)

Nyní se zaměříme na první dílčí výzkumnou otázku, tj. budeme sledovat, zda se v analyzovaných článcích používají statistické testy i v případech, kdy to analyzovaná data neumožňují, respektive to není vhodné (srov. výše část věnovanou

[24] Maximální doba mezi sběrem dat a publikací činila 12 let, pětina textů pak vycházela z dat, která byla maximálně jeden rok stará.

Graf 4. Podíly článků zaměřených na kvantitativní analýzu (2005–2014), kde se nesprávně užívají statistické testy (dle časopisu)

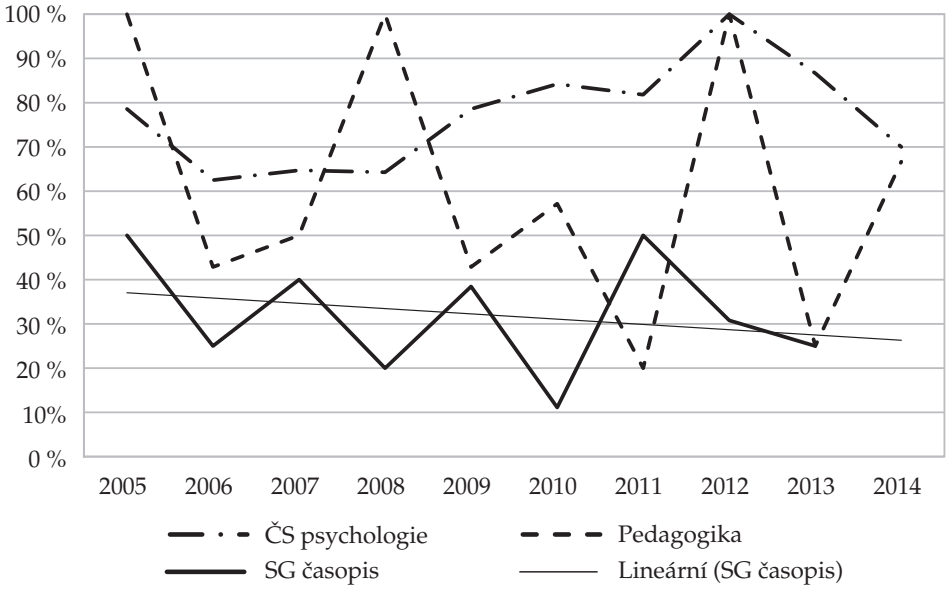


metodologii a dva exkurzy). Základní přehled situace ve všech sledovaných časopisech v posledních 10 letech podávají grafy 4 a 5.

V případě Sociologického časopisu autoři zhruba v pětině případů užívali statistické testy pro data z kvótních výběrů a cca v desetině případů pro data, která byla získána pro celou zkoumanou populaci. Připomeňme, že používání testů pro celopopulační data není nutně chybná strategie, o její vhodnosti se vede v literatuře diskuse (viz první exkurz v tomto článku). *Celková míra nesprávného užívání statistických testů tedy činila v případě Sociologického časopisu cca jednu pětinu*.[25] Tato hodnota je ve srovnání s dalšími dvěma časopisy nízká, u Pedagogiky činí podíl chybné aplikace statistických testů 48 % a v případě Československé psychologie pak dokonce 75 %![26] Při srovnání s výsledky studie Morrisona a Henkela [1969] pak můžeme konstatovat, že produkce v Sociologickém časopise v posledním desetiletí je z pohledu korektního užívání statistické významnosti

[25] Pokud nebereme v potaz užívání testů pro populační data. S ohledem na jejich minimální výskyt v ostatních dvou časopisech, je pro srovnání vynechání této kategorie korektní.
[26] Cuberek a Frömel [2011] uvádí pro období 2009–2010 dokonce údaj 90 %. Rozdíl lze vysvětlit tím, že autoři striktně při nejasném popisu zařazovali text do kategorie nenáhodných výběrů. Postup v tomto článku byl ten, že pokud byl uveden byť jen název výzkumu, byla dohledána detailní informace o tomto výzkumu a dle toho bylo provedeno zařazení.

Graf 5. Vývoj podílu článků zaměřených na kvantitativní analýzu (2005–2014), kde se užívají statistické testy pro data z censů a nenáhodných výběrů (dle časopisu)



Poznámka: Pro Sociologický časopis byl doplněn lineární trend.

lepší, než byla v American Sociological Review v padesátých a šedesátých letech minulého století, obdobně jsou výsledky v zásadě příznivé při srovnání se studií Erin Leahey [2005], která v American Sociological Review a American Journal of Sociology do roku 2000 hovoří o cca polovině článků. Nicméně rozhodně nelze podlehnout alibistickému odůvodňování nesprávné praxe s odkazem, že v jiných disciplínách (či dříve) byla tato praxe častější, ostatně následná analýza článků Sociologického časopisu poukáže na skutečnost, že vývoj v delším čase není povzbudivý (byť trend v posledním desetiletí je mírně klesající).

Detailnější analýza Sociologického časopisu (1995–2014)

Nahlédneme-li na dvacetiletí Sociologického časopisu, pak se míra chybné aplikace statistických testů ještě sníží (klesne na 17 %[27]). Z toho logicky plyne, že v posledním desetiletí se situace zhoršila. Pro detailnější analytický náhled po pětiletkách využijeme Tabulku 1.

[27] Opět nejsou brány v potaz případy, kde se testy užívají pro populační data.

Tabulka 1. Podíly článků v Sociologickém časopise zaměřených na kvantitativní analýzu (1995–2014), kde se nesprávně/problematicky užívají statistické testy (pro pětiletá období; sloupcová procenta)

Data	1995–1999	2000–2004	2005–2009	2010–2014	Celkem
Záměrný výběr	8,8	13,2	19,6	22,7	16,7
Census			15,2	6,8	6,2
Velká data				2,3	0,6
Celkem	8,8	13,2	34,8	31,8	23,5

Zatímco v první zkoumané pětiletce (1995–1999) činila míra nesprávné aplikace statistických testů 9 %, v poslední zkoumané pětiletce (2010–2014) činí již 25 % (tj. téměř trojnásobek!). Uzavřeme, že tento vývoj nelze považovat za příznivý, i když v poslední sledované pětiletce došlo k mírnému zlepšení.

Doplňkově ještě zmíníme výsledky nesprávného užívání statistických testů dle afiliace prvního autora článku v Sociologickém časopise. Nejčastější je překročení mezí pro aplikaci statistických testů u autorů z českých vysokých škol (30 %), u autorů z Akademie věd a těch, kteří působí mimo ČR, jde o pětinu případů. Trošku paradoxně nejméně selhávají autoři z ČR působící mimo akademická pracoviště (jen ve 14 % případů).[28] Na okraj dodejme, že při srovnání článků psaných česky a anglicky jsou výsledky v zásadě totožné.

Nesprávná interpretace statistické významnosti a mechanické využívání statistiky

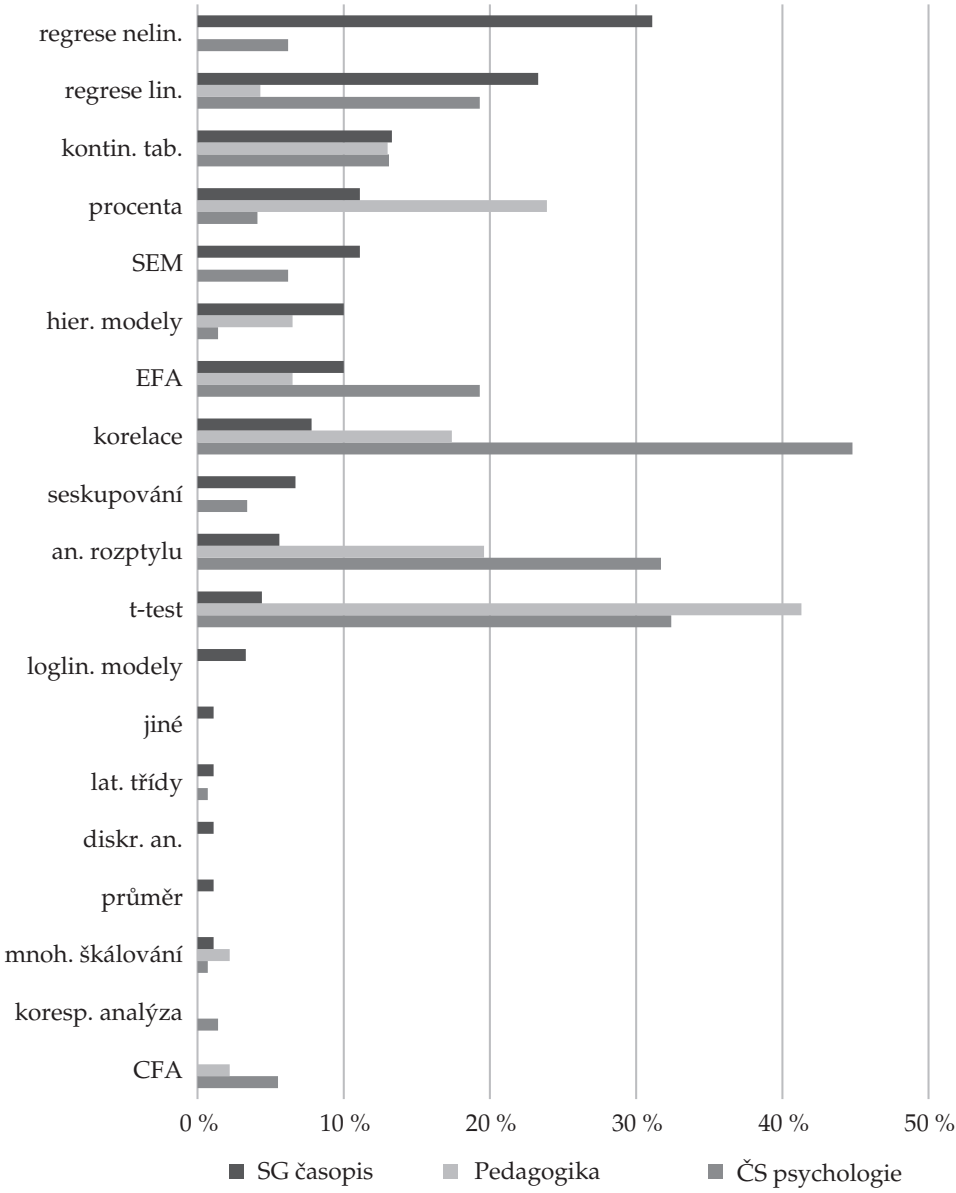
Srovnání statistických technik užívaných v jednotlivých časopisech (2005–2014)

Ještě než se zaměříme na druhou dílčí výzkumnou otázku, provedeme popis statistických technik, které autoři publikující v jednotlivých časopisech využívají. Toto rozlišení pak bude využito pro další analýzy a také pro pochopení rozdílů ve výsledcích mezi jednotlivými časopisy (jde fakticky o jeden ze základních klíčů pro pochopení výsledků). Základní přehled frekvencí užívání jednotlivých statistických technik ve třech sledovaných časopisech za posledních 10 let podává Graf 6.

V Sociologickém časopise je téměř ve třetině případů užito nelineární regrese (typicky binární logistická regrese), ve více než pětině textů je užita lineár-

[28] Zde je ale nutno poznamenat, že tito autoři byli prvními autory jen u sedmi článků a v jediném bylo využito statistického testování pro nevhodná data (tvořila celý základní soubor). Nelze z toho tedy dovozovat žádné závěry.

Graf 6. Užívání statistických technik v článcích publikovaných v letech 2005–2014 (dle časopisu)



Poznámka: Základem (100 %) je počet článků.
Vysvětlivky: EFA = explorační faktorová analýza, CFA = konfirmační faktorová analýza, SEM = strukturní modelování.

ní regrese. V Československé psychologii vévodí užívání korelací (45 %), t-testů a analýzy rozptylu (shodně 32 % článků). Pro Pedagogiku jsou pak nejtypičtější technikou t-test (42 % článků), procenta (24 %), analýza rozptylu (20 %) a korelace (17 %). Pro korektnost srovnání doplňme, že zatímco v Československé psychologii bývá typicky užito korelací na počátku textu jako první techniky (na ni navazují typicky t-testy či regresní postupy) u Pedagogiky bývají naopak korelace často jedinou použitou technikou.

V každém případě lze z přehledu nejčastěji používaných technik usoudit, že sociologická analytická praxe je odlišná od praxe pedagogů a psychologů (ty jsou obdobné, jen s tím, že psychologové užívají většinou více technik a častěji užívají techniky složitější, viz popis dále). Zajímavý pohled přináší srovnání užívání složitějších vícerozměrných statistických technik (tj. faktorové analýzy, strukturních modelů, hierarchických lineárních modelů, analýzy latentních tříd, shlukové analýzy, diskriminační analýzy a loglineárních modelů). V případě Pedagogiky se tyto techniky téměř nevyskytují (jednou byla užita konfirmační faktorová analýza, jednou mnohorozměrné škálování a třikrát hierarchické modely[29]). Zajímavý je rozdíl mezi Sociologickým časopisem a Československou psychologií – psychologové užívají více explorační faktorovou analýzu (pětina publikovaných textů), naopak sociologové využívají častěji hierarchických lineárních modelů.

Popsané rozdíly je nutné vzít v potaz při čtení dalších výsledků – lze například očekávat, že typicky u jednodušších technik věnují autoři méně pozornosti popisu výsledků (očekávají, že jim čtenář rozumí), naopak u složitějších (a nově) užívaných technik lze očekávat detailnější popis (tedy lze očekávat, že u jednodušších výsledků bude chybět věcná interpretace jako celek či interpretace měr věcné významnosti, případně tyto vůbec nebudou použity, srov. výsledky dále). Některé další údaje o vývoji užívání statistických technik v v Sociologickém časopise za posledních 20 let lze nalézt v příloze článku.

Srovnání nesprávné interpretace a mechanické aplikace statistické významnosti v jednotlivých časopisech (2005–2014)

Výše uvedený popis užívaných technik v jednotlivých sledovaných časopisech usnadní interpretaci výsledků odlišné míry nesprávného užívání statistických technik, respektive jejich mechanické aplikace (druhá dílčí výzkumná otázka). Lze očekávat zejména výskyt mechanických aplikací (nejčastěji zřejmě hvězdiček ve výstupech statistických procedur)[30] a nesprávnou interpretaci výsledků, typic-

[29] Tyto tři texty nadto napsali autoři oborově zakotvení v sociologii a primárně publikující v Sociologickém časopise.

[30] Toto očekávání plyne z běžné praxe, kdy autoři kopírují do článků a beze změny publikují tabulky produkované statistickým software. Tento software pro pohodlí uživatelů většinou hvězdičky do tabulek zobrazuje.

ky záměnu věcné a statistické interpretace, tj. označování statisticky významných výsledků za výsledky „významné", „signifikantní", „důležité". Dodejme, že problém těchto nesprávných užití statistiky může mít zejména tyto důsledky:

a) Svádí autora textu k mechanické práci s výsledky a rezignaci na jejich věcnou interpretaci (srov. výsledky v další části tohoto článku).
b) Vede k přeceňování vyhodnocení statistických testů a k umenšování významu či přehlížení dalších aspektů výsledků.
c) Svádí čtenáře (zejména statisticky nepoučeného) k nesprávnému zhodnocení výsledků. Zejména výrazy „významný" či „důležitý" či hvězdičky a jiné mechanické aplikace mohou nepoučeného čtenáře vést k chybnému pochopení výsledků prezentované analýzy a záměně věcné a statistické významnosti.[31]

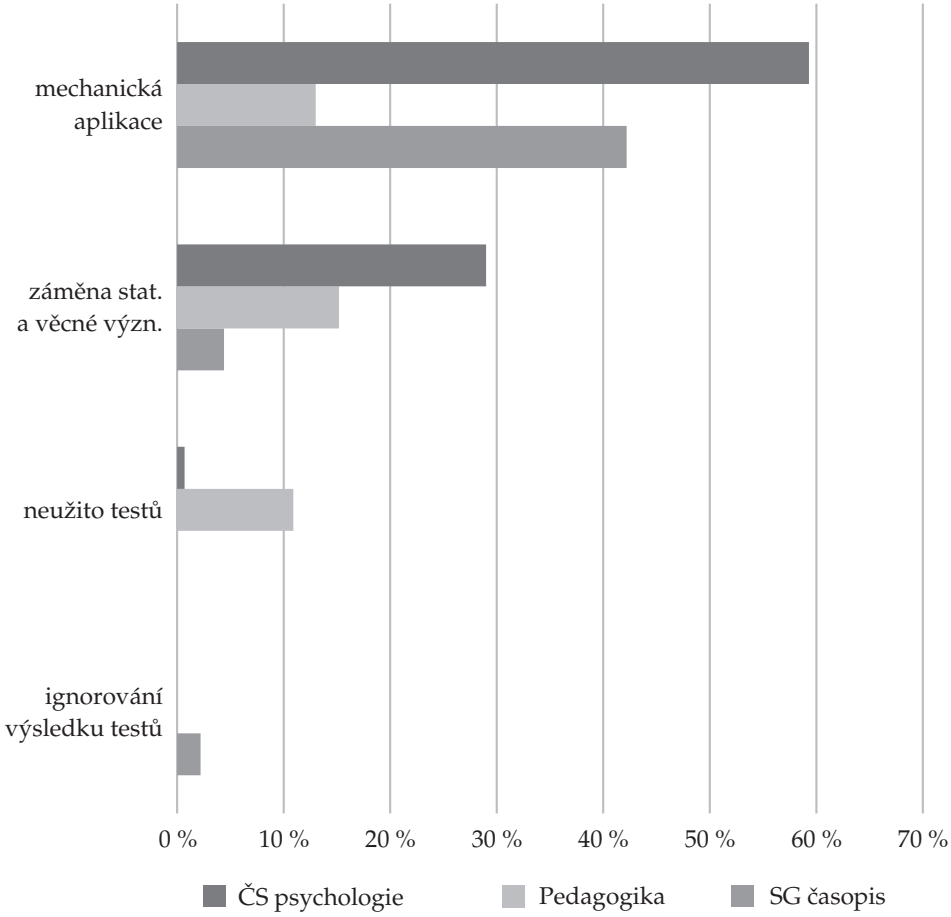
Základní popis výskytu mechanické aplikace statistiky a její případné nevhodné interpretace ve sledovaných časopisech zachycují grafy 7–9.

Zanedbatelným problémem je ignorování výsledků statistických textů, respektive interpretace v rozporu s výsledky (označení statisticky neprůkazného výsledku za statisticky významný a vice versa). Obdobně se s výjimkou časopisu Pedagogika téměř nevyskytují případy, kdy autoři mají data pocházející z náhodných výběrů a k jejich zpracování statistické testy nepoužívají (naopak častý je opačný případ, jak bylo ukázáno v předchozí části tohoto článku). Mechanické aplikace statistiky (respektive mechanické označování výsledků) a nesprávná interpretace jsou naopak poměrně časté. *S mechanickou aplikací se v posledním desetiletí nejčastěji setkáváme na stránkách Československé psychologie* (téměř 60 % textů obsahovalo aspoň jednu mechanickou aplikaci), *méně často v Sociologickém časopise* (42 %) *a poměrně řídce na stránkách Pedagogiky* (13 %). V případě mechanické aplikace lze sledovat v případě Sociologického časopisu i Československé psychologie klesající tendenci. Pro vysvětlení rozdílů mezi časopisy lze částečně využít předchozí popis statistických technik, které jsou v jednotlivých časopisech užívány. Pedagogika představuje časopis, kde se techniky svádějící k mechanické aplikaci (typicky stepwise přístup v regresní analýze) téměř neužívají a obdobně málo jsou užívány též korelace a explorační faktorová analýza (autoři typicky užívají hvězdiček k označování statisticky významných korelací a zakrývání malých hodnot či tučného označení vysokých hodnot faktorových zátěží). I rozdíl mezi úrovní mechanické aplikace statistiky na stránkách Československé psychologie a Sociologického časopisu v posledním desetiletí lze obdobně vysvětlit (psychologové používají častěji korelační přístupy a explorační faktorovou analýzu).[32]

[31] Dodejme, že reálně mohou nastat čtyři možné kombinace: 1. výsledek statisticky i věcně nevýznamný, 2. výsledek statisticky významný, věcně však nikoli, 3. výsledek statisticky nevýznamný, věcně však zajímavý a 4. výsledek statisticky i věcně významný.

[32] Ještě jedno systémové vysvětlení je možné. Článek v Sociologickém časopise může mít maximálně 35 normostran, výzkumná studie v Československé psychologii má však limit

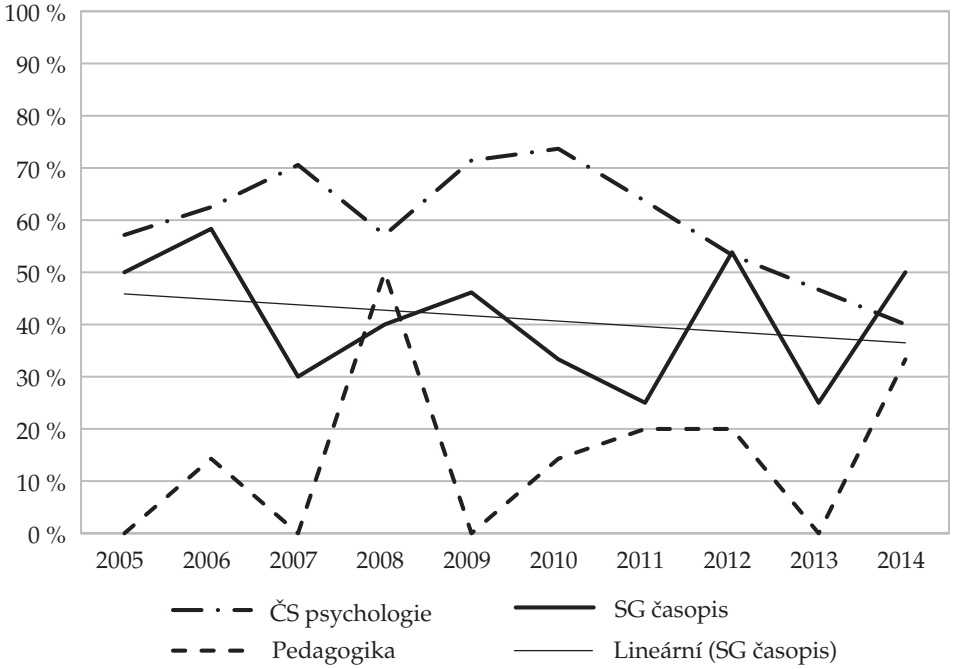
Graf 7. Nesprávné užívání statistických technik a jejich mechanická aplikace v článcích publikovaných v letech 2005–2014 (dle časopisu)



Poznámka: Základem (100 %) je počet článků.

jen 20 normostran (v Pedagogice pak jen 15 normostran). Nadto texty v Československé psychologii obsahují více různých analýz (tento výsledek vyplývá z další analýzy sebraných dat). Autoři textů pro Československou psychologii jsou tak nepřímo pobízeni ke stručnosti a jejím projevem může být i častější využití mechanické aplikace statistiky. Za toto vysvětlení děkuje autor vedoucí redaktorce časopisu Československá psychologie Ivě Šolcové.

Graf 8. Vývoj podílu mechanického užívání statistických testů v článcích publikovaných v letech 2005–2014 (dle časopisu)



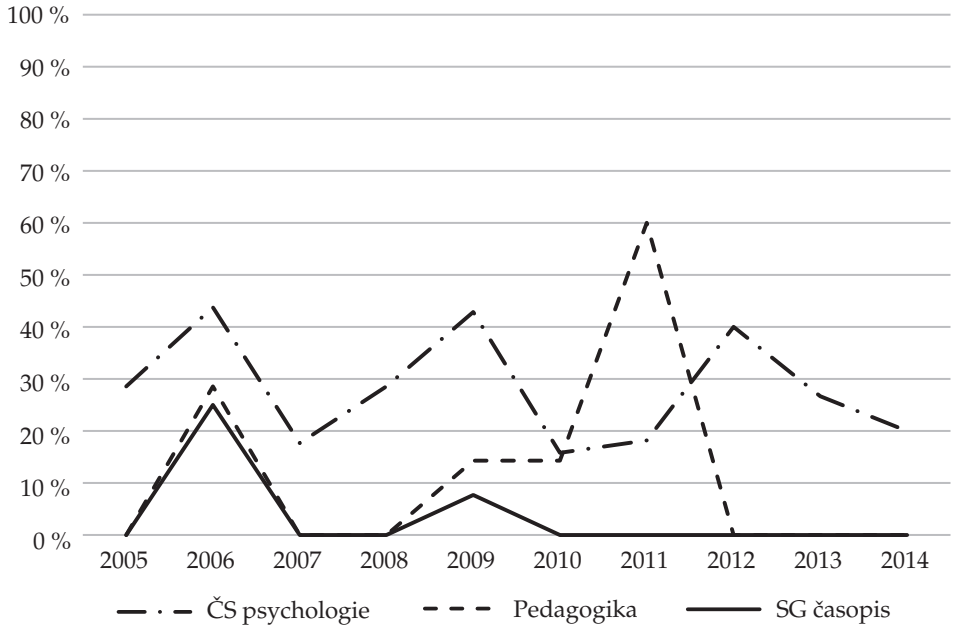
Poznámka: Základem (100 %) je počet článků, pro Sociologický časopis byl doplněn lineární trend.

Nejčastějším případem mechanické aplikace je využívání hvězdiček v korelační či regresní analýze. Samozřejmě, že někdo může namítnout, že používání hvězdiček pro označování statisticky významných výsledků není nijak problematické a v zásadě je téměř uzancí. Nicméně problémů je zde mnoho. Zmiňme základní z nich:

a) Autor (i čtenář) na základě hvězdiček podlehne dojmu, že „ohvězdičkované" výsledky stojí za pozornost a jiné nikoli, nadto k výsledkům s hvězdičkou připojí následně komentář, že jde o „významný výsledek" a zamění statistickou a věcnou významnost.

b) Autoři neužívají hvězdiček jednotně. Nejčastější systém (jedna hvězdička pro 5% hladinu významnosti, dvě pro 1% a tři pro 0,1%) bývá různě variován a bohužel často bez vysvětlení (ve formě poznámky pod tabulkou). Nadto je často vysvětlení nesprávné a situaci ještě zhoršuje, např. „0,1% chybou". Asi nejhorší

Graf 9. Vývoj podílu záměny statistické a věcné významnosti v článcích publikovaných v letech 2005–2014 (dle časopisu)



Poznámka: Základem (100 %) je počet článků.

variantou je, když autor místo zobrazení výsledku (typicky korelačního či regresního koeficientu) uvede hvězdičku či jiný mechanický symbol (často již zmíněný „n.s“). Nikdo tak nemá možnost posoudit a interpretovat posouzený koeficient, jen ví, zda byl, či nebyl statisticky významný (a to je pohříchu málo). Psychologové pak typicky skrývají malé faktorové zátěže (často opět bez uvedení hodnoty, která je hraniční) nebo tyto zátěže zobrazují tučně či kurzívou.

Z výše uvedených příkladů nevhodné praxe se lze přiklonit k doporučení, aby se analytici mechanickému označování výsledků vyhýbali, a pokud už k němu přikročí, pak důsledně užívali jeden styl a ten jasně v textu popsali (ostatně obdobné doporučení lze nalézt i v manuálu APA [2010: 136]).

Druhým nešvarem při interpretaci výsledků analyzovaných textů je záměna statistické a věcné významnosti. Při provádění obsahové analýzy se původní záměr, tj. rozlišit, kdy autor používá pouze nesprávný výraz (typicky slovo „významný“ pro statisticky významný výsledek) a kdy zcela zaměňuje statistickou

a věcnou významnost, ukázal jako nerealizovatelný. V mnoha textech nelze tuto nuanci rozlišit, proto byly sledovány oba fenomény společně. Jejich výskyt není zanedbatelný, v Československé psychologii je nalezneme ve 29 % textů, v Pedagogice v 15 % a v Sociologickém časopise ve 4 %. Pro vysvětlení rozdílů zde lze opakovat důvody již popsané u popisu odlišné frekvence užívání mechanických aplikací snad s tím rozdílem, že autoři v Sociologickém časopise obecně bývají zkušenější analytici než autoři v Pedagogice, a proto se těchto chyb v interpretaci dopouštějí méně často. *Celkově lze uzavřít, že nesprávné interpretace a mechanické aplikace jsou poměrně časté, typičtější je mechanická aplikace než záměna statistické a věcné významnosti.*

Detailnější analýza Sociologického časopisu (1995–2014)

Po obecnějším popisu a porovnání situace tří sledovaných časopisů opět obraťme pozornost na detailnější analýzu Sociologického časopisu. S ohledem na poměrně příznivé výsledky (plynoucí z předchozího srovnání časopisů) je provedeno pouze jednoduché srovnání mechanických aplikací v čase a stručný rozbor mechanických aplikací u nejčastějších statistických technik. Vývoj mechanického užívání statistiky v čase zobrazuje Tabulka 2.

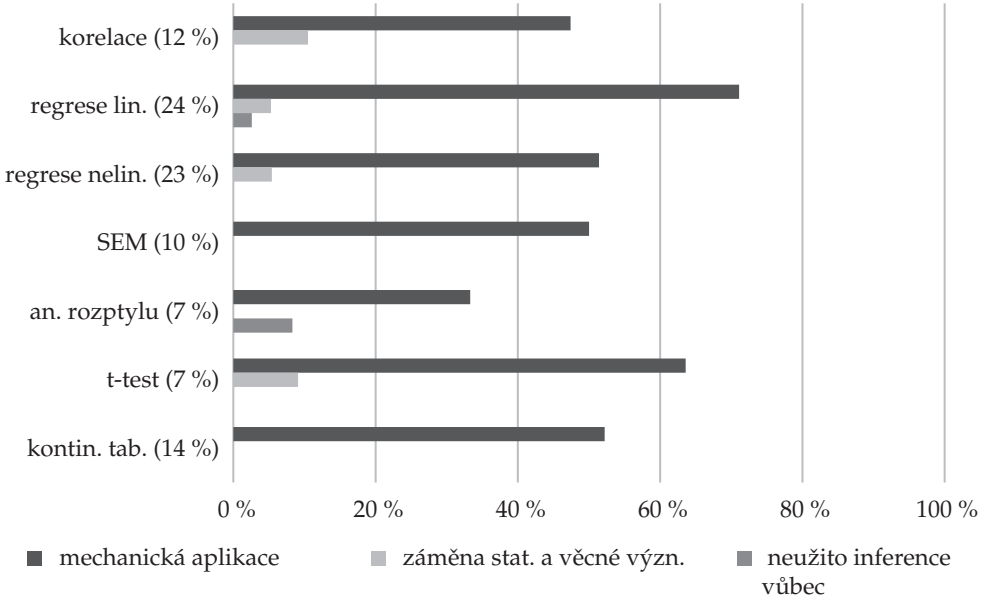
Z tabulky plyne poměrně jednoduchý závěr, že *míra mechanické aplikace statistiky se v textech Sociologického časopisu v čase v zásadě nemění a zůstává zhruba na úrovni 40 %, nicméně v posledním desetiletí dochází k mírnému poklesu* (srov. Graf 8). Zajímavější vhled nabízí analýza chybné interpretace či mechanické aplikace dle jednotlivých nejčastěji užívaných statistických technik (srov. Graf 10).

Z grafu vyplývá, že nejčastější je mechanická aplikace statistiky v případě lineární regrese (71 % publikovaných analýz obsahovalo minimálně jeden příznak mechanické aplikace), časté jsou mechanické aplikace u t-testu (64 %) a zhruba v polovině případů se s nimi setkáváme u korelací, nelineární regrese, strukturních modelů a kontingenčních tabulek. Nejvyšší výskyt u lineární regrese je poměrně snadno vysvětlitelný, protože zde kromě oblíbených hvězdiček používají někteří autoři neméně oblíbený postup mechanického výběru nejlepší podmnožiny prediktorů (téměř výhradně stepwise postup).

Tabulka 2. Podíly článků v Sociologickém časopise zaměřených na kvantitativní analýzu (1995–2014), kde se mechanicky používá statistika (pro pětiletá období; sloupcová procenta)

	1995–1999	2000–2004	2005–2009	2010–2014
Mechanická aplikace	41,2	39,5	45,7	38,6

Graf 10. Nesprávné užívání statistiky a mechanická aplikace v Sociologickém časopise v letech 1995–2014 (dle statistické techniky)



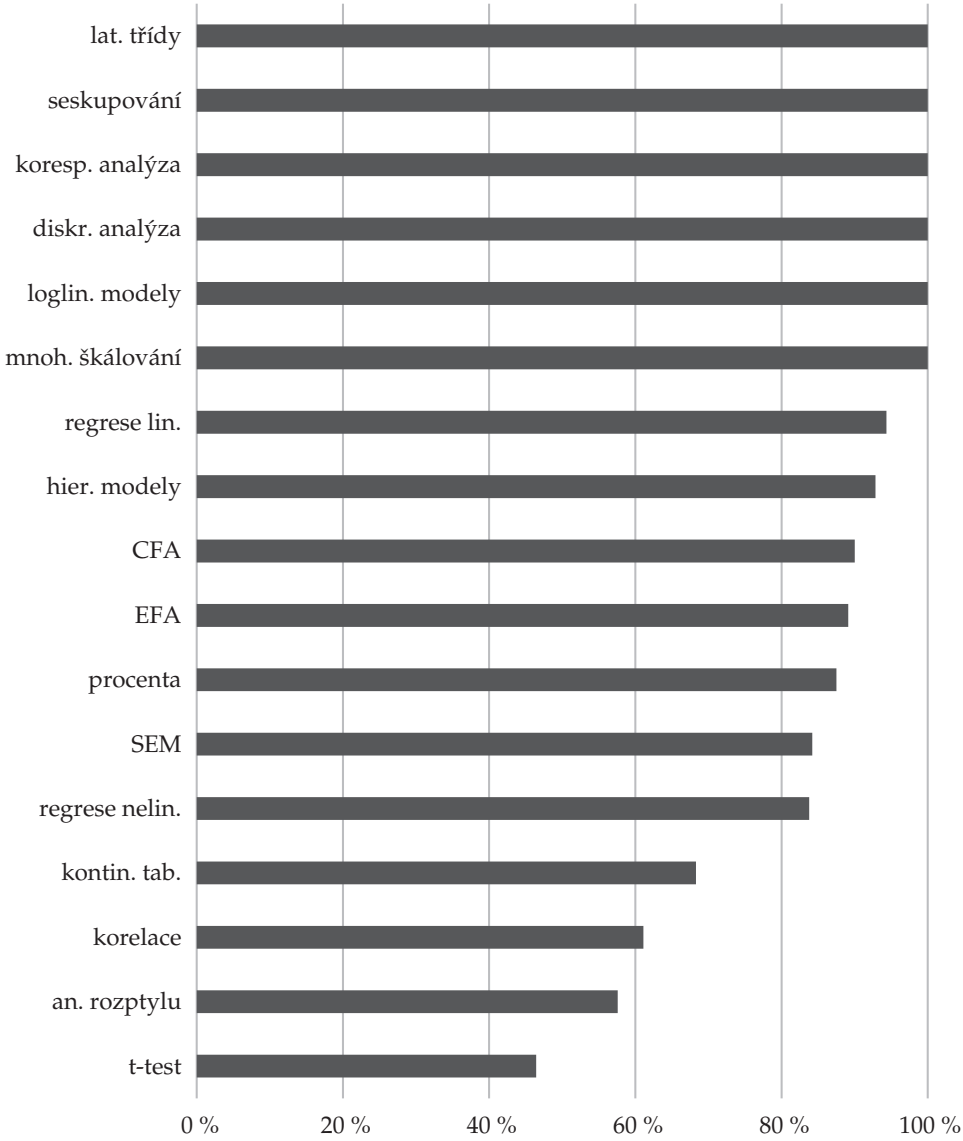
Poznámky: Základem (100 %) je počet článků. Procento za technikou udává podíl kvantitativních článků, v nichž je technika v daném období v Sociologickém časopise použita.

Na okraj k výsledkům prezentovaným v této části článku dodejme, že doplňkově byl sledován výskyt alternativ, které jsou doporučovány k výsledkům klasických statistických testů pro vyhodnocení kvantitativních dat. Konkrétně to byly tyto přístupy:

1. intervaly spolehlivosti,
2. síla testu,
3. resamplingové přístupy (bootstrap, jackknife) a
4. postupy bayesovské statistiky.

S výjimkou Československé psychologie, kde někteří autoři publikují intervaly spolehlivosti (průměrů a regresních koeficientů) a publikují též výsledky síly použitých testů (méně časté), nebyly výše popsané alternativy zachyceny (respektive z 363 analyzovaných textů obsahoval jediný resampling a jediný bayesovský postup). S ohledem na tuto skutečnost bude v budoucnu soustředěna pozornost na prosazování těchto alternativ do sociálněvědní kvantitativní analýzy.

Graf 11. Míra věcné interpretace výsledků u jednotlivých statistických technik v článcích publikovaných v Pedagogice, Sociologickém časopise a Československé psychologii v letech 2005–2014 (dle statistické techniky)



Poznámka: Základem (100 %) je počet článků.

Používání měr věcné významnosti a věcná interpretace výsledků

Srovnání frekvence věcné interpretace výsledků v jednotlivých časopisech (2005–2014)

Poslední část analýzy v tomto článku je věnována zhodnocení praxe při věcné interpretaci výsledků kvantitativních studií publikovaných ve třech sledovaných časopisech. Kromě zhodnocení základní interpretace publikovaných analýz je věnována pozornost i užívání měr věcné významnosti a jejich interpretaci, tj. praxi, kterou předvídají manuály odborných asociací AERA [2006] a APA [2010]. V rámci třetí dílčí výzkumné otázky bude nejprve zkoumána věcná interpretace (dle jednotlivých použitých statistických technik). Pro jednoduchost lze tentokrát základní výsledky prezentovat slovně. *V letech 2004–2015 chyběla věcná interpretace minimálně u jedné publikované analýzy v 9 % článků publikovaných v Sociologickém časopise, ve 39 % textů v Pedagogice a v polovině textů v Československé psychologii*. Částečně lze tento rozdíl opět vysvětlit odlišnými používanými technikami (lze očekávat častější opominutí věcné interpretace u jednodušších technik, srov. dále Graf 11) a odlišným rozsahem textů (nejkratší mohou být v Pedagogice, nejdelší v Sociologickém časopise). Nicméně velkou roli zde budou nejspíše hrát i oborové zvyklosti, tj. skutečnost, že recenzenti vyžadují či nevyžadují věcnou interpretaci výsledků. Graf 11 zobrazuje míru věcné interpretace dle jednotlivých použitých statistických technik napříč třemi sledovanými časopisy.

Z grafu vyplývá, že platí jednoduché tvrzení, že autoři ponechávají bez věcné interpretace spíše výsledky jednodušších statistických technik (t-testů, korelací a analýzy rozptylu), naopak složitější vícerozměrné techniky bývají téměř bez výjimky interpretovány.

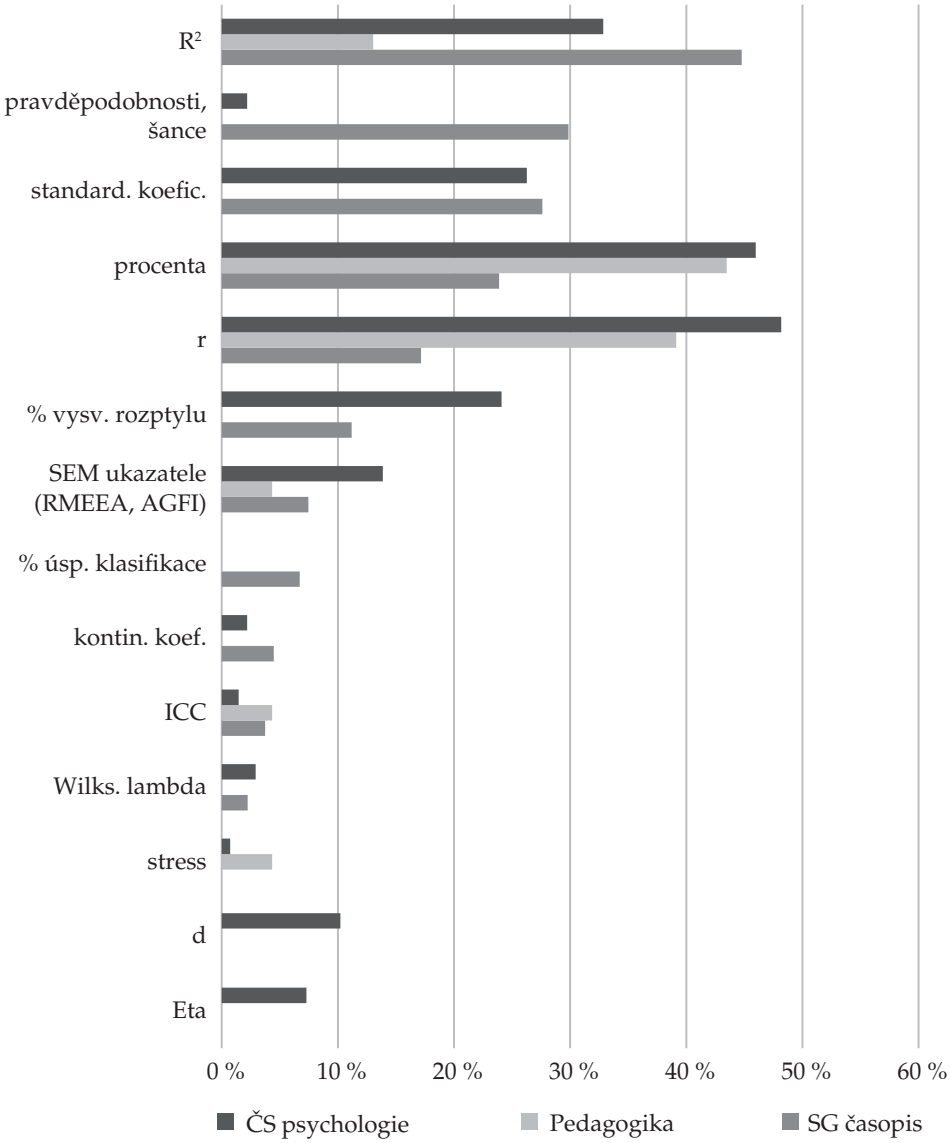
Srovnání frekvence užívání měr věcné významnosti a jejich interpretace výsledků v jednotlivých časopisech (2005–2014)

Po krátkém zhodnocení věcné interpretace výsledků zaměříme pozornost na používání měr věcné významnosti a jejich interpretaci ve třech sledovaných časopisech. Základní přehled užívání jednotlivých měr věcné významnosti pro léta 2005–2014 zobrazuje Graf 12, v následujícím Grafu 13 je pak ukázán časový vývoj podílu článků, kde míra věcné významnosti minimálně u jedné provedené analýzy uvedena není.

Nejčastěji užívanou mírou věcné významnosti v Sociologickém časopise je index determinace[33] (v posledním desetiletí se objevuje ve 45 % textů), hojně je tento ukazatel užíván též na stránkách Československé psychologie (33 %). V textech z Pedagogiky a Československé psychologie vévodí korelační koeficient (39 %,

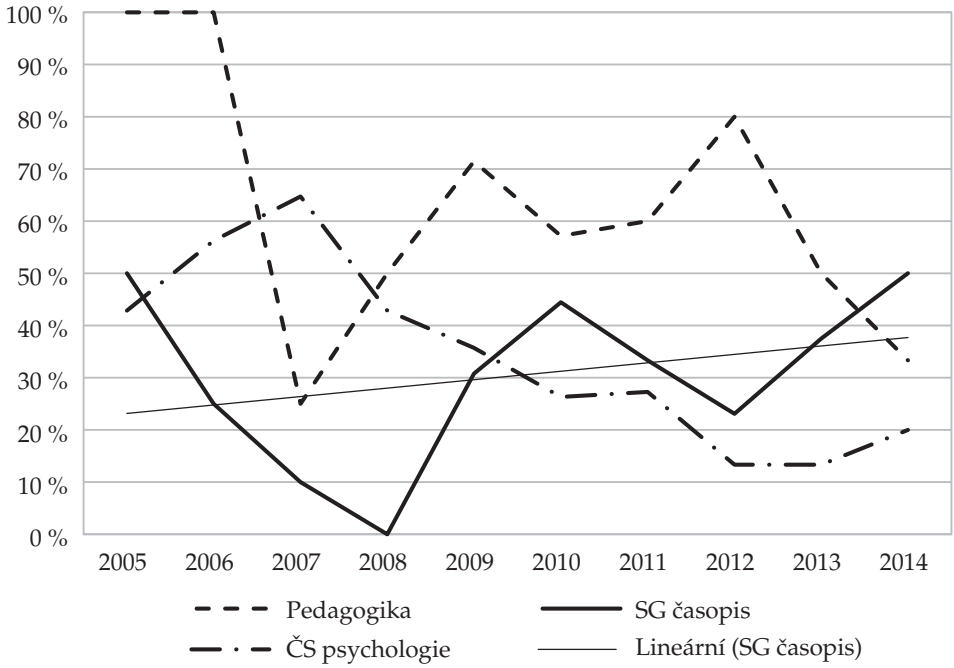
[33] Do této analytické kategorie byl kódován jak ukazatel pro lineární regresní analýzu, tak jeho obdoby pro nelineární regrese.

Graf 12. Užívání měr věcné významnosti v letech 2005–2014 (dle časopisu)



Poznámka: Základem (100 %) je počet článků.

Graf 13. Vývoj podílu článků v letech 2005–2014, kde míra věcné významnosti minimálně u jedné provedené analýzy uvedena není (dle časopisu)



Poznámka: Pro Sociologický časopis byl doplněn lineární trend.

respektive 48 %). *Zhruba ve čtvrtině textů Sociologického časopisu i Československé psychologie autoři užívají standardizované koeficienty* (typicky pro regresní analýzy či strukturní modely), v Pedagogice nebyly užity v žádném článku za sledované období. Za povšimnutí stojí míry věcné významnosti doporučované pro analýzu rozptylu (Eta, respektive Eta^2) a pro t-testy (Cohenovo d[34]). Tyto míry jsou užívány jen na stránkách Československé psychologie. Celkově shrnuto platí, že *v Pedagogice se vyskytuje minimálně jedna míra věcné významnosti v polovině textů, v Sociologickém časopise a Československé psychologii je to výrazně častěji* (87 %, respektive 88 %). Vývoj užívání měr věcné významnosti ve sledovaném desetiletí je ve třech sledovaných časopisech výrazně odlišný (srov. Graf 13). V případě Československé psychologie dochází ke zlepšení (srov. diskusi o institucionálním

[34] Připomeňme, že právě tento ukazatel je použit v názvu článku, protože by bylo možné očekávat s ohledem na jeho jednoduchost, že jeho užívání bude časté.

kontextu vědy v příloze 1), podíl textů, kde míry věcné významnosti chybějí, tedy klesá. V časopise Pedagogika je stav v zásadě setrvalý (první dva sledované roky nebyly míry používány vůbec, protože autoři užívali jen velice jednoduché statistické techniky). V Sociologickém časopise dochází minimálně od roku 2007 ke zhoršení, tj. podíl textů, kde míry věcné významnosti chybí, roste.

Pouhé vypočtení a publikování míry věcné významnosti však není v souladu s doporučeními AERA [2006] či APA [2010]. Namísto této mechanické aplikace (obdoby mechanické práce se statistickou významností) se doporučuje míry věcné významnosti interpretovat. Tato praxe již ale bohužel příliš častá není. *V Pedagogice i Sociologickém časopise je míra věcné významnosti interpretována zhruba ve třetině případů, v Československé psychologii pak v necelé polovině*. Je tedy zřejmé, že psychologická komunita alespoň částečně zaznamenala odbornou mezinárodní diskusi ze stránek odborných časopisů vtělenou do doporučení APA [2010]. Odborná komunita pedagogů dle výsledků provedené analýzy doporučení AERA [2006] zatím nezohlednila. S ohledem na absenci specifičtějších měr věcné významnosti (Eta^2, Cohenovo d aj.) a nízkou míru těchto měr v Sociologickém časopise lze konstatovat, že česká sociologická komunita tuto debatu také nezaregistrovala.

Detailnější analýza užívání měr věcné významnosti v Sociologickém časopise (1995–2014)

Poslední analytický vhled této části bude zaměřen na užívání měr věcné významnosti a jejich interpretaci na stránkách Sociologického časopisu za posledních 20 let s tím, že rozbor bude krátce zaměřen též na nejčastěji užívané statistické techniky a v nich používané míry věcné významnosti. S ohledem na vysokou míru věcné interpretace výsledků v Sociologickém časopise, nemá smysl tuto dále analyzovat. V Tabulce 3 jsou uvedeny po pětiletých cyklech podíly článků, ve kterých bylo minimálně jednou užito míry věcné významnosti.

S výjimkou druhé pětiletky (2000–2004) je využívání měr věcné významnosti velice časté (necelých 90 %). Je namístě si položit otázku, k jaké změně došlo v kvantitativní produkci od roku 2000. Autor textu favorizuje vysvětlení o prvním projevu vzdělanostní expanze, která přivedla na stránky Sociologického časopisu novou generaci autorů, již se časem (během dalších 5–10 let) naučili standardům oboru, a tím se situace vrátila do „normálu". Nicméně tento výkyv určitě stojí za další zkoumání. Ostatně pohled na frekvenci interpretace měr věcné významnosti by zasloužil též další zkoumání. Platí, že více než třetinová míra interpretace měr věcné významnosti v první pětiletce (1995–1999), byla pro následujících 10 let snížena zhruba na polovinu a teprve v posledních pěti letech (2009–2014) opět vzrostla na více než třetinu.

Tabulka 4 shrnuje typicky užívané míry věcné významnosti pro nejčastěji užívané statistické techniky v Sociologickém časopise.

Tabulka 3. Podíly článků v Sociologickém časopise zaměřených na kvantitativní analýzu (1995–2014), kde je užito minimálně jedenkrát míry věcné významnosti (pro pětiletá období; sloupcová procenta)

	1995–1999	2000–2004	2005–2009	2010–2014
Použita míra věc. význ.	88,2	68,4	87,0	86,4

Tabulka 4. Užívání měr věcné významnosti v letech 1995–2014 v Sociologickém časopise (dle statistické techniky)

Statistická technika	Míra věcné významnosti a procentní podíl využití v rámci statistické techniky
Lineární regrese (24 %)	R^2 (76 %), standardizovaný koeficient (45 %)
Nelineární regrese (23 %)	R^2 (54 %), standardizovaný koeficient (16 %), míra úspěšnosti klasifikace (26 %)
Analýza rozptylu (7 %)	Žádná
T-test (7 %)	Průměr (18 %)
Kontingenční tabulky (14 %)	Kontingenční koeficient (26 %)

Poznámka: Procento za technikou udává podíl článků, v nichž je technika v daném období v Sociologickém časopise použita.

U lineární i nelineární regrese je nejtypičtější mírou věcné významnosti index determinace, nicméně jeho používání u nelineárních regresí je méně časté. Standardizovaných koeficientů požívají autoři v případě lineární regrese v necelé polovině analýz, v případě nelineárně regrese pak jen v osmině případů. U dalších technik se míry věcné významnosti užívají řídce (t-test, kontingenční tabulky) nebo vůbec (analýza rozptylu). Pro úplnost dodejme, že výklad o kontingenčních koeficientech, koeficientu Eta či Cohenově d je dnes obsažen i v běžných učebnicích statistiky [srov. např. Hendl 2012]. Používání měr věcné významnosti je doporučováno mj. s ohledem na možnost srovnávání výsledků a provádění metaanalýzy. Pokud autoři tyto míry neuvádějí, pak srovnání ani metaanalýzu neumožňují.

Celkově lze uzavřít, že na stránkách Sociologického časopisu nedochází v posledních 20 letech v čase k výraznějším změnám v užívání měr věcné významnosti a věcné interpretaci výsledků. Diskuse, která probíhá zejména na stránkách mezinárodních psychologických a pedagogických časopisů, se v sociologii téměř neodehrává a není ani české sociologické komunitě tlumočena.

Diskuse výsledků, omezení výzkumu a náměty na další zkoumání

Výsledky představené v tomto článku poukazují na skutečnost, že v české sociálněvědní a příbuzné kvantitativní produkci zdaleka není vše v pořádku. Poměrně často dochází k využívání statistických testů pro data, která toto neumožňují (v Sociologickém časopise se jedná cca o pětinu případů, v Pedagogice o polovinu a v Československé psychologii pak dokonce o více než tři čtvrtiny). Alarmující je, že v případě Sociologického časopisu má tento nešvar narůstající tendenci. Kromě toho dochází poměrně často k mechanické aplikaci statistiky či nesprávné interpretaci (ve více než 40 % článků v Sociologickém časopise za posledních 10 let) a absentuje věcná interpretace výsledků (zejména v Československé psychologii). Míry věcné významnosti jsou používány poměrně často, nicméně v jejich užívání jsou také mezery (novější ukazatele se nepoužívají), nadto poměrně málo dochází k jejich interpretaci. I když ze srovnání tří časopisů (Československá psychologie, Pedagogika a Sociologický časopis) vychází Sociologický časopis téměř ve všech ohledech jako nejlepší (autoři používají složitější techniky, nejméně často používají statistické testy pro nevhodná data, málo nesprávně interpretují výsledky a poměrně často užívají míry věcné významnosti), lze i zde nalézt prostor ke zlepšení. Je nutné připomenout, že obdobná studie zaměřená na European Sociological Review [Bernard, Chakhaia, Leopold 2017] ukázala, že přední evropský časopis vykazuje ještě více pochybení, zejména že autoři často opomíjejí věcnou intrepretaci výsledků (je ovšem namístě připomenout, že kódování tohoto fenoménu obou studií bylo výrazně odlišné a nadto se studie zaměřená na European Sociological Review zaměřila jen na texty uplatňující regresní přístupy).

Pro korektnost závěrů je třeba upozornit na jasná omezení předložené studie. Tato studie srovnává produkci tří předních českých časopisů za posledních deset let (2005–2014). V případě Sociologického časopisu jsou navíc zveřejněny i výsledky za předchozí desetiletí pro možnost zachycení změn v čase. Nicméně nelze automaticky výsledky této studie považovat za výsledky platné pro tři odborné komunity působící v ČR (tj. pedagogy, psychology a sociology). Důvody vedoucí k tomuto jsou následující. Dílem dochází k překryvu autorů publikujících v časopisech – psychologové a sociologové publikují někdy v Pedagogice, nicméně jde o řídký jev. U Sociologického časopisu je nutné zdůraznit, že díky třetině anglicky vydávaných čísel má největší zastoupení zahraničních autorů, v případě Československé psychologie a Pedagogiky jsou typicky cca z 15 % zastoupeni autoři ze Slovenska. Dále je nutno zdůraznit, že analýza se opírá jen o publikované texty. Jde tedy jistě o zkreslený výsledek – texty zamítnuté editory či v recenzním řízení obsaženy nejsou (autor se v budocunu tomuto hodlá věnovat). Bylo by tedy možné konstatovat (ovšem s jistou mírou nepřesnosti), že článek spíše vypovídá o československé komunitě, ostatně psychologický časopis se k této tradici hlásí i svým názvem. Je třeba dále poznamenat, že zejména v posledním desetiletí dochází k častému publikování mimo ČR, a díky tomu nelze výsledky v této studii brát jako jednoznačně platné pro odborné komunity (lze očekávat, že nároky zahraničních kvalitních časopisů budou vyšší, a díky

tomu by výsledky pro texty publikované v zahraničí byly příznivější).[35] Nicméně i stávající výsledky umožňují debatu a případné úpravy publikačních standardů jednotlivých časopisů po vzoru zahraničních asociací, zejména APA [2010].[36] Výsledky slibně naznačují, že standardy používané v Československé psychologii mají pozitivní vliv. Obdobně je zapotřebí reflektovat zjištěné skutečnosti ve výuce statistických předmětů (a zřejmě i ve výuce kvantitativní metodologie), kdy je nutno studenty jednak naučit využívat statistiky ve správných situacích, ale i věcně interpretovat výsledky, ne pouze mechanicky reprodukovat výstupy statistických testů, nadto přebrané bez dalších úprav ve formě nabízené statistickým softwarem. Pro tyto účely by bylo vhodné sestavit čítanku příkladů dobré praxe, aby budoucí sociální vědci znali příklady hodné následování a vyhnuli se tak případným špatným příkladům (student typicky, byť v rozporu se základní maximou „Nikdy nepřísahej na slova učitelova", věří svému učiteli i všemu, co je otištěno v odborných knihách nebo článcích). Doplňkově by bylo vhodné také odhalit, v čem spočívají typická neporozumění statistickým koncepcím u studentů, a dle toho upravit výuku.[37]

Bylo by určitě vhodné provést obdobnou analýzu pro více časopisů etablovaných v České i Slovenské republice v sociálněvědní oblasti (určitě by bylo vhodné provést výzkum slovenské Sociológie) a perspektivně by bylo pro srovnání zajímavé zaměřit pozornost i na přírodovědecké, lékařské či technické časopisy. Lze předvídat, že díky tomu zjistíme, že v problémech s nesprávným užíváním statistiky a její interpretací nejsme sami. To ale není omluvou pro chybnou praxi, spíše podnětem pro její zlepšování.

Petr Soukup *působí na katedře sociologie na Fakultě sociálních věd Univerzity Karlovy. Odborně se zaměřuje na pokročilé techniky analýzy kvantitativních dat, badatelsky se orientuje na sociologii vzdělání. V posledních letech vydal několik článků orientovaných na téma statistické a věcné významnosti.*

Literatura

AERA. 2006. „Standards for Reporting on Empirical Social Science Research in AERA Publications." *Educational Researcher* 35 (6): 33–40, https://doi.org/10.3102/0013189X035006033.

[35] Autor textu hodlá tuto hypotézu v budoucnu dalším výzkumem v oblasti sociologické produkce prověřit.

[36] Lze doporučit stručné přetlumočení klíčových požadavků z APA manuálu ve vztahu ke statistické analýze, které podal ve svém článku Cumming et al. [2012].

[37] I zde hodlá autor textu dalším výzkumem navázat.

APA. 2010. *Publication Manual of the American Psychological Association*. 6th edition. Washington, DC: American Psychological Association.
Bernard, F., L. Chakhaia, L. Leopold. 2017. „'Sing Me a Song with Social Significance': The (Mis)Use of Statistical Significance Testing in European Sociological Research." *European Sociological Review* 33 (1): 1–15, https://doi.org/10.1093/esr/jcx044.
Berk R. A., B. Western, R. E. Weiss. 1995a. „Statistical Inference for Apparent Populations." *Sociological Methodology* 25: 421–458, https://doi.org/10.2307/271073.
Berk R. A., B. Western, R. E. Weiss. 1995b. „Reply to Bollen, Firebaugh, and Rubin." *Sociological Methodology* 25: 481–485, https://doi.org/10.2307/271077.
Blahuš, P. 2000. „Statistická významnost proti vědecké průkaznosti výsledků výzkumu." *Česká kinantropologie* 4 (2): 53–72.
Bollen, K. A. 1995. „Apparent and Nonapparent Significance Tests." *Sociological Methodology* 25: 459–468, https://doi.org/10.2307/271074.
Bryan, M. L., S. P. Jenkins. 2015. „Multilevel Modelling of Country Effects: A Cautionary Tale." *European Sociological Review* 0 (0): 1–20, https://doi.org/10.1093/esr/jcv059.
Cuberek, R., K. Frömel. 2011. „K problematice výzkumného výběru a testování nulové hypotézy." *Československá psychologie* 55 (5): 468–477.
Cumming, G., F. Fidler, P. Kalinowski, L. Pav. 2012. „The statistical recommendations of the American Psychological Association Publication Manual: Effect sizes, confidence intervals, and meta-analysis." *Australian Journal of Psychology* 64 (3): 138–146, https://doi.org/10.1111/j.1742-9536.2011.00037.x.
DeVaney, T. A. 2001. „Statistical Significance, Effect Size, and Replication: What Do the Journals Say?" *The Journal of Experimental Education* 69 (3): 310–320, https://doi.org/10.1080/00220970109599490.
Firebaugh, G. 1995. „Will Bayesian Inference Help? A Skeptical View." *Sociological Methodology* 25: 469–472, https://doi.org/10.2307/271075.
Haller, H., S. Krauss. 2002. „Misinterpretations of Significance: A Problem Students Share with Their Teachers?" *Methods of Psychological Research Online* 7 (1): 1–20.
Harlow, L. L., S. A. Mulaik, M. L. Steiger. 1997. *What if There Were No Significance Tests?* Mahwah (NJ): Erlbaum.
Hendl, J. 2012. *Přehled statistických metod: analýza a metaanalýza dat. Čtvrté, rozšířené vydání.* Praha: Portál.
Hox, J., R. Schoot, S. Matthijsse. 2012. „How Few Countries Will Do? Comparative Survey Analysis from a Bayesian Perspective." *Survey Research Methods* 6 (2): 87–93.
Kalton, G. 1983. *Introduction to Survey Sampling*. California: SAGE, https://doi.org/10.4135/9781412984683.
Kish, L. 2014. *Survey Sampling*. Wiley.
Kline, R. B. 2004. *Beyond the Statistical Testing. Reforming Data Analysis Methods in Behavioral Research.* Washington, DC: American Psychological Association, https://doi.org/10.1037/10693-000.
Leahey, E. 2005. „Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology." *Social Forces* 84 (1): 1–24, https://doi.org/10.1353/sof.2005.0108.
Levy, P. S., S. Lemeshow 2008. *Sampling of Populations: Methods and Applications*. 4th ed. Wiley, https://doi.org/10.1002/9780470374597.
Little, R. J. 2004. „To Model or Not to Model? Competing Modes of Inference for Finite Population Sampling." *Journal of the American Statistical Association* 99 (466): 546–556, https://doi.org/10.1198/016214504000000467.
Mittag, K. C., B. Thompson. 2000. „A National Survey of AERA Members' Perceptions of Statistical Significance Tests and Other Statistical Issues." *Educational Researcher* 29: 14–20.

Morrison, D. E., R. E. Henkel. 1969. „Significance Tests Reconsidered." *The American Sociologist* 4: 131–140.
Morrison, D. E., R. E. Henkel. 1970. *The Significance Test Controversy: A Reader.* London: Butterworth.
Nuzzo, R. 2014. „Scientific Method: Statistical Errors." *Nature* (506): 150–152, https://doi.org/10.1038/506150a.
Oakes, M. 1986. *Statistical Inference: A Commentary for the Social and Behavioral Sciences.* Chichester: Wiley.
Rubin, D. R. 1995. „Bayes, Neyman, and Calibration." *Sociological Methodology* 25: 473–479, https://doi.org/10.2307/271076.
Sarndal, C. E. 1978. „Design-based and Model-based Inference in Survey Sampling." *Scandinavian Journal of Statistics* 5: 27–52.
Selvin, H. 1957. „A Critique of Tests of Signifikance in Survey Research." *American Sociological Review* 22: 519–527, https://doi.org/10.2307/2089475.
Soukup, P. 2010. „Nesprávná užívání statistické významnosti a jejich možná řešení." *Data a výzkum – SDA Info* 4 (2): 77–104.
Soukup, P. 2013. „Věcná významnost výsledků a její možnosti měření." *Data a výzkum – SDA Info* 7 (2): 125–148, https://doi.org/10.13060/23362391.2013.127.2.41.
Soukup, P., L. Rabušic. 2007. „Několik poznámek k jedné obsesi českých sociálních věd – statistické významnosti." *Sociologický časopis / Czech Sociological Review* 43 (2): 379–395.
Stegmueller, D. 2013. „How Many Countries for Multilevel Modeling? A Comparison of Frequentist and Bayesian Approaches." *American Journal of Political Science* 57 (3): 748–761, https://doi.org/10.1111/ajps.12001.
Sterba, S. K. 2009. „Alternative Model-Based and Design-Based Frameworks for Inference From Samples to Populations: From Polarization to Integration." *Multivariate Behavioral Research* 44 (6): 711–740, https://doi.org/10.1080/00273170903333574.
Trafimow, D. 2014. Editorial. *Basic and Applied Social Psychology* 36 (1): 1–2, https://doi.org/10.1080/01973533.2014.865505.
Trafimow, D., M. Marks. 2015. „Editorial." *Basic and Applied Social Psychology* 37 (1): 1–2, https://doi.org/10.1080/01973533.2015.1012991.
Wasserstein, R. L., A. L. Lazar. 2016. „The ASA's Statement on p-Values: Context, Process, and Purpose." *The American Statistician* 70 (2): 129–133, https://doi.org/10.1080/00031305.2016.1154108.
Ziliak, S. T., D. N. McCloskey. 2008. *The Cult of Statistical Significance. How the Standard Error Costs Us Jobs, Justice, and Lives.* The University of Michigan Press.