UNIVERZITA KARLOVA

FAKULTA SOCIÁLNÍCH VĚD

Institut sociologických studií

P a D (statistická a věcná významnost a jejich praktické užívání v českých sociálních vědách)

Autor práce: **Petr Soukup** Školitel: **Prof. PhDr. Hynek Jeřábek, CSc.**

Rok: 2017

Prohlášení

1. Prohlašuji, že jsem předkládanou práci zpracoval/a samostatně a použil/a jen uvedené prameny a literaturu.
2. Prohlašuji, že práce nebyla využita k získání jiného titulu.
3. Souhlasím s tím, aby práce byla zpřístupněna pro studijní a výzkumné účely.

V Praze dne 19. 3. 2017 Petr Soukup

Abstrakt

Předložená disertační práce se zaměřuje na problematiku používání statistické a věcné významnosti v českých sociálních vědách. Větší část práce je věnována popisu problematiky koncepce statistické významnosti a výkladu koncepce věcné významnosti. Pátá kapitola je pak zaměřena na empirickou refllexi užívání statistické a věcné významnosti v českých sociálních vědách, konkrétně byl výzkum proveden skrze obsahovou analýzu článků publikovaných v posledním desetiletí (2005-2014) na stránkách časopisů Československá psychologie, Pedagogika a Sociologický časopis. Pro detailnější vhled byla provedena též obsahová analýza textů ze Sociologického časopisu i pro předchozí desetiletí (1995-2004).

Předložená práce se věnuje koncepcím statistické a věcné významnosti a jejich využívání v oblasti českých sociálních věd. Cílem práce je jednak teoreticky popsat současné „standardy“ v oblasti užívání statistické a věcné významnosti, jednak skrze výzkum článků publikovaných ve třech zmíněných časopisech empiricky zhodnotit, zda publikované články dodržují „standardy“ v oblasti statistické a věcné významnosti výsledků.

Základné výzkumná otázka zní: Nakolik jsou dodržovány současné „standardy“ v oblasti užívání statistické a věcné významnosti v české sociálně vědní produkci prizmatem produkce tří předních českých časopisů? (kapitola 5)

Kromě základní výzkumné otázky byly další (pomocné) výzkumné otázky zaměřeny na následující skutečnosti:

1) Jaká jsou základní omezení koncepce statistické významnosti, z pohledu získaných dat tj. kdy nelze statistické testy (případně interval spolehlivosti) s ohledem na výzkumný design využívat ? (kapitola 2)
2) Jaké jsou nedostatky samotné koncepce statistické významnosti a jaké jsou nejčastější problémy při jejím využívání výzkumníky ve světovém kontextu ? (kapitola 3)
3) Jaké existují alternativy ke koncepci statistické významnosti ? (kapitola 4)
4) Jak lze zhodnotit věcnou významnost výsledků, jaké míry jsou běžně doporučovány?

Empirická část práce se zaměřuje na kvantitativní obsahovou analýzu článků (obsahujících kvantitativní analýzu dat) publikovaných ve třech zmíněných časopisech v posledních 10

letech (celkem 363 článků). Hlavním cílem obsahové analýzy bylo zjistit, zda jsou dodržovány standardy v oblasti statistické a věcné významnosti.

Pozornost byla konkrétně zaměřena na tyto tři aspekty statistických analýz v publikovaných článcích:

I. vhodnost využití statistických testů pro analyzovaná data,

II. nevhodné interpretace a užití statistických testů, zejména záměnu věcné a statistické významnosti a mechanické aplikace statistických testů a

III. věcnou interpretaci výsledků, užívání měr věcné významnosti a jejich interpretaci.

Výsledky empirické části ukazují na skutečnost, že ne vždy jsou dodržovány standardy práce pro kvantitativní analýzu dat. Poměrně časté je užívání statistických testů pro data, která nepocházejí z náhodných výběrů (v Sociologickém časopise zhruba ve třetině případů, v Československé psychologii ve vice než třech čtvrtinách), v Sociologickém časopise lze vysledovat v posledních 20 letech narůstající tendenci tohoto fenoménu. Autoři článků poměrně často používají mechanické postupy statistického testování (hvězdičky, stepwise přístupy), v případě Sociologického časopisu bylo použito více než 40 % článků. Užívání měr věcné významnosti je poměrně rozšířené, nicméně zejména jednoduché míry (např. Cohenovo d, či Eta) jsou užívány jen v Československé psychologii a zejména v posledních letech. V tomto časopise na rozdíl od dalších dvou lze nadto vysledovat nárůst užívání měr věcné významnosti a lze odhadovat, že doporučení (od asociace APA či v publikovaných článcích Urbánka) měla pozitivní efekt. Doplňkově byl zkoumán vliv změny financování české vědy na dodržování standardů v oblasti statistické a věcné významnosti. Lze konstatovat, že na produkci třech sledovaných časopisů nelze prokázat žádné negativní vlivy systému založeného na evidování výsledků v RIV a přepočtu bodů za publikace na finanční prostředky.

Klíčová slova: statistická významnost, věcná významnost, míra věcné významnosti, výběr, Sociologický časopis

Obsah

1. Úvod

Předložená disertační práce se zaměřuje na problematiku používání statistické a věcné významnosti v českých sociálních vědách. Větší část práce (převáženě teoretická, kapitoly 2-4) je věnována popisu problematiky koncepce statistické významnosti a výkladu koncepce věcné významnosti. Pátá kapitola je pak zaměřena na empirickou refllexi užívání statistické a věcné významnosti v českých sociálních vědách, konkrétně byl výzkum proveden skrze obsahovou analýzu článků publikovaných v posledním desetiletí na stránkách časopisů Československá psychologie a pedagogika. S ohledem na mírně netradiční strukturu práce (kapitoly 2-5 jsou publikované či pro publikaci připravené články) má předložená práce delší úvod a také závěr je delší než bývá obvyklé, aby bylo možné doplnit některé skutečnosti, které nelze na ploše časopiseckých článků z důvodu jejich omezeného rozsahu obsáhnout.

Cíl práce a výzkumná otázka

Předložená práce se věnuje koncepcím statistické a věcné významnosti a jejich využívání v oblasti českých sociálních věd. Cílem práce je jednak **teoreticky popsat současné „standardy“[1] v oblasti užívání statistické a věcné významnosti**, jednak skrze výzkum článků publikovaných ve třech předních českých časopisech (**Československá psychologie, Pedagogika a Sociologický časopis**) **empiricky zhodnotit, zda publikované články dodržují „standardy“** v oblasti statistické a věcné významnosti výsledků.

Základné výzkumná otázka zní: **Nakolik jsou dodržovány současné „standardy“ v oblasti užívání statistické a věcné významnosti v české sociálně vědní produkci prizmatem produkce tří předních českých časopisů?**

Odpověď na tuto výzkumnou otázku podává čtvrtý (poslední článek), který tvoří disertaci.

Kromě základní výzkumné otázky byly další (pomocné) výzkumné otázky zaměřeny na následující skutečnosti:

[1] Nejde o standardy v úzkém slova smyslu, které by byly zavedeny všeobecně závaznou normou, nicméně jde o nejrůznější doporučení odborných asociací (zejména APA, AERA a ASA) i některých předních odborníků¨. Dále uvedený exkurz podává přehled těchto standardů.

5) Jaká jsou základní omezení koncepce statistické významnosti, z pohledu získaných dat tj. kdy nelze statistické testy (případně interval spolehlivosti) s ohledem na výzkumný design využívat (více viz první článek tvořící disertaci v kapitole 2)?
6) Jaké jsou nedostatky samotné koncepce statistické významnosti a jaké jsou nejčastější problémy při jejím využívání výzkumníky ve světovém kontextu (více viz druhý článek tvořící disertaci v kapitole 3)?
7) Jaké existují alternativy ke koncepci statistické významnosti (více viz druhý článek tvořící disertaci v kapitole 3)?
8) Jak lze zhodnotit věcnou významnost výsledků, jaké míry jsou běžně doporučovány (více viz třetí článek tvořící disertaci v kapitole 4)?

Exkurz: Zdroje publikačních standardů a jejich obsah

Výraz publikační standardy si zaslouží bližšího vysvětlení. Neexistuje žádná závazná norma (ať národní či mezinárodní povahy), která by stanovovala, co je třeba při publikaci výzkumných výsledků dodržet. Proto výraz standardy je možná lehce nadsazený Nicméně jednotlivé asociace vědců mají ve svých publikačních manuálech či doporučeních nejrůznější požadavky, obdobně tak činí některé časopisy ve svých popisech publikačních náležitostí a mnohá doporučení obsahují i jednotlivé články, které se věnují problematice statistické a věcné významnosti. S ohledem na zaměření předkládané práce (oborově je zkoumána pedagogická, psychologická a sociologická produkce) je třeba primárně vycházet z doporučení příslušných asociací a je nutné též přihlédnout k doporučením statistiků, jejichž nástroje jsou využívány při prezentaci kvantitativního výzkumu. Základními dokumenty, z nichž vychází souhrn publikačních standardů, o které se opírá tato práce (zejména v empirické části) tak vychází zejména z posledního vydání publikačního manuálu APA [2010], manuálu AERA [2006] a stanoviska ASA [Wasserstein & Lazar, 2016]. Tyto tři materiály (a k nim doprovodné komentáře) se v zásadě shodují že minimem při prezentaci kvantitativních výsledků, pokud autor používá statistické testy by mělo být následující:

1) plná prezentace výsledků statistických testů (tj. hodnota testového kritéria, případné počty stupňů volnosti a vypočtená hodnota statistické významnosti),

2) věcná interpretace získaných výsledků a

3) výpočet a interpretace měr věcné významnosti (effect size, srov. kapitolu 4 této práce).

Kromě takto pozitivně vymezených požadavků lze vymezit i některé negativně. Autoři by neměli při publikaci zaměňovat statistickou a věcnou významnost výsledků a neměli by používat mechanické postupy z oblasti statistického testování, např. hvězdičkování výsledků, stepwise postupy (více o problému se lze dočíst v kapitole 3 této práce). V zásadě identická doporučení lze nalézt i v mnoha časopiseckých článcích, z poslední doby pro oblast sociologie například v textu Bernardi a kol. [2017].

Samozřejmě lze nalézt i extrémnější doporučení, například zákaz používání statistických testů vůbec [Lotus, 1996; Trafimow, 2014; Trafimow & Marks, 2015], více se tématu věnuje třetí kapitola. Nicméně zde rozhodně nepanuje většinová shoda jednotlivých odborných komunit, i manuál APA konstatuje [APA, 2010], že v literatuře existuje diskuse o testech statistické významnosti a existují pochybnosti o jejich používání. Ke standardům pro účely této práce přiřazuji i požadavek na používání statistických testů pro data, která jsou pro tyto účely vhodná (více viz kapitola 2 této práce a dva exkurzy v páté kapitole). Zmíněné texty APA, AERA a ASA o této věci v zásadě mlčí (výjimkou jsou doprovodná stanoviska jednotlivých statistiků ke stanovisku ASA), nicméně např. kniha Kline, která je přímo doporučena manuálem APA a přímo APA vydána je v této věci zcela explicitní [Kline, 2013: 74]. S ohledem na empirickou možnost zkoumání standardů publikace tedy zaměřuji svou pozornost na tyto okruhy:

1) používání statistických testů pro data, která to umožňují,

2) absence mechanické aplikace použití statistických testů a záměn mezi statistickou a věcnou významností,

3) věcnou interpretaci získaných výsledků a

4) výpočet a interpretace měr věcné významnosti.

Tyto otázky jsou detailněji diskutovány v kapitole druhé až čtvrté a empiricky poté využity v páté kapitole na analýzu produkce publikované v časopisech Československá psychologie, Pedagogika a Sociologickém časopise.

Východiska a zakotvení práce

Svým zaměřením jde o práci na hranici několika disciplín resp. subdisciplín. Věcně práce vychází zejména z poznatků sociálněvědní statistiky a kvantitativní metodologie, nicméně svým směřováním a provedenými výzkumy z tohoto rámce částečně vystupuje a pohybuje se

též v oblasti sociologie vědy. I přes toto ne zcela jednotné oborové zaměření, které je ovšem dle názoru autora práce plně funkční, je práce vedena jednotným tématem a v zásadě i jedinou otázkou: „Používají sociální vědci a vědkyně při zpracování kvantitativních dat korektně postupy, které byly vyvinuty v rámci metodologie jejich disciplíny ale též v jiných disciplínách?“ Samozřejmě na takto obecně položenou otázku by nebylo možné poskytnout odpověď na ploše jedné disertační práce, proto bylo téma pragmaticky zúženo na:

1) české sociální vědy (konkrétně na sociologii, psychologii a pedagogiku),
2) konkrétní reprezentace těchto vědních disciplín v jejich předních časopisech (Sociologický časopis, Pedagogika resp. Československá psychologie) a
3) pouze na koncepce statistické a věcné významnosti, které jsou pro zpracování kvantitativních dat klíčové a jsou nejhojněji používány.

Úvodní myšlenky a vymezení cílů práce

Jak již bylo uvedeno zcela na počátku, hlavní výzkumnou otázku, kterou řeší tato práce, lze formulovat poměrně snadno:“ „Používají sociální vědci a vědkyně při zpracování kvantitativních dat korektně statistickou a věcnou významnost?“ Aby bylo možné tuto komplexní otázku zodpovědět pokud možno vyčerpávajícím způsobem[2], je nutné se zaměřit na několik relativně samostatných oblastí. Konkrétně se tak staly cílem práce tyto úkoly:

1) Zmapovat poznání v oblasti statistické významnosti, poznání limitů této koncepce a jejího možného chybného užívání (kapitola 2 a 3).
2) Zmapování v ČR téměř neznámé koncepce věcné významnosti a způsobů jejího měření (kapitola 4).
3) Zhodnotit skrze obsahovou analýzu používání statistické a věcné významnosti v českých sociálních vědách (kapitola 5).
4) Formulovat doporučení reagující na získané teoretické i praktické znalosti, přičemž se zaměřit na doporučení pro publikační praxi i výuku (kapitola 6).

Každý popsaný úkol je náplní příslušných kapitol (viz uvedené odkazy) a v rámci jednotlivých kapitol (zejména u kapitoly páté, založené na realizovaném výzkumu) je detailněji popsána metodologie použitá pro zpracování.

[2] Autor si je plně vědom skutečnosti, že zcela nelze žádnou složitější výzkumnou otázku zodpovědět, nadto odpověď je vždy poplatná době a kontextu, v němž byla poskytnuta. Ostatně autor plánuje i v budoucnu v práci na tématu pokračovat a dále jej rozvíjet.

Historie vzniku práce

Práce vychází ze zhruba 12-leté snahy autora věnovat se tématu statistické a věcné významnosti a jejich uplatňování ve vysokoškolské výuce a výzkumu[3]. Započetí práce na tématu lze datovat do roku 2003, kdy se autor práce začal zamýšlet nad tím, jak je běžně vyučována výzkumná metodologie a statistika a jak reálně poučky z těchto předmětů využívají studenti, ale i výzkumníci. Zhruba v této době vznikla první kostra článku nazvaného nakonec „Několik poznámek k jedné obsesi českých sociálních věd - statistické významnosti“ [Soukup, Rabušic 2007], který se dle reakcí a diskusí na následných seminářích stal velice používaným textem a posloužil k zamyšlení o současné praxi jak pro studenty, tak i výzkumníky. Na tento text navázaly další tématicky svázané články dílem publikované [Soukup, 2010, 2013, 2016], dílem dosud nepublikované (pátá kapitola disertace, text je ve druhém kole recenzního řízení). Lze tedy konstatovat, že předložená práce je založena na čtyřech článcích, z nichž tři již byly publikovány [Soukup, Rabušic 2007; Soukup, 2010; Soukup, 2013] a čtvrtý je v druhém kole recenzního řízení. Na tomto místě je korektní upozornit, že první článek byl vydán ve spoluautorství. Základní myšlenky tohoto textu pochází od autora disertace (Petr Soukup), spoluautor (Ladislav Rabušic) pomáhal tyto myšlenky precizovat, což odráží i pořadí autorů připsané u článku.

Několik poznámek ke struktuře práce

Práce se zaměřuje na statistickou a věcnou významnost, nicméně pozornost věnovaná těmto tématům není zcela symetrická. Na první pohled je věnováno více pozornosti statistické významnosti (kapitoly 2 a 3) než významnosti věcné (kapitola 4), nicméně tato proporce má svou logiku. Koncepce statistické významnosti je známější, proto není na rozdíl od věcné významnosti detailně vyložena, ale je jí věnována pozornost jako první. S ohledem na to, že se již desítky let diskutuje o omezené možnosti používat statistickou významnost a také o její chybné interpretaci ve výzkumnické praxi, jsou právě tato témata zařazena do kapitol 2 a 3. V následné části (kapitola 4) je pak detailně vyložena koncepce statistické významnosti a pozornost je zaměřena na představení klíčových měr věcné významnosti. Důvodem je skutečnost, že v českém prostředí jde o téma v zásadě neznámé. Empirická část, tvořená posledním článkem (kapitola 5), pak obě témata propojuje, neboť pojednává výsledky

[3] Tímto nemá být řečeno, že překládaná práce je výsledek systematické 12-leté práce autora, protože to by jistě bylo málo. Autor se v průběhu času věnoval aktivně výuce a výzkumu a publikoval více než patnáct odborných textů na zcela jiná témata.

výzkumu zaměřeného na užívání statistické a věcné významnosti v předních českých sociálně vědních časopisech (Československá psychologie, Pedagogika a Sociologický časopis) v období 2005-2014 (resp. pro Sociologický časopis pro období 1995-2014). Obě koncepce se pak promítají i do závěru (kapitola 6) do doporučení pro publikační (návrh základních standardů pro publikace) i výukovou praxi (z oblasti výzkumné metodologie a sociálněvědní statistiky).

Epistemologická východiska práce

Při zpracování stejného tématu různými autory, mohou vzniknout zcela odlišné texty zejména díky jejich odlišnému epistemologickému či ideovému zázemí. Je proto namístě přiznat východiska, o které se práce opírá a ukázat na autory, na jejichž ramenech[4] se autor pokouší „balancovat“. Z hlediska sociologických paradigmat je autorovi nejbližší pozitivistický přístup, tj. přístup, který do sociologie zavedl August Comte, rozvedl jej Emile Durkheim a navázal na něj zejména strukturní funkcionalismus vedený Talcottem Parsonsem a R. K. Mertonem. Pokud by měl autor volit jednoho ze sociologických teoreriků, který je mu myšlenkově nejbližší, byl by to jistě právě posledně jmenovaný Merton, který dokázal kombinovat teoretické poznání s empirickým výzkumem a je pokládán za autora koncepce teorií středního dosahu. Společně s pozitivistickým přístupem je autorovi blízké přesvědčení o možnosti využití matematických metod v oblasti sociálních věd, konkrétně sympatizuje s tzv. matematickou sociologií. Za jejího zakladatele můžeme považovat P. F. Lazarsfelda[5], který se svým celoživotním profesním působením zasloužil nejen o založení, ale i bouřlivý rozvoj této disciplíny[6]. Samozřejmě nelze podlehnout iluzi, že jedině matematika (případně statistika jako její aplikovaná oblast) se má využívat v sociologii (či obecně sociálních vědách). Ale dle přesvědčení autora jsou kvantitativní postupy v sociálních vědách používány oprávněné a umožňují získávat zjištění, která jinými způsoby získat nelze. Samozřejmě, že používání kvantitativních přístupů nemá být samoúčelné a ani chybné (právě na toto se zaměřuje práce samotná). V České republice jsme zejména v posledních 10 letech svědky neuvěřitelného nárůstu vědní produkce. Nelze též pominout bezprecendentní rozvoj výpočetní techniky a příslušného software, který umožňuje i podprůměrně znalému analytikovi zpracovávat objemné datové soubory sofistikovanými přístupy. Díky tomu jsou

[4] Tímto příměrem autor vědomě využívá textu R. K. Mertona [1965].
[5] V literatuře lze samozřejmě nalézt odkazy na starší inspirace a tak bývají za zakladatele tohoto přístupu někdy označováni statistici jako byl Quetelet či Condorcett.
[6] Dnes jde o svébytnou disciplínu se svým časopisem (vychází od r. 1971), konferencemi, kanonickými knihami. V Česku není dle autora nikdo, kdo by se této disciplíně systematicky věnoval.

často texty, které se halí do pláště vědeckosti, „vyráběny“ jako běžné produkty na pásu (proces Mcdonaldizace tedy zasáhl i českou vědu), často bez hlubšího rozmyšlení (jak před samotným výzkumem), tak při jeho provádění a jeho zpracování). Cílem předloženého textu je mj. varovat před postupy automatizované produkce výsledků skrze statistické postupy a vybídnout české studenty a výzkumníky k přemýšlení nad tím co dělají, a co jejich výzkumná data vlastně znamenají. Autor práce si je vědom omezenosti svých znalostí a zkušeností, proto již na počátku přiznává, že kromě soustavného studia literatury z oblasti kvantitativní metodologie a sociálně vědní statistiky, se snažil jednotlivé části práce konzultovat s odborníky z ČR i ze zahraničí na jednotlivá zde popsaná témata. Samozřejmě tyto konzultace nijak nezbavují autora odpovědnosti za případné chyby obsažené v této práci.

Metodologie práce

První tři články tvořící disertaci jsou převážně přehledovými studiemi stávající odborné literatury z oblasti kvantitativní metodologie a statistiky a kromě krátkých historických exkurzů nabízí přehled současných přístupů a doporučení. S ohledem na šíře jednotlivých témat jsou v některých případech problémy podány poměrně stručně. Poslední, empirická část (čtvrtý článek věnovaný výzkumu časopisů) pak zcela navazuje na tři předchozí a kategorie zavedené v předchozích textech využívá jako kategorie vhodné pro obsahovou analýzu provedenou na publikovaných textech (tam, kde nebylo toto empiricky možné, samozřejmě tyto kategorie využity nejsou).

S ohledem na skutečnost, že pátá kapitola popisuje empirický výzkum založený na obsahové analýze článků, je zde ve stručnosti popsána metodologie provedeného výzkumu, detaily jsou pak v páté kapitole.

Základním cílem bylo skrze obsahovou analýzu zhodnotit kvantitativní analýzy publikované v Sociologickém časopise v posledních dvaceti letech (2005-2014), těchto článků bylo celkem publikováno 162.

Pro všechny články všech zmíněných časopisů (celkem 363 textů) byla provedena kvantitativní 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 (zda je pro analyzovaná data možné použít statistické techniky), interpretaci výsledků (zda jsou korektně využívány doporučované míry věcné významnosti a nedochází pouze k mechanické práci s tabulkami a grafy).

Konkrétně byla pozornost zaměřena na tři oblasti:

I. používání statistických testů pro data, kde tyto využívat nelze,
II. nesprávná užívání statistické významnosti, zejména interpretační pochybení a mechanickou práci s daty a výsledky analýz,
III. věcnou interpretaci výsledků, používání měr věcné významnosti a jejich interpretaci.

Doplňkově bylo též 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. Nicméně s výjimkou Československé psychologie (kde se častěji objevují intervaly spolehlivosti), nebyl tento výskyt zaznamenán, a proto není ve výsledcích uváděn.

Seznam použitých zkratek a symbolů

AERA –American Educational Research Association

APA – American Psychological Association

ASA - American Statistical Association

d – označení Cohenova d, jedné z nejznámějších měr věcné významnosti, sloužící pro srovnání průměrů ve dvou skupinách

P – označení pro pravděpodobnost sloužící k vyhodnocení statistických testů (zjednodušeně statistická významnost)

Seznam literatury citované v úvodu

APA. 2010. Publication Manual of the American Psychological Association. 6th edition. Washington, DC : American Psychological Association.

AERA. 2006. Standards for Reporting on Empirical Social Science Research in AERA Publications

Bernardi, 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.

Kline, R. B. 2013. Beyond the statistical testing. Reforming data analysis methods in behavioral research,2nd edition. Washington, DC: American Psychological Association.

Loftus, G. R. 1996 Psychology will be a Much Better Science When We Change the Way We Analyze Data. Current Directions in Psychological Science vol.1: 161-171.

Merton, Robert K. (1965). *On the Shoulders of Giant.* Free Press.

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), s. 125-148. doi: 10.13060/23362391.2013.127.2.41.

Soukup, P. 2016. Užívání statistické a věcné významnosti v časopise Pedagogická orientace a Pedagogika v posledních deseti letech: pohled statistika. *Pedagogická orientace* .26(2):182-201.

Soukup, P., L. Rabušic. 2007. Několik poznámek k jedné obsesi českých sociálních věd – statistické významnosti. *Sociologický časopis/Czech Sociological Review*. 43 (2): 379-395.

Trafimow, D., M.Marks (2015) Editorial. Basic and Applied Social Psychology, 37(1): 1-2.

Trafimow, D. 2014. Editorial. Basic and Applied Social Psychology, 36(1): 1–2.

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.

2. Několik poznámek k jedné obsesi českých sociálních věd – statistické významnosti (Soukup, Rabušic)

Aforismus o statistice: There are three kinds of lies: lies, damned lies, and statistics (Disraeli).

K sepsání tohoto článku nás přivedlo poznání, že značná část českých sociálních vědců, nemluvě o značné proporci studentů, je posedlá statistickou významností. Testy statistické signifikance v jejich povědomí (neboť tak „pochopili" smysl testování v kurzech statistiky) slouží jako všemocné zaklínadlo. Jsou přesvědčeni, že bez testů statistických hypotéz není možné získat vědecky relevantní poznatky. Domnívají se, že tyto testy musí aplikovat na všechny výsledky bez ohledu na to, zdali jejich data pocházejí z pravděpodobnostního (náhodného) výběru, vyčerpávajícího zjišťování (z cenzu) nebo výběru nenáhodného (kvótního, záměrného, samovýběru). Jsou přesvědčeni, že testy významnosti jim řeknou, co je v datech důležitého, prostřednictvím nalezené statistické signifikance se snaží prokazovat těsnost vztahu dvou proměnných. Nic z toho ovšem statistická významnost neumí.

Cílem tohoto článku není hanět statistiku (to by asi bylo vzhledem ke vzdělání a zaměření obou autorů velmi zvláštní), ale upozornit na meze jejího používání a na její případné zneužívání zejména v (české) sociologii. V článku se věnujeme problematice používání statistických testů a podmínkám, při kterých je možné testy užívat. Naznačujeme i možnosti, jak si poradit, pokud statistické testy není možné použít, případně jak postupovat, pokud není statistické testování vůbec na místě. V neposlední řadě se krátce věnujeme i problematice práce s váženými daty, která jsou tak oblíbená, ale při práci s nimi se často chybuje. Při ukázkách se budeme snažit používat obecné postupy, někdy však uvedeme i konkrétní návod postupu v ČR zřejmě nejrozšířenějším paketu SPSS. Hned na počátku ovšem upozorňujeme čtenáře, že nelze na ploše jednoho článku popsat vše a že tento článek chápeme jako první vlaštovku, na kterou naváží další detailnější pojednání o jednotlivých problémech užívání statistické indukce.[7]

Začněme nejdřív s připomenutím samozřejmostí, na něž se ovšem v praxi často zapomíná. Teorie statistické indukce byla průvodně vyvinuta pro náhodné experimenty, posléze byla převzata pro data z náhodných (pravděpodobnostních) výběrů. U těchto dat slouží ke zobecňování výsledků z výběrového souboru na základní soubor. V základním modelové

[7] Partii statistiky, která pojednává o statistické indukci se zpravidla říká matematická statistika, v česky i anglicky psané literatuře se setkáme též s pojmy statistická inference či induktivní statistika. Neřešíme v tomto článku terminologické rozdíly mezi těmito pojmy a bereme je víceméně za synonyma.

situaci se předpokládá, že bude použita pro velké výběry[8] z velkých (ideálně nekonečných) základních souborů.

Nelze tedy bez dalšího její běžné postupy aplikovat v jiných situacích. Použití běžných postupů statistické indukce je podmíněno čtyřmi požadavky, které si postupně uvedeme, přičemž jsme si vědomi faktu, že v praxi je nesplnění těchto požadavků zpravidla propojeno.

1. Co znamená požadavek *velkého náhodného výběru*? Jde v podstatě o tři požadavky najednou. Jednak (1) aby měl výběr dostatečný počet jednotek, jednak (2) aby bylo provedeno vybírání ze základního souboru náhodně. Posledním, ale rozhodně ne nedůležitým, požadavkem je, (3) aby šlo o výběr. Předběžně lze poznamenat, že pouze výběry, kde je vybíráno alespoň v řádu desítek jednotek cca od 30–50, lze označit za „velké".[9] Náhodné výběry jsou pak jen ty, kde o vybrání či nevybrání jednotky rozhoduje náhoda, samozřejmě ve statistickém slova smyslu, kdy náhodou rozumíme souhrn drobných, ne zcela zjistitelných či zcela nezjistitelných příčin, které způsobují, že dopředu neumíme jistě stanovit výsledek náhodného pokusu. Provedení náhodného výběru se ovšem ve vědě řídí striktními pravidly. Důležité samozřejmě je, aby bylo vůbec vybíráno. Pokud máme data, která pocházejí z úplného zjišťování, nemůžeme o statistické indukci vůbec hovořit.
2. A co je to *velký základní soubor*? Protože většina reálně prováděných výběrů je prováděna jako výběr bez vracení (podobně jako například ve Sportce nelze v jednom tahu dvakrát vytáhnout totéž číslo), ale statistika používaná pro statistickou indukci vychází ze vzorců výběrů s vracením, je dobré zajistit, aby základní soubor byl alespoň 100krát větší než zamýšlený výběrový soubor.[10] Není-li toto splněno, lze samozřejmě statisticky testovat, ale s jinými než běžně dostupnými vzorci. V takovýchto situacích již zpravidla nelze užívat tolik oblíbené statistické balíky, popřípadě je nutno používat jejich speciální moduly, které nejsou běžně známé a uživatel je často nemá ani zakoupeny.

[8] Mnohé statistické testy lze užívat i pro malé výběry, nicméně pak je potřebné striktně splnit jejich předpoklady (typicky normalitu a homoskedasticitu), což bývá v sociálních vědách zcela výjimečné. Proto je zde cum grano salis vylíčena statistická inference (klasické testy jako t-testy, analýza rozptylu) jako procedura vhodná primárně pro velké výběrové soubory.

[9] Naši studenti se nás často ptají, jak velké výběrové soubory by měli mít pro své diplomové práce. Zde je každá rada drahá, nicméně určitý návod, s nímž se ztotožňujeme, podává Blaikie [2003: 166]: 300 je adekvátní, 500 je lepší a 1000 by bylo ještě lepší.

[10] Je to samozřejmě pouze orientační návod. Například pro dospělou populaci ČR ve věku 15 let a více (asi 7 miliónů osob) nemusíme mít výběrový soubor o velikosti 70 000 jednotek). A naopak, pro výzkum studentů nějaké fakulty, která má 3 000 studentů, by výběrový soubor o 30 jednotkách byl bezesporu velmi malý. Velikost výběru ovlivňují i další charakteristiky výzkumu, především hloubka analýzy, kterou chceme provádět.

Poté co jsme uvedli základní předpoklady pro používání běžných testů statistické indukce (řadíme sem zejména t-testy (jednovýběrový, dvouvýběrový, párový), analýzu rozptylu, chí-kvadrát test pro podíl (četnosti); obdobně platí naše závěry i pro užívání intervalů spolehlivosti (pro střední hodnotu, pro rozdíl dvou středních hodnot, pro podíl, resp. relativní četnost, apod.) se budeme věnovat jednotlivým případům, kdy jsou požadavky pro její použití nesplněny, a budeme se snažit ukázat, jak v takových případech postupovat. Pro doplnění také ukážeme některé nesprávně postupy, aby bylo možno se jich vyvarovat. Ještě před detailním rozborem různých problémů statistické významnosti si zaslouží tento pojem uvedení definice a nutné je definovat i pojem reprezentativity, který je s ním nedílně spojen. V literatuře lze nalézt mnohé velmi složité definice statistické významnosti, pro účely tohoto článku volíme definici od Blahuše [Blahuš 2000: 55], protože ten zřejmě jako jediný český autor podrobnější definici nabízí.

„Tvrzení, že výsledky jsou ‚statisticky významné' na hladině a = 0,05 má přesně následující význam.

a) U náhodného reprezentativního výběru znamená, že riziko zobecnění z náhodného reprezentativního výběru na celý základní soubor je nejvýše 0,05 (tj. 5 %). Tedy např. riziko, že v základním souboru studentů není procento spokojenosti vyšší než 50 %.

Jde o riziko tzv. chyby I. druhu, že nesprávně zamítneme statistickou nulovou hypotézu H0. Tj. zde hypotézu, že rozdíl mezi skutečným procentem spokojených v základním souboru a zadaným procentem 50 % je nulový. Jinak též, že chybně zamítneme hypotézu, že rozdíl mezi hodnotou u výběru (60 %) a pesimisticky předpokládanou možnou hodnotou v základním souboru (50 %) je jen náhodný. Tedy chybně učiníme závěr, že z výběru lze provést zobecnění (zde, zobecnění, že v souboru studentů je počet spokojených větší než 50 %)." Blahuš [2000: 55] k tomu dodává, že statistická významnost tedy znamená pouze, že výsledek je „‚statisticky zobecnitelný' z reprezentativního-randomizovaného výběru na základní soubor, a to se zvoleným rizikem"[11]. Nyní se pokusme o definici reprezentativity. V běžných učebnicích statistiky pouze najdeme konstatování, že je důležité, aby výběr byl reprezentativní pro možnost induktivního usuzování z výběru na základní soubor [např. Knoke, Bohrnstedt, Mee 2002: 15]. Zřejmě jedinou cestou, jak *dosáhnout reprezentativity,* je provést *náhodný výběr ze všech jednotek základního souboru* za situace, kdy *všechny jednotky mají stejnou pravděpodobnost vybrání*. [Knoke, Bohrnstedt, Mee 2002: 15]. Takto lze reprezentativitu zajistit více technikami, podrobněji se tématu věnujeme v části B.

Dále budou v článku detailněji popsány tyto situace:

A) úplná zjišťování,

B) nenáhodné výběry,

C) malé výběry,

[11] Blahuš podává definici i pro případ náhodných experimentů, ty jsou ovšem v sociologii řídké, proto tuto část definice neuvádíme.

D) výběry z malých populací,
E) velké výběry, agregace dat, mezinárodní datové soubory a
F) používání vah u náhodných výběrů.

Úplná (vyčerpávající) zjišťování (cenzy) a používání statistické indukce

Pokud neprovádíme výběr, ale máme informaci o všech jednotkách základního souboru, pak samozřejmě nepotřebujeme usuzovat na situaci v základním souboru a nepotřebujeme statistickou indukci. Předpokládejme, že jsme provedli šetření, v němž byl zjišťován průměrný čistý měsíční příjem ekonomicky aktivních obyvatel ČR. Zjistili bychom, že příjem mužů je vyšší než příjem žen. V takové situaci jde o skutečný rozdíl, není třeba žádného testu, abychom tuto skutečnost prokázali. V praxi je provádění úplných šetření málo časté, ale u malých základních souborů k němu může odcházet. Při používání statistické indukce bychom dospěli mnohdy k závěrům o *statisticky nevýznamných* rozdílech, ač o rozdíly ve skutečnosti jde.

S fenoménem úplných zjišťování jsou spojeny ještě dva dílčí problémy. První nastává, pokud provedeme úplné zjišťování, ale z nejrůznějších důvodů nezískáme informaci o všech jednotkách ve výběru. U šetření založených na dotazování hovoříme o míře návratnosti (*response rate*). Někteří autoři jsou toho názoru, že podmínkou aplikace statistické indukce je i vysoká návratnost – Blaikie [2003: 167] hovoří o 85 %, byť jedním dechem dodává, že v současnosti je naplnění tohoto požadavku velmi obtížné (v českých pravděpodobnostních výzkumech se pohybuje mezi 40–60 %). V této situaci rozhodně nelze postupovat tak, že budeme náš soubor považovat za výběr (navíc náhodný) ze základního souboru a budeme používat statistickou indukci. Toto pravidlo platí i v případě, že se nám podaří získat informaci i například jen o polovině jednotek. A co v tomto případě lze dělat? Nutné je za pomoci analýzy zjistit, z jakých důvodů a případně kdo neodpověděl. Poté je možné za pomoci postupů obdobných vážení (viz bod F) se pokusit napravit možná zkreslení. Obdobný problém nastává v případě, kdy máme malý základní soubor (řádově desítky až stovky jednotek), ale z určitých důvodů se rozhodneme nezjišťovat informace o všech jednotkách. Tento případ popisujeme v bodě D) a vysvětlujeme, jak postupovat při využívání statistické indukce.

Druhým problémem nebo spíše zapeklitostí je možnost zkoumání některých vztahů pomocí statistických testů u úplných zjišťování. V případě, že chceme zjišťovat, zda spolu souvisí dvě nominální proměnné, můžeme samozřejmě sestrojit kontingenční tabulku a spočítat koeficient kontingence. Z tabulky a hodnot koeficientu pak víme, zda proměnné spolu souvisí a případně, zda je souvislost výrazná (těsná), či nikoliv. Obdobně u dvou

pořadových proměnných můžeme spočítat Kendallovo Tau (nebo stále populárnější koeficient gamma), u kvantitativních proměnných pak Pearsonův korelační koeficient (pro předpokládané lineární vztahy). Namístě již ale nejsou testy, zda příslušný koeficient je nulový v základní populaci, či nikoliv. Například pokud vyjde z úplného zjišťování korelační koeficient mezi příjmem a počtem let vzdělání 0,48, jde o středně silnou souvislost a žádný test o nulovosti korelačního koeficientu nepřichází v úvahu. Korelační koeficient o velikosti 0,48 zjištěný ze základního souboru prostě nulový není.

Je tedy důležité ještě jednou připomenout, že v případě vyčerpávajících šetření není namístě užívat statistickou indukci vůbec.

Nenáhodné výběry

Jak již bylo uvedeno na počátku, byla statistická indukce a její jednotlivé postupy vyvinuty pro případ náhodných výběrů (nejdříve zejména pro nejjednodušší případ prostého náhodného výběru). Pro tento případ platí veškeré vzorce pro intervalové odhady, testy statických hypotéz. Co ale dělat, pokud nemáme náhodný výběr? Nejdřív si vůbec stručně připomeňme, jaké výběry můžeme v praxi provádět. Zpravidla se užívá následující dělení výběrů:[12]

náhodné	**úsudkové/nenáhodné**
- prostý	- kvótní
- vícestupňový	- záměrné atd.
- oblastní apod.	

Z uvedeného schématu vyplývá, že v praxi tolik oblíbený *kvótní výběr není výběrem náhodným*, a tudíž statistická indukce nemá při jeho používání místo. Samozřejmě expert, který provádí kvótní výběr, může na základě svých znalostí a zejména zkušenosti být schopen stanovit chybu takového výběru a používat tak „obdoby“ induktivních postupů. V praxi se však často bezhlavě (a obáváme se, že i bez vědomostí) na výsledky z kvótního výběru aplikují postupy statistické indukce. Což je samozřejmě špatně. Pochopitelně si nemyslíme, že je špatné používat kvótní výběry (spíše naopak), špatné je však u dat z takovýchto výběrů užívat induktivní statistiku bez rozmyslu.

[12] Podrobný popis jednotlivých variant výběrů přesahuje možnosti tohoto textu, zájemce odkazujeme zejména na texty Čermáka a Vrabce [Čermák, Vrabec 1998a, 1998b, 1999], na Thompsona [Thompson 2002] nebo klasický text Kishe [Kish 1965].

V souvislosti s typy výběrů je nutné připomenout, že náhodným výběrem není ani *samovýběr***,** tj. situace, kdy výběr jednotky závisí na jednotce samé a nikoliv na náhodě (příkladem nechť je anketa). Z tohoto důvodu u samovýběrů nemá induktivní statistika místo; dokonce žádné zobecňování na základní soubor nemá u samovýběrů místo!

Poslední poznámku k náhodnému výběru činíme z praktických důvodů. Je velice jednoduché teoreticky říkat, že nejlepší je náhodný výběr, protože pak lze užívat induktivní statistiku se všemi jejími kouzly, a odmítat všechny nenáhodné výběry. V praxi je však velice problematické náhodný výběr provést. Jak získat oporu výběru (tj. seznam všech jeho jednotek), chceme-li například provádět výběr celé dospělé populace v ČR? Shromažďování osobních údajů podléhá zákonu č. 101/2000 Sb., o ochraně osobních údajů, a je nadto nesmírně obtížné. Provést výzkum, kde bychom chtěli takovou oporu získat, by mohlo trvat také několik let. V praxi se nadto při realizaci náhodných výběrů (zpravidla vícestupňových) setkáváme s velkým podílem odmítnutých rozhovorů. Můžeme pak výběr, kde nám polovina respondentů odmítne, považovat za náhodný? Domníváme se, že nikoliv.

Jednou z možností, jak bez opory výběru provést náhodný výběr, je technika *náhodné procházky* (*random walk*),[13] která je v praxi občas užívána a nelze proti ní mít výrazné námitky. Jen ve stručnosti popišme, jak může probíhat. Tazatel vyráží z náhodně vybraného místa, například křižovatky ve městě. Má například určenu cestu, první ulici vpravo, pak třetí doleva a zde do třetího domu, 1. patra. Zde náhodně vybere byt a v tomto bytě dotazovanou osobu například metodou prvních narozenin. Obdobně pokud je malý základní soubor a těžko dosažitelný, lze použít *techniku sněhové koule* (*snow ball technique*). Ta spočívá v tom, že vybereme prvního respondenta a on nám sám doporučí dalšího. Jisté nebezpečí zde spočívá v tom, že některé charakteristiky jednotek výběru budou systematicky zkresleny, neboť lidé doporučují ty, které sami znají a kteří jsou (tak trochu) jako oni sami.

Malé výběry

Vzorce statistické indukce používané běžně ve statistice (a nadto vyučované běžně ve statistických kurzech) vycházejí z předpokladu, pokud naše proměnná nemá striktně normální rozdělení resp. není ve skupinách[14] zajištěna homoskedasticita, pak by výběr (resp. i jednotlivé podskupiny, za které děláme závěry) měl mít minimálně 30–50 jednotek. Samozřejmě může dojít k situaci, kdy takový výběr nemáme a v praxi zejména u

[13] Nezaměňujme toto označení se stejně označeným stochastickým procesem při modelování časových řad, i když podobnost lze jistě nalézt.

[14] Z těchto důvodů se zpravidla design výzkumu, zejména velikost výběru, odvíjí od požadavku na dostatečně velký počet respondentů ve skupinách, za něž mají být samostatně zjišťovány výsledky, popřípadě má dojít ke srovnání s výsledky jiných skupin. Bývá zvykem, že pro vyšší ‚kvalitu' výsledků se minimální vzorek pro jednotlivé skupiny stanovuje na cca 80–100 respondentů.

specifických podskupin k takové situaci nezřídka dochází. Důvodem této skutečnosti mohou být často peníze, protože zadavatel výzkumu rád získává co nejvíce informací za co nejméně peněz (na tom není nic špatného), a nutí tím výzkumníky snižovat velikost výběru. Nicméně použít v tomto případě klasické postupy statistické indukce není správné.

Jaké alternativy se nám nabízejí, chceme-li přeci jen čerpat z hluboké studnice poznatků induktivní statistiky?

1) Užití *speciálních testových statistik* (nebo jednodušeji testů[15]) vyvinutých pro malé výběry při nesplnění parametrických předpokladů.
2) Užití *neparametrických metod* s „přesnými" hodnotami testových kritérií.[16]

Ad 1) Přístupy založené na speciálních statistikách, odvozených pro malé výběry jsou spíše „hrátkami" statistiků a každý výpočet založený na těchto statistikách by bylo nutné provádět s kalkulačkou v ruce (popřípadě si naprogramovat vlastní proceduru), protože tyto postupy nejsou zahrnuty ve standardních statistických paketech. Z tohoto důvodu nelze tuto cestu pro zpracování větších úloh a počítání více analýz doporučit. Zájemce o tyto postupy lze odkázat na texty [Kahounová 2000; Řehák, Řeháková 1986: 62, 120].

Ad 2) Použití neparametrických metod s „přesnými" hodnotami testových kritérií je zřejmě v případě malých výběrů praktičtější variantou. Nutno poznamenat, že pro větší výběry má testové kritérium neparametrických testů (resp. funkce odvozená od tohoto testového kritéria) nejčastěji přibližně rozdělení normální nebo jiné běžně užívané rozdělení. Testování u větších výběrů probíhá porovnáním testového kritéria s kvantily běžně užívaných rozdělení. Pro malé výběry tato konvergence k běžně známým rozdělením neplatí a existují speciální tabulky [Anděl 2003, 2005; Blatná 1996; Likeš, Laga 1978], v nichž lze nalézt hodnotu, s níž porovnáváme vypočtené testové kritérium. Také statistické pakety zohledňují tento přístup a za pomoci simulačních metod nebo jiných přístupů umožňují provádět přesné testování. Zřejmě nejběžnější statistický software v ČR SPSS má samostatný modul nazvaný *Exact Tests*, který slouží k testování statistických hypotéz u neparametrických testů v případě malých výběrů. Jen pro připomenutí uveďme, že neparametrické metody slouží k testování hypotéz zejména v případech, kdy proměnné nemají požadovaný charakter (zejm. nejsou kardinální normálně rozdělené). Daní, kterou platíme za přechod od požadavku na kardinální normálně rozdělené proměnné k ordinálním proměnným, je nižší síla neparametrických testů (schopnost zamítnutí testované hypotézy za situace, že tato ve skutečnosti neplatí) oproti

[15] Každý běžný statistický test je založen na testové statistice, její hodnota se vypočítá z výběrových dat a srovná se s kvantily příslušného statistického rozdělení a učiní se závěr.
[16] Důležitá nejsou až tak přesná testová kritéria, ale důležité je že se u malých výběrů neužívá aproximace testových kritérií za pomoci nejběžnějších rozdělení (jako je zejména normované normální), ale využívá se simulačních postupů (resp. speciálních tabulek) k nalezení přesnějších hodnot, s nimiž má být porovnáno testové kritérium.

jejich parametrickým protějškům [Hendl 2004]. Známe neparametrické obdoby t-testů (jedno-, dvouvýběrového a párového) a analýzy rozptylu, korelačních koeficientů apod. Pro ilustraci špatného a správného testování rozdílu ve středních hodnotách uveďme příklad na srovnání průměrného příjmu mužů a žen v případě malého výběru (tabulka 1).

Tabulka 1. Rozdíly platů mužů a žen v malých výběrech[17]

Nesprávně použitý t-test poskytne tyto výsledky:

Dvouvýběrový t-test s nerovností rozptylů

	muži	*ženy*
Stř. hodnota	8644,615	6572,222
Rozptyl	6523877	8359444
Pozorování	13	9
t stat	1,73262	
P(T<=t) (1)	0,051196	
t krit (1)	1,745884	
P(T<=t) (2)	0,102391	
t krit (2)	2,119905	

Ženy ($n_1 = 9$), muži ($n_2 = 13$), vypočteno v Excelu

Zdroj: vlastní výpočty, náhodný výběr z dat ISSP 1999, n=22.

Závěr zní: *nulovou hypotézu nelze zamítnout*, nelze říci, že průměry příjmů mužů jsou vyšší než příjmy žen. Vypočtenou významnost 0,102 je nutno dělit dvěma, abychom získali jednostranný test, neboť naše alternativní hypotéza byla směrovaná (directional). Tedy 0,102/2 = 0,051. Podívejme se na výsledky Mann-Whitney testu dle přesných kritérií. Počítáme dle vzorců [Anděl 2003: 103]:

$U_1 = n_1 * n_2 + n_1 * (n_1-1)/2 + T_1,$

$U_2 = n_1 * n_2 + n_2 * (n_2-1)/2 + T_2,$

kde T_1, resp. T_2 je součet pořadových čísel hodnot pro ženy, resp. muže.

[17] Data použitá pro tento příklad jsou k dispozici na vyžádání u prvního autora.

Testové kritérium se určí jako menší z hodnot U_1 a U_2 a srovnává se s tabulkovými hodnotami. V našem případě je hodnota T1 = 72 a T2 = 181, z toho dopočteno U1 = 54 a U2 =27. V tabulkách [Blatná 1996: 203; Anděl 2003: 266] můžeme zjistit, že pro n_1=9 a n_2=13 je kritickou hodnotou pro jednostranný test na 5% hladině významnosti hodnota 33, pro 2,5% hladinu významnosti pak 28. *Nulovou hypotézu lze zamítnout i na 2,5% hladině* významnosti. *Můžeme vidět, že správná metoda (neparametrická přesná) dává zcela jiné výsledky než nepřesná metoda parametrická.*

Dodejme, že statistické programy umožňují často tyto přesné výpočty (například modul Exact v SPSS), případně za pomoci simulací (metodou Monte Carlo) lze stanovit interval spolehlivosti pro hladinu významnosti.

Výběry z malých populací

V některých výzkumech (zejména akademického charakteru) narážíme na skutečnost, že náš základní soubor je poměrně malý (pod tímto pojmem máme na mysli řádově stovky osob). Samozřejmě že v této situaci by bylo optimální udělat úplné zjišťování, ale to často není z finančních a časových důvodů možné. Přistoupí-li výzkumník k rozhodnutí, že z malé populace udělá výběr (často vzhledem k základní populaci dost velký) a chce zároveň používat postupy statistické indukce pro usuzování na celou populaci, je potřebné modifikovat běžně užívané postupy. Ještě než si přiblížíme tento postup podrobněji, věnujme se krátce problematice výběrů z malých populací.

V případě, že je prováděn výběr z malé populace, je ještě důležitější než v případě výběrů z velkých souborů, aby se jednalo o náhodný výběr. Jakékoliv samovýběry a jejich obdoby je nutno u malých základních souborů naprosto jednoznačně odmítnout. Problémem u výběrů z malých souborů (jejichž základní charakteristiky nejsou zpravidla známy) je skutečnost, že nelze posoudit reprezentativitu. Proto naléháme na požadavek striktně provedeného náhodného výběru, který by reprezentativitu měl zaručovat.

Nyní se opět vraťme k tomu, jak vypadá modifikace statistické indukce v případě výběrů z malých souborů. Předpokládejme, že stojíme před následujícím problémem: Byl proveden výběr o velikosti 150 (dále ve vzorcích symbol n = 150) ze základního souboru o velikosti 300 (dále ve vzorcích N = 300). Problémem, na který narazí výzkumník používající běžnou statistickou indukci obsaženou ve statistickém softwaru, bude skutečnost, že při použití dvouvýběrového testu (nebo analýzy rozptylu v případě více skupin) vychází rozdíly mezi skupinami jako nevýznamné, intervaly spolehlivosti (pro střední hodnoty a relativní četnosti) jsou poměrně široké atd. Důvodem je skutečnost, že vzorce pro klasickou

statistickou indukci vycházejí z předpokladu, že se provádí výběr s vracením,[18] i když v praxi se používá téměř výhradně výběr bez vracení.[19] Vychází se z poučky, že *konečnostní násobitel* (finite population correction factor), kterým se liší[20] vzorec rozptylu u výběru bez vracení a výběru s vracením, se v případě výběru z velké populace blíží svou hodnotou 1 a lze říci, že rozptyl u výběrů s vracením a výběrů bez vracení z velkých populací je v podstatě shodný (a využívá se pro výběry bez vracení jednodušších vzorců pro výběry s vracením). Tento poznatek ale neplatí u výběrů z malých populací, kdy vybíráme podstatnou část základního souboru (i v námi uvedeném příkladě, kde n/N = 1/2, tedy je vybrána jedna polovina základního souboru) a rozptyly výběrů s vracením a bez vracení se liší. A jak vypadá vzorec konečnostního násobitele?

$$K = \frac{(N-n)}{N-1},$$

kde N je velikost základního souboru a n velikost výběru.

Vypočteme-li hodnotu K v námi uvedeném příkladu, dosazením do vzorce získáme hodnotu přibližně rovnou 1/2. Touto hodnotou musíme korigovat rozptyl v klasických vzorcích statistické indukce, v našem příkladě rozptyl snížíme o jednu polovinu. Ukažme si na našem smyšleném výběru 150 respondentů ze 300 na příkladu užití konečnostního násobitele.

Příklad: Určete 95% intervalový odhad podílu (relativní četnosti) osob, které byly ve vězení déle než 10 let, když z výběru jste získali bodový odhad relativní četnosti hodnoty p = 0,3 (resp. po vynásobení jedním stem 30 %). Připomeňme, že máme základní soubor o velikosti 300 a z něj vybíráme 150 osob.

Výpočet

Nejdříve počítejme, jak je běžné:

Intervalový odhad rel. četnosti = $p \pm u_{0,975} * \sqrt{(p*(1-p)/n)}$ kde $u_{0,975}$ je 97, 5% kvantil normovaného normálního rozdělení (který jak známo má hodnou přibližně 1,96).

Výsledný intervalový odhad po dosazení do vzorce je mezi 0,23–0,37 (tedy mezi 23% a 37%).

[18] Postup, kdy jakoby pomyslně vybíráme z osudí a jednotka (např. respondent) po vybrání je vrácena zpět do osudí a může být vybrána i vícekrát.
[19] Postup, kdy jakoby pomyslně vybíráme z osudí a jednotka (např. respondent) po vybrání není vrácena zpět do osudí a nemůže být vybrána i vícekrát.
[20] Přesnější je vyjádření, že konečnostní násobitel ukazuje, kolikrát se liší rozptyl u výběru bez vracení oproti výběru s vracením.

Jelikož ale máme výběr z malé populace, musíme korigovat rozptyl konečnostním násobitelem. Rozptyl je ve vzorci vyjádřen výrazem pod odmocninou, vzorec výše uvedený můžeme korigovat násobením odmocninou z konečnostního násobitele, pro úplnost uveďme celý vzorec:

Intervalový odhad rel. četnosti = $p \pm u_{0,975} * \sqrt{(p*(1-p)/n)}*\sqrt{K}$

Výsledný intervalový odhad je mezi 0,25 a 0,35 (25 % – 35 %), je užší. Konkrétně je užší násobkem odmocninou z jedné poloviny (konečnostního násobitele). V našem případě činí zhruba 70 % původního intervalu.

Poznamenejme, že obdobný postup můžeme uplatnit i pro interval spolehlivosti pro průměr (například pro průměrný příjem domácnosti), popřípadě pro rozdíl mezi průměry dvou skupin (výběrů) a pro klasické testy srovnávající průměry (t-testy).

Zobecníme-li výše uvedené, máme-li výběr z malého základního souboru, spočteme dle výše uvedeného vzorce konečnostní násobitel, resp. odmocninu z něj. Poté postupujeme následovně:

1) V případě intervalového odhadu zkrátíme interval vynásobením odmocninou z konečnostního násobitele. Náš příklad byl ukázán na oboustranném intervalovém odhadu, ale samozřejmě korekci odmocninou konečnostního násobitele můžeme použít i u jednostranných intervalových odhadů.
2) V případě statistických testů vynásobíme testové kritérium převrácenou hodnotou odmocniny z konečnostního násobitele, zvýší se hodnota testového kritéria. Tu pak budeme muset porovnat s hodnotou kvantilu příslušného statistického rozdělení a učinit závěr o zamítnutí-nezamítnutí testované hypotézy.[21]

Velké výběry, agregace dat, mezinárodní datové soubory

V některých případech je výzkumník ve zdánlivě dobré situaci, protože má k dispozici velký výběrový soubor. V případě výběrů v řádu tisíců pak již na první pohled téměř u všech charakteristik vychází významné rozdíly (které mohou být například u průměrů na 5bodových škálách na úrovni menší než 0,1), téměř všechny závislosti měřené nejrůznějšími koeficienty jsou významné apod. I taková je odvrácená tvář statistické indukce. Výzkumník může mít radost, že nalézá rozdíly a závislosti, ale je opravdu právě toto zjištění jeho cílem? Nemá jít spíše než o statisticky významné rozdíly o rozdíly věcné navíc určité velikosti?

[21] Staromilci budou hledat v tabulkách, pokrokáři pak v tabulkových kalkulátorech anebo ještě lépe ve statistických paketech, kde dokonce mohou využít luxusu hledání hladiny významnosti, na které je ještě přípustné zamítnout testovanou (nulovou) hypotézu (to, co se značí jako Sig., P, P-level apod. v běžných výstupech).

Samozřejmě že ano, a proto cílem tohoto oddílu má být varování před svody, že výsledky jsou automaticky dobré, pokud vychází jako statisticky signifikantní.

Ukažme si, jak lze relativně uměle dosáhnout velkých výběrů včetně popsaných efektů typu „vše souvisí se vším, každý rozdíl je významný“. V komerční praxi často užívané (u tzv. kontinuálních výzkumů) spojování dat může přinést kýžené výsledky. Provádíme-li měření za pomoci stejného dotazníku každý týden, měsíc apod., není nic snazšího než data z různých okamžiků spojit a začít testovat výsledky na spojených datech. Problém je, že měření z různých časových okamžiků mohou být zatížena různými chybami, které nevědomky sčítáme, takže pak zdaleka neplatí pravidlo statistické indukce o poklesu chyby s nárůstem velikosti výběru. Dalším přítomným fenoménem samozřejmě může být časový vývoj sledovaného ukazatele pouze v některé ze sledovaných skupin apod. Otázkou také je, co nám říká například výsledek o odlišnosti spokojenosti mužů a žen s jejich mobilním telefonem na datech spojených za poslední tři roky. Proto než bezhlavě testovat na spojených datech za dlouhá časová období je často lepší testovat spíše na kratších úsecích například neparametrickými metodami.

Obdobným nešvarem s možná ještě horšími výsledky je testování na datech spojených z několika „různých“ výzkumů. V zásadě se nabízejí tyto možnosti:

1) data z jednoho výzkumu, která byla sebrána několika institucemi (příkladem v ČR může být Media projekt),

2) data z jednoho výzkumu sbírána v různých zemích (příkladem může být projekt EVS, WVS, ISSP apod.),

3) data spojená ad-hoc výzkumníkem z několika projektů, které obsahují tytéž otázky (často měřené v nejrůznějším čase).

Ponechme teď stranou variantu 3, která se podobá variantě uvedené v předchozím odstavci (částečně ale i variantám 1) a 2)) a zaměřme se na varianty 1) a 2), jež jsou si v mnohém podobné. Co je problémem takového spojování? Předně skutečnost, že zpravidla každý subjekt, který sbírá data, se dopouští určitých systematických chyb, jež se při spojování dat mohou jen umocnit. Navíc v případě nekvality jednoho ze spojovaných datových souborů je ihned tato nekvalita zanesena do celkových dat. Ještě větší problém souvisí s různými postupy při výběrech v různých organizacích a případném následném vážení. Zejména v mezinárodních projektech dochází často k situaci, kdy výsledný soubor má u některých zemí váhovou proměnnou, u jiných ji však nemá.

Výhodou spojených souborů je samozřejmě efekt výše popsaný, tedy vše na sobě závisí, vše se od sebe liší. Ale má například opravdu smysl testovat rozdíly mezi muži a ženami na evropské/celosvětové úrovni? Osobně se domníváme, že daleko lepší je provést dílčí analýzy na národních úrovních a pak postupy metaanalýzy a dospět ke všeobecným závěrům. Při

práci se spojenými daty je také zapotřebí víc než kdy jindy kontrolovat, zda jsou data v pořádku, a to nejen na úrovni spojeného souboru, ale zejména na úrovni jednotlivých spojovaných souborů. Ze všech těchto důvodů proto varujeme před podlehnutím kouzlu spojování souborů a před svodem laciného získávání statisticky významných výsledků, které jsou však ve skutečnosti dosti *bezvýznamné*.

Aby bylo ukázáno, že statistika myslí i na případy, které zde popisujme, uveďme si, že *s daty spojenými za relativně homogenní skupiny* (za něž lze považovat i jednotlivé země) *lze pracovat za pomoci víceúrovňových modelů* (hierarchických modelů[22]). Používání těchto modelů zatím není v české sociologii samozřejmostí [Soukup 2006; Hamplová 2005 a Hendl 2004], více nalezneme v zahraniční literatuře [např. Hox 1995, 2002; Norušis 2004; Raudenbusch, Bryk 2002].

Používání vah u náhodných výběrů

Než si ukážeme, kdy se nesprávně používá statistická indukce na vážených datech, zkusme se krátce zamyslet nad tím, kdy k vážení dochází a zda je třeba tuto proceduru užívat. Pokusme se nalézt společné rysy případů, ve kterých výzkumník zkouší vytvořit váhu a posléze ji aplikuje na svá data a pracuje s váženými daty. V praxi se nejčastěji vyskytují tyto dva případy:

1) úprava výběru takovým způsobem, aby jeho vybrané (zpravidla demografické) charakteristiky (resp. proporce z hlediska těchto charakteristik) odpovídaly hodnotám (proporcím) v základním souboru.
2) Spojení souboru ze „základního“ výběru s dodatkovým výběrem (tzv. boostem).

Samozřejmě, že někdo může namítnout, že případ 2) je speciálním případem varianty 1) (a někdy i jejím důsledkem), ale pro jeho odlišnou logiku si o něm řekneme zvlášť.

Ad 1) Jak vypadá tvorba váhy u 1. varianty?[23] Vše demonstruje tabulka 2:

[22] Anglická terminologie užívá pojmů multilevel modelling, hierarchical modelling, ale např. v ekonometrické literatuře random-coefficient model apod. Přehled nejrůznějších pojmů pro tyto modely lze nalézt v Raudenbusch, Bryk [2002: 5–6].

[23] Omlouváme se za poněkud triviální výklad této problematiky, ale z edukativních důvodů uvádíme problematiku práce s váženými dat popisem procesu vzniku vah. Čtenáře znalé těchto postupů prosíme, nechť přeskočí tyto triviální popisy a pokračují četbu dalším odstavcem odhalujícím úskalí práce s váženými daty postupy statistické indukce. Totéž platí i pro dále uvedený popis postupu ad 2).

Tabulka 2. Ukázka vážení dat podle jedné proměnné

	ZŠ	SŠ bez maturity	SŠ s maturitou	VŠ	Součet
I. Struktura výběru	21%	40%	35%	4%	100%
II. Struktura základní soubor (ČSÚ)	23%	41,5%	28%	7,5%	100%
III. váha pro jednotlivé skupiny (II/I)	1,095	1,0375	0,800	1,875	X
IV. (I*III) váha*počet ve výběru	23%	41,5%	28%	7,5%	100%

Zdroj: vlastní smyšlený výběr, proporce základního souboru ČSÚ 1999.

Výzkumník stanoví váhy pro jednotlivé respondenty dle jejich vzdělání, tak aby po spuštění váhy (při práci s váženými daty) odpovídaly proporce údajům z ČSÚ. Až potud je postup zcela správný. Samozřejmě za předpokladu, že výběr byl opravdu náhodný, o čemž se lze přesvědčit zejména za pomoci chí-kvadrát testů dobré shody, které umožňují testovat hypotézu o náhodném vychýlení struktury ve výběry [Herzmann et al. 1995: 93; Anděl 2005: 271; Zvára, Štěpán 2002: 195].

Pokud bychom statistickou indukci používali jen pro zobecnění výsledků za celek na celou populaci, byla by práce s váženými daty v pořádku. Problém ale nastává v okamžiku, když například budeme v našem případě zkoumat rozdíly mezi vzdělanostními skupinami. Budeme-li v takovém případě pracovat s vahou, počítáme s jinými (umělými) počty respondentů v jednotlivých skupinách (například u vysokoškoláků cca 1,9krát vyššími) a výsledky statistických testů najednou mohou být významné jen díky tomuto umělému navýšení. Namístě je testování bez spuštěné váhy. Docházíme tak k zesložitění práce s váženými daty, pro jednu úlohu váhu užijeme, pro jinou nikoliv. Situace je v praxi ještě komplikovanější, protože váha se zpravidla nestanoví jen za pomoci jedné charakteristiky (proměnné), ale za pomoci více charakteristik. Opomíjíme složitý praktický problém, kde získat sdružené distribuce těchto více charakteristik za celou populaci (každý, kdo se o to pokoušel, o tom ví své). I z těchto důvodů bychom rádi problematice vážení a možností statistické indukce u vážených dat věnovali samostatný navazující článek.

Ad 2) Věnujme ještě krátce pozornost situaci, kdy jsme provedli výběr a poté ještě dodatečný výběr[24] například jen jedné skupiny (například osob s VŠ vzděláním). Motivací

[24] Zde narážíme na meze české terminologie výběrových šetření, logicky by se nabízel pojem dovýběr, ale to je pojem užívaný pro doplňování počtu respondentů v případě, že máme méně respondentů, než jsme zamýšleli a rozhodujeme o něm ex-post. Dodatečný výběr málo zastoupených skupin je naopak zvolen zpravidla již na

těchto dodatečných výběrů je fakt, že určité skupiny jsou málo zastoupeny, a proto uměle navýšíme jejich zastoupení ve výběru. Jak konstruujeme váhu v tomto případě? Zjistíme strukturu z hlediska relevantní charakteristiky v základním výběru (který musí být samozřejmě náhodný a reprezentativní) a váhu stanovíme tak, aby tato struktura byla zachována i na datech po sloučení základního a dodatečného výběru. Zatímco v případě označeném 1) jsme na výběr aplikovali strukturu základního souboru, nyní aplikujeme na spojená data ze dvou výběrů struktur výběrového souboru. Samozřejmě tomuto kroku může předcházet aplikace postupu dle 1) na základní výběrový soubor, nicméně toto „dvojí“ vážení spíše nedoporučujeme.

A nyní opět zauvažujme nad užitím statistické indukce u takto vážených dat. I zde platí, že chceme-li usuzovat na výsledky za celou populaci, užijeme váhu (srovnatelné výsledky bychom měli získat i v případě, že použijeme data bez váhy bez dodatečného výběru). V případě, že chceme testovat rozdíly mezi vzdělanostními skupinami, měli bychom opět pracovat s neváženými daty (i když v tomto případě bychom se nedopustili takové chyby, jako v případě 1), neboť u skupin zahrnutých v dodatečných výběrech bychom pracovali s nižšími počty respondentů a méně rozdílů by bylo statisticky významných, což je jistě méně nebezpečné). Opět uveďme, že v praxi se situace zpravidla komplikuje, protože není prováděn jeden dodatečný výběr, ale je jich prováděno více. Vážení je v takovém případě logicky složitější a práce s váženými/neváženými daty samozřejmě také.

Shrnutí poznatků a výzvy do budoucna

Poté co jsme nastínili možné konkrétní problémy, ke kterým může docházet při nesprávném mechanickém užívání klasických postupů statistické indukce, ještě dodejme, že problémy spojené se statistickou významností tímto nekončí. V metodologické literatuře najdeme mnoho výtek proti konceptu statistické významnosti, radiální autoři dokonce navrhují přestat tento koncept užívat. Jsme si vědomi, že není možné v jednom článku tuto diskusi plně postihnout, proto připravujeme článek, který bude doplňkem tohoto textu a jehož cílem bude ukázat na obecné meze statistické významnosti a také na koncepci věcné významnosti a jejího měření.

Závěrem jen poznamenejme, že věcná významnost a možnosti jejího měření zatím není standardním obsahem statistických, ale ani obecnějších metodologických učebnic. Nicméně to není důvodem, abychom se tématu dále nevěnovali, nebo ho dokonce zcela vytěsňovali z metodologických diskusí.

---

počátku výzkumu (tedy ex-ante) a provádí se společně se základním výběrem. Z tohoto pohledu není možná adjektivum „dodatečný“ příléhavé a bylo by lépe užít adjektivum „doplňkový“.

Závěr

V tomto článku jsme se zabývali problematikou statistické indukce a jejích možností. Vyšli jsme z našich pedagogických i výzkumnických zkušeností, které ukazují, že tato problematika je jednou z nejhůře pochopených pasáží statistiky. Četba zahraničních učebnic analýzy dat pro sociální vědce naznačuje [viz např. Blaikie 2003; de Vaus 2002; Field 2005], že v tom nejsme (naštěstí) tak úplně sami – a Blahušův [Blahuš 2000] nabádavý článek zase ukazuje, že ani badatelé ve vědách, které mají blízko k vědám přírodním, na tom nejsou o mnoho lépe. V článku varujeme před (českou) obsesí používat postupy statistické inference vždy, za každou cenu a bez ohledu na typ dat, která analyzujeme.

Z důvodů jisté problematičnosti celého konceptu statistické indukce v sociálních vědách někteří zahraniční autoři radikálně navrhují tyto postupy zcela vypustit ze statistické analýzy. Nejdeme tak daleko, avšak nabádáme k jejich uvážlivé aplikaci. Na to, že v nereflektované aplikaci statistické indukce může být ještě hlubší problém, poukazuje Field, jehož dlouhou citací náš článek uzavíráme.

> „Většina statistik používaná v sociálních vědách je založena na lineárních modelech. Většina výsledků publikovaných v časopisech jsou statisticky signifikantní výsledky. Jelikož se sociálněvědní badatelé většinou učili používat technik, které jsou na těchto lineárních modelech založené, znamená to, že publikované výsledky jsou ty, které lineární modely využily. Což znamená, že data a vztahy, které mohou být zpracována na základě nelineárních modelů, jsou povětšinou mylně ignorována – mylně proto, že na nelineární data byly aplikovány lineární přístupy, takže výsledky badatelům ‚nevyšly'… Je proto možné, že poznatky v některých oblastech vědy se vyvíjejí zkresleně" [Field 2005: 22].

Literatura citovaná ve 2. kapitole

Anděl, J. 2003. *Statistické metody*. Praha: Matfyzpress.

Anděl, J. 2005. *Základy matematické statistiky*. Praha: Matfyzpress.

Blahuš, P. 2000. „Statistická významnost proti vědecké průkaznosti výsledků výzkumu.“ *Česká kinantropologie* 4 (2): 53–72.

Blaikie, N. 2003. *Analyzing Quantitative Data*. London: Sage.

Blatná, D. 1996. *Neparametrické metody*. Praha: Vysoká škola ekonomická.

Čermák, V., Vrabec, M. 1998a. *Teorie výběrových šetření* – část 1. Praha: Vysoká škola ekonomická.

Čermák, V., Vrabec, M. 1998b. *Teorie výběrových šetření* – část 2. Praha: Vysoká škola ekonomická.

Čermák, V., Vrabec, M. 1999. *Teorie výběrových šetření* – část 3. Praha: Vysoká škola ekonomická.

Field, A. 2005. *Discovering Statistics Using SPSS*. London: Sage.

Hamplová, D. 2005. „Základní principy víceúrovňových modelů.“ *SDA Info* 7 (2): 1–2.

Hendl, J. 2004. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Praha: Portál.

Herzmann J., I. Novák, I. Pecáková. 1995. *Výzkumy veřejného mínění*. Praha: Vysoká škola ekonomická.

Hox, J., J. 1995. *Applied Multilevel Analysis*. Amsterdam: TT-Publikaties.

Hox, J., J. 2002. *Multilevel analysis: techniques and applications*. Mahwah (N.J.): Earlbaum.

Kahounová, J. 2000. *Praktikum k výuce matematické statistiky I.* Praha: Vysoká škola ekonomická.

Knoke, D., G. W. Bohrnstedt, A. P. Mee. 2002. *Statistics for Social Data Analysis*. Belmont (CA): Wadsworth/Thomson Learning.

Kish, L. 1965. *Survey Sampling*. New York: Wiley-Interscience.

Likeš, J., Laga, J. 1978. *Základní statistické tabulky*. Praha: Státní nakladatelství technické literatury.

Norušis, M. 2004. *Advanced Statistical Companion*. SPSS. Upper Saddle River (N.J.): Prentice Hall.

Raudenbush, W. S., A. S. Bryk. 2002. *Hierarchical linear models: applications and data analysis methods*. London: Sage.

Řehák, J., B. Řeháková. 1986. *Analýza kategorizovaných dat*. Praha: Academia.

Soukup, P. 2006. „Proč užívat hierarchické lineární modely.“ *Sociologický časopis* 42 (5): 987–1012.

Thompson, S. K. 2002. *Sampling*. New York: Wiley-Interscience.

Vaus D. A. de. 2002. *Analyzing Social Science Data*. 50 Key Problems in Data

Analysis. London: Sage.

Zvára, K., J. Štěpán. 2002. *Pravděpodobnost a matematická statistika*. Praha: Matfyzpress.

3. Nesprávná užívání statistické významnosti a jejich možná řešení

„Všechny modely jsou špatné.“ (George E. P. Box)

Úvod

V poslední době se díky lepšímu programovému vybavení stále více setkáváme s mechanickou aplikací statistických metod založenou na pouhém uvádění statistické významnosti bez hlubšího porozumění modelu a bez řádné interpretace výsledků. Statistická významnost slouží jako mocné zaklínadlo mnohých výzkumníků: je-li výsledek statisticky významný, je prý vše v pořádku. Je tomu tak ale doopravdy? Je skutečně statistická významnost samospasitelným lékem, který nám zajistí vědeckost? Cílem tohoto článku je ukázat, že rozhodně nikoliv. Ostatně na to nás upozorňují statistici i metodologové již zhruba 80 let. Při užívání statistické významnosti bychom si měli být vědomi všech jejích nedostatků (o tom je pojednáno na počátku článku). Poté jsou vyloženy statistické alternativy (doplňky) ke statistické významnosti, tedy jisté nástroje v rámci paradigmatu. V závěru se pak pokoušíme o shrnutí doporučení, která vedou ke korektnímu využívání statistické významnosti či příbuzných koncepcí.

Jsme si vědomi toho, že článek neobsahuje některé důležité problémy, které souvisí se statistickou významností. Konkrétně se jedná o výklad a případné srovnání Bayesovské statistiky s klasickými testovacími postupy. Obdobně není rozvedena metoda bootstrapu a resamplingu, která se prosazuje zejména v posledních letech. S ohledem na skutečnost, že tyto postupy nejsou v české sociologii příliš známé, budou obsaženy v samostatných článcích navazujících na tento text. Výjimkou je popis informačních kritérií, který je v článku obsažen.

Statistická významnost a její slabiny

Stručná historie

Statistická významnost resp. testování nulové hypotézy (v angličtině null hypothesis statistical testing nebo jen NHST) je velmi staré. Původním autorem myšlenky je zřejmě John Arbuthnott [1710], který svým testem mínil prokázat existenci boží prozřetelnosti. Dnešní vědci si kladou cíle méně smělé[25] a využívají koncepce, kterou rozpracovali statistici ve 20. a 30. letech minulého století, zejména Ronald Fisher, Jerzy Neyman a Egon Pearson. Jejich odkaz dodnes využívají statistici a vědci ve všech empirických oborech. Mnozí pokládají tento koncept za vynikající, naopak lze nalézt i odpůrce, kteří navrhují vymícení statistické významnosti z vědy. Detailnější pozornost koncepci Ronalda Fischera resp. Neymanovo-Pearsonovskému pohledu je věnována v dalším textu.

Definice statistické významnosti

Nejdříve si připomeňme postup testování hypotéz, který ukazuje na definici statistické významnosti. Začněme definicí pojmu statistické hypotézy: **Hypotézou rozumíme tvrzení o rozdělení pozorované náhodné veličiny**. Důležitý zvláštní případ nastává, je-li rozdělení výběrové statistiky známé: v takovém případě lze **hypotézu formulovat přímo jako tvrzení o hodnotě parametru příslušného rozdělení**[26, 27]. Je nutné zdůraznit, že **hypotéza se týká základního souboru**, z něhož jsme vybírali nebo který experimentálně zkoumáme (např. všech dospělých osob v Česku), **zatímco její testování se odehrává pouze na vybraných jedincích**, které jsme skutečně zkoumali. **Smyslem testování je správně zobecnit z vybrané podmnožiny (výběru) na celek**.

Klasicky formulujeme vždy dvě hypotézy o situaci v základním souboru: takzvanou **nulovou** (H_0) a k ní opačnou **alternativní** (H_1) hypotézu. Nulová hypotéza se tak nazývá proto, že

[25] Arbuthnott se snažil prokázat boží prozřetelnost skrze odhalování zákona vyrovnávajícího počet narozených mužů a žen a prokazoval také nepřirozenost polygamie z tohoto zákona plynoucí. Statistická významnost je tedy de facto původně sociálněvědním konceptem, ač by to zřejmě dnešní matematici či statistici nepřiznali.
[26] Například lze předpokládat (tj. utvořit hypotézu) o tom, že průměrný příjem je 23 tisíc Kč, politická strana má podporu 25 % voličů.
[27] Ve statistické teorii se odlišují hypotézy jednoduché a složené (platí jak pro nulovou tak alternativní), alternativní hypotézy pak mohou být jednostranné i dvoustranné, v detailech odkazujeme na Anděl [2003, 2005].

obvykle tvrdí, že proměnné v populaci na sobě nezávisí nebo že mezi skupinami v populaci nejsou rozdíly (v průměrech, mediánech apod.). Předpokládá tedy nulovost efektu vyvolaného nějakým zásahem (např. žádný rozdíl mezi skupinami experimentálních objektů, s nimiž se různě zachází). Takto se totiž původně objevila v dílech sira Ronalda Fishera o vyhodnocování biologických experimentů, (viz dále). Alternativní hypotéza oproti nulové hypotéze naopak tvrdí, že existuje nějaký (zobecnitelný) rozdíl mezi sledovanými skupinami, případně že dva fenomény spolu souvisí. Testování statistických hypotéz provádíme tak, že z výběrových dat je vypočtena testová statistika a na základě porovnání s kvantily rozdělení této statistiky (za předpokladu platnosti nulové hypotézy) se zjistí, zda je na dané hladině spolehlivosti možno nulovou hypotézu zamítnout.

Koncept testování byl původně vyvinut pro pravé experimentální uspořádání, dnes však je hojně využíván i pro data pocházející z kvaziexperimentů (není zaručeno náhodné zařazení jedinců do skupin podrobených různým experimentálním „ošetřením“ – například když se účastníci mohou sami rozhodnout, zda se experimentu podrobí) a pseudoexperimentů (neprovádí se „ošetření“, nelze tedy použít kontrolní skupiny – například standardní jednorázové sociologické výzkumy). Nejenže pak není úplně vhodná původní terminologie, ale i při interpretaci výsledků je potřeba uvážit odlišné okolnosti vzniku dat.

Interpretace statistické významnosti

Klasicky se při výkladu v učebnicích statistiky uvádí rozhodovací tabulka uplatňovaná při testování hypotéz (viz tabulka 1), nicméně většinou se jejímu obsahu a významu nevěnuje patřičná pozornost.

Tabulka 1. Platnost hypotéz o základním souboru a možná rozhodnutí na základě testování[28]

	Rozhodnutí	
Platí	**H_0**	**H_1**
H_0	OK (P=1- α)	Chyba prvního druhu (P= α)
H_1	Chyba druhého druhu (P= β)	OK (P= 1-β) Síla testu

[28] V tabulce užíváme nejběžnějšího značení, symbolu H_0 pro nulovou hypotézu a H_1 pro hypotézu alternativní.

Jak je vidět z tabulky, rozhodnutí testu o hypotéze nemusí být vždy v pořádku. K chybě prvního druhu dochází, když je nulová hypotéza zamítnuta, přestože H_0 platí. Obdobně chyba druhého druhu nastává, když nulová hypotéza zamítnuta není, přestože neplatí. Kvalita testu je dána pravděpodobnostmi, s jakými tyto chyby mohou nastat (α a β v tabulce 1). Platí, že pro daný výběrový soubor obvykle nelze současně minimalizovat pravděpodobnosti obou druhů chyb. Z tohoto důvodu se statistici rozhodli omezit riziko chyby prvního druhu na rozumnou velikost, nejčastěji na 5 % (α=0,05); tuto hodnotu zavedl do statistického diskurzu Ronald Fisher [1925, 1926] (více v další části "Nedostatky" statistické významnosti). Zamítání nulové hypotézy se tedy děje nejčastěji s 5 % rizikem, tj. stanovujeme pravděpodobnost zamítání nulové hypotézy při její platnosti v základním souboru na maximální hodnotu 0,05. Protože chybu druhého druhu nemáme jasně pod kontrolou, volíme v případě, že nedokážeme na základě hodnoty testové statistiky zamítnout nulovou hypotézu, opatrný závěr: "Nezamítáme H_0" namísto závěru "zamítáme H_1 a přijímáme H_0". Dnešní počítače zjistí pro příslušný test a náš datový soubor pravděpodobnost chyby prvního druhu (tzv. Sig., P-value apod.), a tu srovnáváme s klasickou hodnotou 0,05 a pro zamítnutí H_0 (a přijetí H_1) požadujeme, aby vypočtená hodnota byla nižší. Takto vypočtená pravděpodobnost chyby prvního druhu je právě statistická významnost. Jinak řečeno **statistická významnost je pravděpodobnost, s jakou bychom – za předpokladu pravdivosti nulové hypotézy – mohli obdržet data odporující nulové hypotéze stejně či ještě více než pozorovaná data.** [Euromise, kapitola 7, obdobně Zvára Štěpán 2002: 167]. Jde tedy o podmíněnou pravděpodobnost získání dat s testovou statistikou stejnou jako je naše nebo „horší" při platnosti nulové hypotézy v základním souboru **$P(D/H_0)$ a nikoliv** o pravděpodobnost platnosti nulové hypotézy při existenci našich dat **$P(H_0/D)$** [Cohen 1994, Loftus 1996]. Naše uvažování je běžně řízeno následující logikou: Je-li statistická významnost nízká (většinou menší než 5 %), nulová hypotéza pro základní soubor nejspíš neplatí[29], protože získat náš výběr z takového základního souboru je velmi nepravděpodobné (ale ne nemožné!).

Demonstrujme výše uvedený postup statistického rozhodování na jednoduchém příkladu, aby byla jasná interpretace statistické významnosti na základě výstupů ze statistických softwarových produktů.

Příklad 1: Odlišnosti mužů a žen ve frekvenci konzumace alkoholických nápojů

[29] V praxi při malé hodnotě vypočtené statistické významnosti (nejčastěji pod 0,05) říkáme, že výsledek je statisticky významný a naopak při větší nebo rovné 0,05 říkáme, že je statisticky nevýznamný.

V příkladu vycházíme z dat získaných v České republice v rámci výzkumu ISSP 2007, který byl tématicky zaměřen na volný čas. Respondenty byly osoby starší 17-ti let. Pro popis jednoho z výsledků získaných ve výběrovém souboru (čítajícím 1222 respondentů, z toho 545 mužů a 677 žen) byla zvolena kontingenční tabulka zobrazující frekvenci konzumace alkoholu u mužů a žen (viz tabulka 2).

Tabulka 2. Frekvence konzumace alkoholických nápojů mužů a žen v ČR (2007)

		Pohlaví		Celkem
		muži	ženy	
Pijete alkohol:	Ano, téměř denně	8,4%	1,3%	4,7%
	Ano, 1-2x týdně	30,7%	10,1%	20,1%
	Ano, ale pouze příležitostně	39,1%	40,4%	39,8%
	Velmi zřídka	14,8%	25,2%	20,2%
	Vůbec	7,0%	23,1%	15,3%
Celkem		100,0%	100,0%	100,0%

Zdroj: ISSP 2007, n=1210

Poznámka: Uvedena jsou sloupcová procenta. Výpočet byl proveden v systému SPSS.

Z popisné statistky (sloupcových procent) plyne, že v námi sledovaném výběru platí, že muži pijí alkohol častěji než ženy. Za pomoci postupů vycházejících z testování statistických hypotéz zkusíme vyřešit otázku, zda je tento závěr možné zobecnit na celou dospělou populaci ČR[30]. Konkrétně užijeme chi-kvadrát test o nezávislosti znaků[31]. Nulová hypotéza (H_0) tohoto testu tvrdí, že mezi zkoumanými proměnnými (v našem případě frekvence konzumace alkoholu a pohlaví) neexistuje v základním souboru (dospělé populaci) souvislost. V našem případě lze formulaci nulové hypotézy upravit, neexistence souvislosti znamená věcně, že muži i ženy v ČR konzumují alkohol stejně často (tedy rozložení

[30] S ohledem na to, že naše data jsou pouze z výběru nelze automaticky závěry zjištěné ve výběru zobecňovat na populaci. Důvodem proč toto zobecnění nelze přímo provádět je existence výběrové chyby. I když tuto chybu pro naše data neznáme, lze ji za pomoci známých vzorců odhadnout.

[31] Pro popsaný problém (odlišnost frekvence konzumace alkoholu mužů a žen) by bylo možné užít i další statistické testy (jistě vhodnější s ohledem na zkoumanou otázku). Cílem tohoto textu není upozornit na skutečnost, že při užívání statistické významnosti dochází k užívání nesprávných testů (to by byl námět na samostatný článek). Vycházíme z přesvědčení, že uvedený test je velmi známý a běžně vykládaný v základních kurzech statistiky.

frekvence konzumace alkoholu u mužů i žen je v populaci totožné). Alternativní hypotéza (H_1) naopak tvrdí, že zkoumané veličiny spolu v základním souboru souvisí, v našem konkrétním případě lze interpretovat souvislost v tom smyslu, že rozložení frekvence konzumace alkoholů u mužů a žen je odlišná. Výsledek statistického testování uvedených hypotéz shrnuje tabulka 3.

Tabulka 3. Chi-kvadrát test nezávislosti (frekvence konzumace alkoholických nápojů vs. pohlaví)

	Testové kritérium	Stupně volnosti	Sig
Pearsonův chi-kvadrát	163,320	4	,000

Zdroj: ISSP 2007, n=1210

V případě využití statistického software běžně dostáváme tabulky obdobné tabulce 3. Výsledek testování je možné vyhodnotit dvojím způsobem. První (běžně neužívaný) vychází z hodnoty testového kritéria a případných stupňů volnosti statistického rozdělení příslušného testového kritéria. V našem případě bychom museli mít k dispozici tabulky rozdělení chi-kvadrát a zjišťovat, jaké jsou kvantily tohoto rozdělení při čtyřech stupních volnosti[32]. Při testování na „5 % hladině významnosti“ (tj. nechceme-li připustit pravděpodobnost chyby prvního druhu větší než 0,05) bychom v tabulkách našli hodnotu kvantilu chi-kvadrát rozdělení pro 4 stupně volnosti o velikosti 9,81 [Zvára Štěpán 2000: 220]. Protože hodnota námi vypočteného testového kritéria (163,3) tuto hodnotu převyšuje, uzavřeli bychom, že zamítáme nulovou hypotézu a přijímáme alternativní, tj. frekvence konzumace alkoholu u mužů a žen v ČR je odlišná.

Běžně se namísto srovnávání testového kritéria s kvantily statistických rozdělení využívá vypočtené pravděpodobnosti chyby prvního druhu (většinou značené Sig., P, P – value či P - level). Postup, který se uplatňuje je následující: **V případě, že vypočtená pravděpodobnost chyby prvního druhu je menší než námi předem stanovená hranice** (nejčastěji 0,05 - viz dále v textu) **zamítáme nulovou hypotézu, v případě opačném nulovou hypotézu**

[32] Stupně volnosti pro uvedený test jsou dány součinem počtu řádků a sloupců vždy zmenšeným o jednotku u příslušné kontingenční tabulky. V našem případě máme pět kategorií frekvence konzumace alkoholu a dvě kategorie pohlaví, počet stupňů volnosti je tedy dán výpočtem 4x1=4.

nezamítáme. Pokud postup použijeme v našem případě, zjišťujeme, že softwarem vypočtená pravděpodobnost chyby prvního druhu[33] je menší než 0,05 a tudíž zamítneme nulovou hypotézu a přijmeme alternativu, tj. muži a ženy v ČR se z hlediska frekvence konzumace alkoholu odlišují.

"Nedostatky" statistické významnosti

Zkusme se zamyslet nad tím, jaké jsou problémy statistické významnosti a obvyklého způsobu zacházení s ní. Pochopení těchto slabin umožňuje tento koncept poučeně používat a případně se jeho využití v určitých situacích ubránit [Soukup Rabušic 2007]. Mezi tyto problémy patří zejména [Cohen 1994, Loftus 1996, Thompson 1998b]:

a) nedostatečná výpověď o základním souboru,

b) nereálnost nulových hypotéz,

c) mechanická práce s klasickou 5% hladinou (hvězdičky, stepwise, nejlepší modely apod.),

d) statisticky významné neznamená důležité,

e) nepublikování statisticky nevýznamných výsledků.

Ad a) Předně nutno konstatovat, že statistická významnost nám přímo neříká nic o základním souboru [Thompson 1998b]. Z výše uvedené argumentace plyne, že statistická významnost vypovídá o výběru (o pravděpodobnosti, s jakou můžeme tento nebo „horší" získat ze základního souboru, kde platí nulová hypotéza).

Ad b) Nulové hypotézy v základní verzi testů většinou tvrdí, že v základním souboru (celé populaci) dvě (nebo i více) proměnné spolu nesouvisí, nebo dvě skupiny (nebo i více) mají stejný průměr. To je ale většinou naprosto nereálné očekávání. Představa naprosté nezávislosti nebo shody průměrů ve skupinách neodpovídá často skutečnosti. Platí-li v současnosti v sociálních vědách poučka: „Vše souvisí se vším", případně: „Všichni jsme odlišní", jak může sociální vědec předpokládat, že vše se vším nesouvisí a různé skupiny nejsou odlišné? V angličtině se proto ujal posměšný název **nil null hypothesis** [Cohen 1994,

[33] V našem konkrétním případě je softwarem zobrazená pravděpodobnost chyby prvního druhu 0,000. Pomineme-li problematičnost zápisu z matematického hlediska, je důležité upozornit, že tento zápis znamená, že vypočtená pravděpodobnost chyby prvního druhu je menší než 0,0005 a díky automatickému zaokrouhlování na tři desetinná místa se zobrazuje 0,000. Některé softwarové produkty tisknou pravděpodobnost chyb prvního druhu v korektnějším formátu (<0,0005).

Loftus 1996], česky bychom mohli užít výraz nicotné nulové hypotézy. Nemělo by být naším cílem namísto triviálního vyvracení nereálných hypotéz formulovat hypotézy, které mají reálný základ? Například namísto hypotézy o nulové korelaci v základním souboru ($R = 0$) formulovat hypotézu o slabé závislosti (řekněme $R \leq 0{,}2$). Obdobně místo hypotézy o nulovém rozdílu průměrů příjmů mužů a žen ($\mu_1 - \mu_{-2} = 0$), formulovat hypotézu o tom, že rozdíl je 2000 Kč nebo nižší ($|\mu_1 - \mu_{-2}| \leq 2000$)? Proč toto neděláme? Předně na to nejsme zvyklí, a za druhé v softwarech je zpravidla možné testovat jen výše uvedené nicotné nulové hypotézy. Bylo by zřejmě dobré změnit naše zvyky a případně pátrat v našich softwarech, zda není možné uživatelsky nastavit reálnější nulové hypotézy. Ne vždy je toto nastavení možné, pak nezbývá než užít ruční výpočty a případně apelovat na tvůrce software, aby příslušné procedury upravili k potřebě uživatelů. Dodejme, že proti autorům, kteří tvrdě kritizují nereálnost běžných nulových hypotéz [Cohen 1994, Loftus 1996], vystoupili jiní, kteří naopak nulové hypotézy hájí [Např. Biskin 1998].

Ad c) Ronald Fisher zavedl do statistiky svým doporučením uzanci, že statisticky významný je výsledek, pokud vypočtená chyba prvního druhu pro naše data je menší nebo rovna 5 % ($\alpha \leq 0{,}05$). Je ale toto doporučení nutno slepě následovat? Zcela jistě nikoliv. Vždyť každý zkušený analytik už zažil situaci, že někdy vyjde vypočtená statistická významnost 0,051 a někdy 0,049. Zatímco v prvním případě nezamítneme nulovou hypotézu, v druhém bez problému zamítáme. Rozdíl v pravděpodobnostech je ale pouhých 0,002, neboli po násobení stem 0,2 %. **Moudrý analytik nepřijímá striktně pravidlo 5 %, ale ani 1 %, 10 % či jiné meze**. Oporou mu může být citát užívaný zejména odpůrci statistické významnosti: „Bůh má určitě skoro stejně rád 0,06 jako 0,05." [Rosnow Rosenthal 1989:1307].

Prostý uživatel často místo statistické významnosti používá jen oblíbené hvězdičky. Jedna hvězdička znamená významnost na 5% hladině, 2 hvězdičky pak 1% a 3 hvězdičky 0,1%[34]. Z uvedeného příkladu (0,051 vs. 0,049) je zřejmé, jak pouhé čtení hvězdiček může být ošidné [Leahey 2005, Selvin 1957]. Počítač samozřejmě není moudrý analytik, ale uplatňuje striktně fisherovské doporučení. Na tomto místě je důležité upozornit i na skutečnost, že čím je větší výběrový soubor, tím je ceteris paribus pravděpodobnější, že se hvězdičky (příp. více hvězdiček) objeví. Tato skutečnost plyne ze zákona velkých čísel, podle něhož střední chyba odhadu při rostoucí velikosti výběrového souboru klesá [více o fenoménu velkých souborů a dopadu na statistickou významnost viz Soukup Rabuši 2007: 389-390].

[34] Situaci často ještě komplikují různé softwarové produkty, které přiřazují hvězdičky jiným než výše uvedeným pravděpodobnostem.

Podobně jako hvězdičky se chovají i počítačové procedury hledající „nejlepší" model. Velice známá a oblíbená je například procedura stupňovité regrese (Stepwise regression), nicméně obdoby jsou známy i pro oblast logistické regrese, loglineárních, strukturních modelů, analýzy přežití atd. V čem spočívá nebezpečí těchto procedur? Opět zde vybírá model počítač a ne analytik. Pro výběr modelu se neuplatňují věcná kritéria, ale kritéria statistická. Negenerují se apriorní hypotézy před výzkumem, ale aposteriorně dovozujeme hypotézy z dat. Počítač sleduje tupou logiku 5 % (výrobcem přednastavené hladiny) nebo jiné (uživatelem málokdy změněné) hladiny významnosti. Proměnná, která splňuje příslušné kritérium, je do modelu zahrnuta, a vice versa. Výsledkem je model snad statisticky bezproblémový, ale věcně mnohdy naprosto nepoužitelný. Ne nadarmo někteří statistici nazývají proceduru stepwise slovem nemoudrá (unwise) [King 1985:669] a upozorňují na nerozumnost jejího užívání [Thompson 2001: 86-88]. Vyslovuji proto silné varování před bezmyšlenkovitým užíváním těchto procedur. Mnohem vhodnější postup je prověřovat jeden model teoreticky odůvodněný, případně několik málo si konkurujících modelů (viz dále v části Porovnávání více modelů za pomoci informačních kritérií). Prozkoumávat celé třídy modelů s cílem nalézt nejvhodnější lze považovat pouze za vhodný explorační nástroj v počátečních fázích výzkumu.

Na závěr varování před slepým následování pravidla „Všechno nebo nic aneb co je nad 0,05 je nevýznamné a vice versa" ještě dodejme, že Fisher sice zavedl doporučení užívat hodnotu 0,05, nicméně to je pouze část jeho doporučení, které bývá nekriticky přijímáno a objevuje se od té doby ve všech učebnicích statistiky. Fisher zavedl přinejmenším dvě doporučení. Jednak již zmíněnou hodnotu 0,05, o které tvrdil, že tato a nižší indikuje v experimentálních designech užitečné efekty [Fisher 1926]. Fisher ovšem dále doporučoval, aby v případě, že vypočtená hladina statistické významnosti přesáhne tuto hodnotu, ale nepřekročí hodnotu 0,20, výzkumník přemýšlel, zda se má na efekt zaměřit v dalších experimentech. U hladin významnosti nad 0,2 pak Fisher konstatoval, že v rámci příslušného experimentu se efekt nedaří prokázat. Poučné na tomto historickém exkurzu je, že bortí mýtus fisherovského slepého pravidla. Připomeňme navíc, že Fisher vyvinul své myšlenky pro experimentální designy a nikoliv výběrová šetření.

Ad d) S výše uvedenou kritikou striktního dodržování 5% hladiny významnosti a případného čtení hvězdiček souvisí další varování. **Neplatí tvrzení, že čím více hvězdiček, tím je výsledek důležitější nebo kvalitnější.** Správně je pouze tvrzení: **Nižší vypočtená hladina významnosti značí vyšší statistickou významnost. Ale nic více**. Tříhvězdičkový výsledek

není hodnotnější než dvouhvězdičkový[35] (je jen méně pravděpodobné, že náš výběr je ze základního souboru, kde platí nulová hypotéza).

Ad e) S častým omylem statisticky významné = důležité souvisí i výzkumná praxe spočívající jen v publikaci pouze „důležitých" (správně: jen statisticky významných) výsledků. V abstraktech vědeckých textů se povětšinou objevují výsledky, které byly statisticky významné, i když v samotném textu se dost často uvádí i poznatky, které se nepodařilo prokázat. Pozoruhodná je i strategie, kdy autor uvede v abstraktu, že mu některé výsledky nevyšly statisticky významné, ale přesto si myslí, že rozdíly u daného fenoménu existují. Věda se tak podivuhodně reprodukuje převážně statisticky významnými výsledky, nevýznamné lépe neuvádět, zejména pak v abstraktu. Tento hon za statisticky významnými výsledky lze vypozorovat jak v zahraničí, tak samozřejmě u nás. Prvním, kdo na problém upozornil, byl zřejmě Rosenthal [1979], který fenomén pojmenoval jako **problém hromadění pouze statisticky významných výsledků** (file drawer problem). Základní problém honby za statisticky významnými výsledky je zejména v tom, že v případě metaanalytických postupů jsou závěry dělány jen na základě statisticky významných výsledků, a kvůli tomu jsou výsledky zkreslené. Studie se statisticky nevýznamnými výsledky se následkem jejich nepublikování do metaanalýz nezahrnou. Lze odhadovat, že tento proces má multiplikativní efekt, protože čím více se v dané oblasti publikuje statisticky významných výsledků, tím spíše se ten, kdo takového výsledku nedosáhne, neodváží své výsledky publikovat[36].

Jak si poradit s „nedostatky" statistické významnosti?

Předně nutno uvést, že výše uvedené problémy nejsou jen nedostatky konceptu, ale i nedostatky v užívání, a chápání tohoto konceptu. Koncept sám je nosný, nicméně je nadužíván [Soukup Rabušic 2007] a případně nesprávně užíván. Chceme-li kosmeticky upravit užívání statistické významnosti, pak je namístě začít užívat místo vypočtené statistické významnosti **intervaly spolehlivosti** (confidence intervals). Chceme-li přistoupit k problému důležitosti našich výsledků (tj. neřešit jen jejich statistickou významnost), musíme se zaměřit na věcnou významnost a poukázat na možnosti měření v této oblasti.

[35] Profesoru Blahušovi děkuji za upozornění na MacDonaldovo (MacDonald 1985: 20) přirovnání tohoto systému k hotelům, kde naopak samozřejmě hvězdičky kvalitu značí. Ve statistice nikoliv a jejich používání je i z tohoto důvodu zavádějící.

[36] Zcela zde pomíjíme možnost „úpravy" výsledků k zajištění jejich statistické významnosti. Tento postup vybočuje z etických standardů vědy.

Následující odstavce věnujeme popisu některým statistickým alternativám ke statistické významnosti, v budoucnu bude publikován článek věnovaný problematice věcné významnosti a jejího měření.

Intervaly spolehlivosti (confidence intervals)

Pokud užíváme a/nebo publikujeme pouze vypočtenou hladinu statistické významnosti, jsou možnosti posouzení výsledků poměrně omezené. Ještě menší jsou, pokud uvedeme pouze, zda je výsledek statisticky významný či nikoliv. V této situaci lze říci, zda rozdíl (závislost) je statisticky významný, a toť vše. Nás ale spíše zajímá jak moc je velký nebo malý (statisticky významný) rozdíl dvou skupin, nebo velká či malá závislost proměnných. Bodovým odhadem velikosti rozdílu nebo závislosti je hodnota vypočtená z výběrových dat. Tato hodnota trpí všemi nedostatky bodového odhadu: jedná se o jediné číslo platné pouze pro náš výběr a je téměř vyloučeno, aby stejná hodnota byla platná i pro základní soubor. Z tohoto důvodu přistupují statistici ke konstrukci intervalů spolehlivosti. Tyto intervaly by při opakovaném vybírání (což je ovšem zcela nerealistický předpoklad) zahrnovaly v určité proporci (nejčastěji se opět užívá fisherovských 95 %) hodnotu odhadované charakteristiky v základním souboru. Lze získat nejen informaci o tom, zda je výsledek statisticky významný (1)[37], ale **navíc lze získat i představu o tom, v jakém rozpětí se může hodnota příslušného parametru[38] pohybovat v celé populaci** (2). Ekvivalentní postup ke srovnání vypočtené hladiny statistické významnosti s hodnotou 0,05 lze u intervalů spolehlivosti aplikovat za pomoci této otázky: Obsahuje interval spolehlivosti hodnotu platnou dle nulové hypotézy?[39] Ukažme srovnání obou postupů na jednoduchém příkladu regresní analýzy.

Příklad 2: Vypočtená hladina statistické významnosti a intervalové odhady regresních koeficientů

Na datech z výzkumu ISSP 1999 řešíme závislost příjmu na pohlaví a letech vzdělání respodenta. Po vyřazení respondentů, kteří neuvedli nebo nemají příjem, získáme v regresních procedurách výstup v tabulce 4:

[37] Tedy informaci, kterou lze získat i statistickým testováním (viz výše).
[38] V souvislosti s diskusemi o věcné významnosti se většinou hovoří o tzv. velikosti efektu (rozdílu, koeficientu apod.).
[39] Pokud ji neobsahuje, zamítáme H_0 a přijímáme H_1, v opačném případě pak nezamítáme H_0.

Tabulka 4. Regresní analýza závislosti příjmu na pohlaví a letech vzdělání respodenta

	Nestandardizované koeficienty		Standardizované	t	Sig.
	B	chyba odhadu	Beta		
konstanta	6885,2	1024,0		6,72	3,27E-11
počet let školní docházky	615,2	61,7	0,31	9,97	3,16E-22
pohlaví	-3590,2	376,5	-0,30	-9,54	1,53E-20

Závislá proměnná: čistý příjem jedince

R^2=0,193, F-test (Sig <0,0005)

Zdroj ISSP 1999, n=841. Výpočet byl proveden v systému SPSS.

Z uvedené tabulky můžeme konstatovat, že léta vzdělání i pohlaví jsou v České republice statisticky významnými prediktory příjmu jedince (Sig. v posledním sloupci je zcela jistě menší než běžně užívaných 0,05). Interpretujeme-li klasicky hodnoty regresních koeficientů (sloupec nadepsaný B), lze říci, že české ženy měly v průměru v roce 1999 o 3590 Kč méně než muži (protože kodování proměnné bylo 1=muž, 2=žena) a že s každým rokem vzdělání si Čech či Češka průměrně přilepší o 615 Kč. Tyto bodové odhady mohou být ale poměrně vzdálené od skutečné hodnoty těchto parametrů v základním souboru. Proveďme proto ještě jednou analýzu s tím, že vypočteme navíc intervalové odhady regresních koeficientů (tabulka 5).

Tabulka 5. Regresní analýza závislosti příjmu na pohlaví a letech vzdělání respondenta včetně intervalových odhadů regresních koeficientů

	Koeficienty	Chyba odhadu	t	Sig.	Dolní mez	Horní mez
konstanta	6885,2	1024,0	6,72	3,27E-11	4875,3	8895,1
počet let školní docházky	615,2	61,7	9,97	3,28E-22	494,1	736,2
pohlaví	-3590,2	376,5	-9,54	1,55E-20	-4329,1	-2851,3

R^2=0,193, F-test (Sig <0,0005)

Zdroj ISSP 1999, n=841. Výpočet intervalů spolehlivosti byl proveden v programu MS Excel.

V tabulce 5 jsou nové dva poslední sloupce. Naznačují nám, že rok školy navíc přinese (s 95% pravděpodobností) v ČR něco mezi 494 a 736 Kč čistého příjmu navíc. Rozdíl mezi českými ženami a muži se s velkou pravděpodobností (opět 95%) pohybuje mezi 2851 a 4329 Kč. Vidíme, že výpověď je mnohem nejednoznačnější, zároveň ale mnohem bližší realitě. Statistika (aspoň ta inferenční) neumí sdělovat své závěry s jistotou, ale jen s určitou mírou pravděpodobnosti. Připomeňme, že v intervalu spolehlivosti je skryta nejen informace o statistické významnosti (pokud by interval obsahoval nulu, nebyl by regresní koeficient statisticky významně odlišný od nuly) ale navíc i informace o možné hodnotě koeficientu v celém základním souboru.

Opět můžeme formulovat doporučení: Bylo by dobré změnit naše zvyky a pátrat v našich softwarech, zda není možné uživatelsky nastavit získání intervalů spolehlivosti kromě (namísto) vypočtené hladiny statistické významnosti. Ne vždy je toto nastavení možné, pak nezbývá než užít ruční výpočty a případně apelovat na tvůrce software, aby příslušné procedury upravili k potřebě uživatelů. Dodejme, že v mnoha případech lze intervaly spolehlivosti počítat ve zcela běžném software, jak bylo ukázáno na příkladu regresních koeficientů vypočtených v tabulkovém kalkulátoru Excel.

Míry asociace a problematičnost intervalových odhadů

Zatímco s intervaly spolehlivosti regresních parametrů a rozdílů v průměrech si lze většinou bez problémů poradit, u měr asociace je situace mnohem složitější. Vypočítáme-li například z výběrových dat hodnotu korelačního koeficientu dle Pearsona 0,7, jaký je jeho intervalový odhad? Jaká je jeho pravděpodobná hodnota v celém základním souboru. Standardně umíme provést test o nulové hodnotě příslušné míry asociace a rozhodnout, zda proměnné na sobě závisí (většinou jen lineárně) či nikoliv. Nicméně daleko zajímavější může být zjištění, jak moc spolu proměnné souvisí v základním souboru, a o tom nám bodový odhad z výběrových dat opět moc nepoví. Problém intervalového odhadu korelačního koeficientu tkví v tom, že tento nemá při neznámé hodnotě jeho skutečné hodnoty, žádné známé statistické rozdělení a tudíž nelze snadno užít kvantity těchto rozdělení. Nicméně je známo, že po fisherově transformaci má Pearsonův korelační koeficient přibližně normální rozdělení a tohoto lze využít pro konstrukci intervalu spolehlivosti. Postup lze popsat takto [Fan, Thompson. 2001: 525]:

1. Vypočti Pearsonův korelační koeficient (r) z výběrových dat.

2. Transformuj tento koeficient na veličinu s normálním rozdělení dle vzorce *z=0,5*ln((1+r)/(1-r))*.
3. U transformované veličiny vypočítej interval spolehlivosti za pomoci vzorce: $z \pm u_{1-\alpha/2}/\sqrt{(n-3)}$.
4. Z dolní a horní meze transformované veličiny za pomoci transformace inverzní ke kroku 2 vypočti interval spolehlivosti dle vzorce: (exp(2x) – 1) / (exp(2x) + 1), kde x je dolní nebo horní mez transformované veličiny z kroku 3.

Tento postup je relativně jednoduchý, ale kdybychom ho chtěli následovat pokaždé, při práci by nás zdržoval. Statistici naštěstí vyvinuli pomůcky, které výpočty provedou za nás[40]. Stačí pouze dosadit hodnotu korelačního koeficientu vypočítaného z výběru a počet respondentů. Pomůcky mají podobu on-line kalkulátorů na webu nebo tabulkových kalkulátorů dostupných přes web[41]. Problémem těchto kalkulátorů je omezení intervalů spolehlivosti na 90% nebo 95%. Pro zájemce jsem vytvořil obecnější pomůcku, která je k dispozici na mých webových stránkách[42]. Nicméně protože v sociologii častěji používáme korelační koeficienty pro ordinální data (Spearman, Kendallovo tau) bylo by vhodnější počítat intervaly spolehlivosti pro tyto koeficienty. Zde nelze užít postupu transformace na veličinu s přibližně normálním rozdělením, ale nejvhodnější je nejspíše technika bootstrapu (ta by byla použitelná i pro Pearsonův koeficient). Detailnější popis tohoto přístupu přesahuje možnosti tohoto textu, zájemce lze odkázat na freewarový statistický program R, kde je tento přístup implementován. Uvedení intervalu spolehlivosti (a jeho následná interpretace) pro korelační koeficient je určitě daleko vhodnější než test o nulové hodnotě koeficientu někdy doprovázený jeho bodovým odhadem.

Shrňme, že intervaly spolehlivosti jsou slibnou alternativou k testům statistické významnosti nulových nerálných hypotéz. Jejich hlavní výhodou oproti testům je, že dopředu **není třeba žádné hypotézy formulovat[43]**. Další výhodou je **možnost používání metaanalytických postupů ze získaných intervalů spolehlivosti**. Intervaly spolehlivosti jsou zejména zastánci věcné významnosti považovány za vhodný nástroj [Rozeboom 1960, Cohen 1988, 1994, Thompson 2002], nicméně jejich užívání doporučují i mnozí jiní autoři [Tversky Kahneman

[40] Z běžně užívaných statistických paketů umí intervaly spolehlivosti pro korelační koeficienty počítat SAS a STATISTICA.
[41] Příklad prvního typu lze nalézt na http://faculty.vassar.edu/lowry/rho.html a druhého typu na např. http://www.childrens-mercy.org/stats/weblog2005/CorrelationCoefficient.asp.
[42] Viz http://samba.fsv.cuni.cz/~soukup/pomucky.
[43] Dodejme, že je samozřejmě žádoucí před prováděním analýz (i celého výzkumu) hypotézy formulovat. Děkuji za upozornění na toto doplnění anonymnímu recenzentovi.

1971, Jones 1984, Jones, Matloff 1986, Perry 1986, Hunter 1990, Robinson, 2003 Brandstätter 1999, Denis 2003].

Síla testu (Power analysis)

Oponenti uvádění statistické významnosti oprávněně namítají, že problém statistického testování může spočívat v tom, že sice držíme pod kontrolou chybu prvního druhu, ale ztrácíme ze zřetele, jaká je síla testu [Tversky, Kahneman, 1971 Jennions, Miller 2003, Denis 2003, Cohen, 1988, 1994, Borkowski, Welsh 2001, McCloskey 1985]. Připomeňme, že **síla testu** (viz tabulka 1 popis pod ní) **je pravděpodobnost** (hodnota pohybující se mezi 0 a 1) **správného přijetí alternativní hypotézy za předpokladu, že je tato v základním souboru platná**. U mnohých výzkumů je tato síla testu velmi malá, ale protože ji výzkumník nezná, nemůže toto posoudit. Výpočet síly testu je záležitostí matematiků-statistiků, v jejichž textech lze nalézt příslušné vzorce [česky např. Anděl 2003, 2005, Kalounová, 2000]. Pro sociální vědce je daleko vhodnější užívání statistického software, který má příslušné vzorce implementovány. Již v roce 1989 napočítal Goldstein [1989] třináct produktů, které umí tyto výpočty provést. V dnešní době jich jsou jich desítky a navíc mají tyto procedury implementovány i některé obecné statistické produkty (např. SPSS, STATA, Statistica). Mnohé programy umožňují samostatné výpočty síly testu, jiné produkty jsou doplňkem existujících statistických produktů (SAS) nebo i tabulkových kalkulátorů (Excel). Detailnější přehledy a srovnání produktů lze nalézt v článcích [Goldstein, 1989, Thomas, Krebs: 1997] a samozřejmě na Internetu po zadání sousloví „power analysis“. Sílu testu je vhodné počítat orientačně ještě před provedením výzkumu a spojit s ní úvahu o velikosti výběrového souboru (viz další část článku o velikosti výběrového souboru). Platí pravidlo, že pokud je síla testu malá, není vhodné výzkum vůbec provádět, protože přijetí alternativní hypotézy není pravděpodobné. V literatuře bývá doporučována hodnota 0,8 jako minimální pro sílu testu [Vacha-Haase; Nilsson 1998:49]. Dodejme, že při této hodnotě je pravděpodobnost chybného nezamítnutí nulové hypotézy při její neplatnosti 0,2, a to není málo. Pro zájemce dodejme, že síla testu závisí na pravděpodobnosti chyby prvního druhu (čím je vyšší, tím je vyšší síla testu), na velikosti výběrového souboru (u větších výběrových souborů je síla testu vyšší) a na rozdílu sledovaného ukazatele v porovnávaných skupinách (vyšší rozdíl implikuje ceteris paribus vyšší sílu testu). Samozřejmě, že síla testu závisí i na reliabilitě měřeného ukazatele.

Dobrá zpráva pro sociologii je, že díky velikosti obvykle používaných výběrových souborů v řádu stovek či tisíců jsou síly používaných testů velké. Mnohem hůře jsou na tom výzkumníci z oblasti psychologie či pedagogiky. Nicméně někdy používáme i v sociologii statistické testy pro poměrně malé soubory, zejména při porovnání různých málo zastoupených skupin. Potom je samozřejmě namístě zjišťovat, jak velkou sílu mají námi prováděné testy a zvážit, zda je má na našich datech smysl provádět.

Minimální velikost výběru (n)

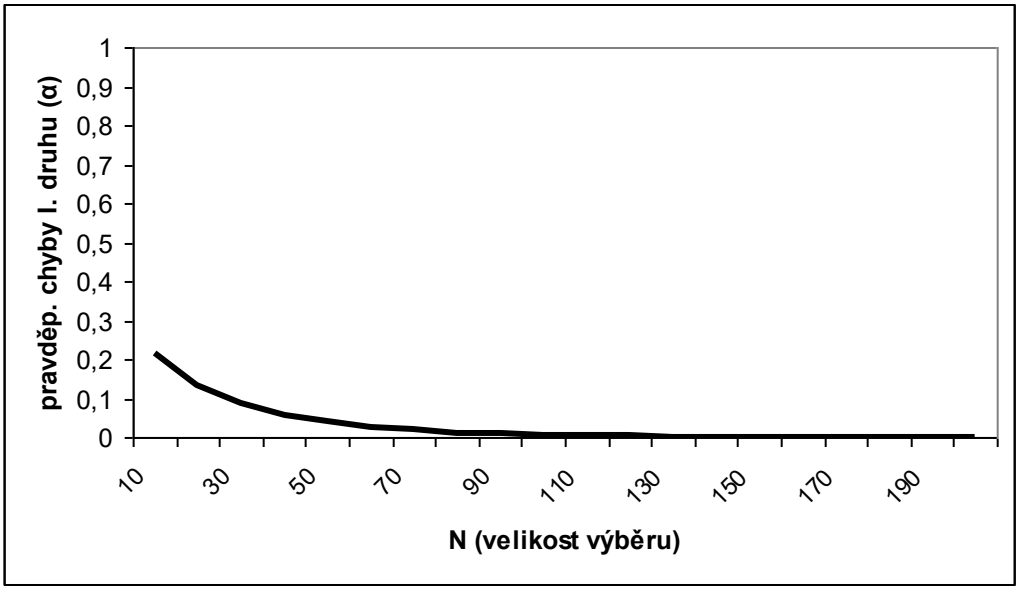
Další doporučovanou pomůckou pro lepší užívání statistických procedur ve výzkumech je plánování velikosti výběru [Snyder Lawson, 1993]. Každý výzkumník, který chce provádět výběrové šetření, by měl dobře zvážit, jak veliký soubor je pro něj vhodný. Nemělo by záležet na libovůli ani na dostupných finančních prostředcích, ale na statistickém zdůvodnění velikosti výběrového souboru. Klasický vzorec pro minimální velikost výběru záleží na velikosti očekávaného rozdílu, tj. efektu (Δ), směrodatné odchylce měřené charakteristiky (σ) a pravděpodobnosti, se kterou budeme chtít odhadovat průměrnou velikost hodnoty měřené charakteristiky pro celou populaci (nejčastěji opět klasických 95 %). Můžeme použít vzorec 3.1.

$$n \geq (u_{1-\alpha/2})^2 * \sigma^2 / \Delta^2 \quad (3.1)$$

Samozřejmě, že běžně neznáme velikost směrodatné chyby měřené charakteristiky a užíváme jejího odhadu z předchozích výzkumů.

Pro ilustraci vlivu velikosti výběrového souboru na chybu prvního druhu uveďme příklad.. Předpokládáme velikost napozorovaného efektu 1, vypočtené velikosti chyb prvního druhu pro různě velké výběrové soubory uvádí graf 1.

Graf 1. Velikost chyby prvního druhu pro jednovýběrový oboustranný t-test pro různé velikosti výběrového souboru (Δ =0,05)[44]



Graf potvrzuje klesající pravděpodobnost chyby prvního druhu s ohledem na velikost výběrového souboru (ceteris paribus), pokles je ovšem nelineární. Na tuto skutečnost (souvislost statistické významnosti a velikosti výběrového souboru) se často poukazuje jako na negativum. Pro velké datové soubory (řádově tisícové a větší) pak nemá testování statistické významnosti valný význam [Soukup, Rabušic 2007].

V případě, že odhadujeme minimální velikost výběrového souboru a naším cílem je odhadnout s určitou přesností proporci jevu v populaci, je situace obdobná:

$$n \geq (u_{1-\alpha/2})^2 * (p*(1-p))^2 / d^2, \quad (3.2)$$

kde p je odhad proporce jevu (opět zpravidla z předchozích šetření).

[44] Výpočty i graf lze nalézt na http://samba.fsv.cuni.cz/~soukup/pomucky.

V sociologických šetřeních je situace vhodné velkosti výběru ještě komplikovaná tím, že většinou není cílem výzkumu měřit jednu charakteristiku. Proto velikost výběrového souboru by měla být větší nebo rovna největší z hodnot, které vypočteme dle uvedených vzorců pro jednotlivé rozdíly a proporce, které hodláme ve výzkumu měřit. Dodejme navíc, že výše uvedené vzorce platí stricto sensu pro prostý náhodný výběr, pro jiné designy výzkumů existují vzorce obdobné a výše uvedené platí s určitou mírou nepřesnosti. Dobrou zprávou pro navrhovatele výzkumů je, že výpočet minimální velikosti vzorku je již implementován do mnohých statistických produktů, které zpravidla umí odhadovat i sílu testu (více viz v předchozí části věnované tomuto tématu). Kromě apriorního odhadu velikosti výběrového souboru je některými autory [Snyder 2000] doporučováno a posteriori vypočítat pro každé zjištění z výzkumu minimální velikost výběrového souboru k prokázání statisticky významného rozdílu nebo závislosti (tzv. **What if strategy**).

Porovnávání více modelů za pomoci informačních kritérií

Jako relativně slibné řešení slabin statistické významnosti se jeví také užívání postupu, kdy není testován jeden model a přijímán nebo odmítán za pomoci hodnot typu 0,09 nebo 0,03. Postup spočívá v tom, že výzkumník negeneruje jednu hypotézu o jednom modelu, ale na základě teoretických předpokladů (a to je pro tento přístup, ale nejen pro něj, klíčové) několik vzájemně si konkurujících modelů tzv. multiple working hypothesis [Anderson, Burnham, Thompson 2000]. Pro jednotlivé modely je poté vypočtena věrohodnostní funkce a na základě ní hodnota informačního kritéria. Nejčastěji jsou užívána Akaikovo informační kritérium (AIC, [Akaike, 1972]) a bayesovské Schwarzovo informační kritérium (BIC, [Raftery, 1995]). Tato kritéria jsou uváděna zpravidla ve formě „čím menší, tím lepší model“ a analytik pak velice jednoduše vybere model, který je nejvhodnější pro jeho data. Tento postup je hojně užíván v rámci strukturních [Urbánek 2000], regresních a loglineárních [Hebák a kol. 2005 a: 111, resp. Hebák a kol. 2005b: 35] a víceúrovňových modelů [Soukup 2006: 1007]. U jednodušších procedur se zpravidla neužívá, ale to samozřejmě neznamená, že jej nelze užít. Bez problémů lze počítat hodnoty informačních kritérií i pro jiné než uvedené procedury, jako například regresní analýzu (lineární i logistickou).

Uveďme základní principy, na kterých stojí koncepce porovnávání více modelů [Burnham , Anderson 2004: 265]:

a) princip **parsimonie** – tj. snaha vybrat co nejjednodušší model obsahující minimální počet parametrů postačujících k popisu reality na základě našich dat,

b) princip **srovnávání více modelů namísto dichotomického rozhodování** o nulové vs. alternativní hypotéze (tento princip je založen na představě, že zpravidla existuje několik teoretických koncepcí, které si vzájemně konkurují a tuto pluralitu by měla odrážet i analýza dat) a
c) princip **získání silného důkazu** – vychází z problému nicotných nulových hypotéz (viz část Nedostatky statistické významnosti) a preferuje vybírat mezi silnými hypotézami o několika modelech (oproti rozhodování nicotná nulová hypotéza vs. smysluplná alternativní hypotéza).

Samozřejmě, že každý z výše uvedených principů by bylo možné diskutovat dále. Burnham a Anderson zároveň upozorňují, že samozřejmě modely jsou z definice zjednodušení reality a vystihují ji jen do určité míry (stejnou skutečnost připomíná i Boxův výrok uvedený na počátku článku).

Před doporučeními ohledně informačních kritérií vyložíme stručně filozofii obou přístupů a upozorníme na odborné diskuse ohledně AIC a BIC. **Akaike** zavedl své informační kritérium jako první a **vyšel zejména z kybernetické informační teorie a statistické teorie věrohodnosti** I přesto má kritérium BIC de facto implicite bayesovský přístup [srov. např. Burnham , Anderson 2004: 283].

Pro úplnost dodejme vzorec užívaný pro výpočet neznámějšího Akaikova kritéria:

$$AIC = -2 \ln (L) + 2k, \quad (1.3)$$

kde L je hodnota věrohodnostní funkce příslušného modelu pro výběrová data a k je počet parametrů odhadovaného modelu.

Ze vzorce je patrné, že se zohledňuje věrohodnost příslušného modelu (tedy pravděpodobnost, že model je vhodný pro naše data, de facto informační hodnota modelu) a dále složitost modelu (tj. princip parsimonie); složitější modely jsou penalizovány navyšováním AIC za každý další parametr o 2 jednotky. Pro vyhodnocení se užívá pravidla, že modely s nižší hodnotou AIC jsou vhodnější. Kromě základního AIC kritéria existuje i modifikované AIC kritérium (AIC_C) pro malé výběry (malý výběr je zde definován jako výběr, kde počet pozorování je méně než čtyřicetinásobkem počtu parametrů modelu).

Prakticky se s kritériem AIC pracuje tak, že se nejdříve stanoví hodnota AIC pro všechny vzájemně si konkurující modely a poté se stanoví rozdíl mezi hodnotou AIC jednotlivých modelů oproti nejnižší hodnotě AIC. Otázku, jak velký rozdíl AIC dvou modelů, lze považovat za významný, řeší doporučení v literatuře (anglicky tzv. rules of tumb) následovně [např. Burnham , Anderson 2004: 271]: rozdíly do 2 jednotek jsou zanedbatelné, rozdíly mezi cca mezi 4-7 jednotkami již stojí za pozornost a rozdíly nad 10 jednotek jsou již výrazné a vedou k jasné preferenci modelu s nižším AIC. Obdobná kritéria definoval i Raftery [1995] pro kritérium BIC.

Z pohledu sociologie vědy je poměrně zajímavé, že kritérium AIC, i když vzniklo dříve, je v sociologii téměř neznámé a dominuje užívání kriteria BIC. Weakliem [2004: 170] uvádí, že v databázi JSTOR nalezl v sociologických článcích 16 užití AIC a stokrát bylo použito kritérium BIC. U článků z oboru ekonomie bylo použito AIC ve 120 případech, BIC ve 100. Vysvětlení pro tuto disproporci je poměrně jednoduché. Zatímco kritérium AIC bylo sociologům detailněji představeno až v roce 2004 (viz dále), kritérium BIC zavedl do sociologie již v polovině 80. let Raftery [1986, 1995]. Od té doby se kritérium BIC stalo standardním sociologickým nástrojem, často problematicky a nesprávně užívaným, obdobně jako statistická významnost [např. Weakliem 1999, Firth Kuha 1999].

Zájemce o detailní informace o kritériu AIC (a srovnání s kritériem BIC) lze odkázat na monotématické číslo Social Methods and Research z roku 2004 (vol. 33).

Přejdeme nyní k druhému informačnímu kritériu, tzv. bayesovského Schwarzova informační kritéria. Kritérium bylo prvně představeno Schwarzem [1978], do sociologie jej zavedl již zmíněný Raftery jako alternativu k napadanému statistickému testování. Na rozdíl od AIC, BIC vychází explicite z Bayesovské statistiky a cílem kritéria je porovnávat poměrově věrohodnost (aposteriorní pravděpoděpodobnost[45]) našeho modelu a modelu saturovaného (tj. modelu obsahujícího všechny myslitelné parametry). Pro úplnost uveďme i vzorec bayesovského Schwarzova informačního kritéria:

$$BIC = -2 \ln(L) + k*\log(N), \quad (1.4),$$

kde N je velikost výběrového souboru.

[45] Tedy pravděpodobnost známou po získání dat, která vychází z apriorní pravděpodobnosti (představ výzkumníka) a zároveň z našeho konkrétního modelu.

S ohledem na uvedenou logiku **platí** i zde, **že se preferují modely, s nižším BIC a prvotně se srovnává se saturovaným modelem, jehož BIC je nulové**. **Modely ze záporným BIC jsou vnímány jako modely lepší než saturovaný** (jsou jednodušší a přesto nemají výrazně menší aposteriorní pravděpodobnost).

Kromě srovnání konkrétního modelu s modelem saturovaným umožňuje BIC i porovnání dvou modelů mezi sebou. Pro tyto situace platí výše uvedená pravidla pro AIC s tím, že kritérium BIC je konzervativnější (preferuje jednodušší modely), kritérium AIC je liberálnější (preferuje složitější modely ceteris paribus). Jako výhody kritéria BIC se zpravidla uvádí (zejména ve srovnání s testováním hypotéz):

a) zohlednění principu parsimonie
b) nezávislost na velikosti výběového souboru (srov. viz část Velikost výběrového souboru a graf 2), resp. modely pro větší soubory dat jsou více penalizovány.

Kritérium BIC ale není bez problémů, detailní diskusi nalezne čtenář v monotématickém čísle Social Methods and Research z roku 1999 (vol. 27). Nejtvrdší kritiku vznesl proti BIC Weakliem [1999], jehož článek je základem zmíněného monočísla. Základní problémy BIC, na které upozorňuje, jsou tyto:

a) tendence k příliš jednoduchým modelům (přílišné uplatnění principu parsimonie,
b) nezohledňuje se rozložení proměnných (bayesovský faktor využitý pro výpočet není závislý jen na N, ale i na rozložení proměnných) a
c) vnucuje výzkumníkům určitou apriorní pravděpodobnost (většinou to ani nevědí) aniž by ji mohli upravit.

Weakliemovu kritiku podpořili i Firth a Kuha [1999], částečně též Gelman a Rubin [1999], kteří se snažili navrhnout kompromisní strategii spočívající v kombinaci užívání BIC a statistických testů vhodnosti modelu (založených většinou na chi-kvadrát rozloženích). BIC samozřejmě obhajoval Raftery [1999] a také Xie [1999]. Raftery trochu paradoxně přiznal, že kritika Weakliema je v mnoha bodech oprávněná, ale přesto jí považuje za příliš tvrdou. Připomněl, že již ve svých textech z 80. a 90. let 20. století navrhl alternativní vzorce pro některé situace, kdy běžné BIC selhává. Není tedy Rafteryho chybou, že se tyto vzorce neužívají a mechanicky se pracuje s jediným vzorcem. Raftery [1999: 413] dále upozornil, že BIC je samozřejmě jen pomůckou výzkumníka a pro výběr modelu by mělo být použito zejména věcných úvah a možné interpretovatelnosti modelu.

Dodejme ještě několik poznámek o problémech při praktickém používání informačních kritérií. Jedním z problémů může být situace, kdy dle jednoho kritéria je nejlepší například model *A* a dle druhého jiný model, například *B* [Soukup 2006: 1007]. Situace však může být ještě komplikovanější, protože informačních kritérií je více (cca 10) a tudíž i „nejlepších" modelů může být více. Zde analytikovi nezbývá než přihlédnout i k jiným pomůckám, jako jsou celkové testy modelů, diagnostika reziduí apod. Nelze doporučit, aby se analytik spolehl pouze na užívání informačních kritérií a užíval je jako doplňkový nástroj[46]. Doplňme, že v rámci strukturních modelů se nabízí i jiná kritéria ke srovnávání modelů, jako je AGFI, RMSEA apod. [Urbánek 2000]. I tato kritéria jsou mnohými autory doporučována jako alternativa ke statistické významnosti [Onwuegbuzie, Levin, Leech, 2003: 1042].

Závěrem uveďme, že je poměrně příznačné, že bayesovské přístupy, které obecně nutí výzkumníka přemýšlet a stanovit si apriorní pravděpodobnosti (na základě věcné úvahy, či předchozích výzkumů) a poté vypočítat aposteriorní pravděpodobnost, sklouzly do mechanického užívání informačních kritérií. BIC je nástrojem vycházející z Bayesovské statistiky, ale k přemýšlení v zásadě nenutí. Automatickou aplikaci BIC (či jiného informačního kritéria) bez zohlednění dalších statistických nástrojů k vyhodnocení dat je nutné odmítnout stejně jako automatické používání statistické významnosti.

Slovní řešení nedostatků v rámci statistické významnosti

Někteří autoři navrhují, aby došlo ke změně výrazu statistická významnost, případně aby se výrazu striktně užívalo takto a nebylo používáno pouze samotného slova významnost. Jaké návrhy lze nalézt v literatuře? První relativně mírný návrh formuloval Thompson [1996]. Jeho doporučení zní: „**Vždy užívejte** sousloví **statistická významnost**, nikdy pouze slova významnost" (tučně autor článku). Toto doporučení plyne z častého omylu významný = důležitý. Aby k tomuto pomýlení docházelo pokud možno co nejméně, navrhuje Thompson výše uvedené pravidlo. Jiní autoři s ním ale polemizují, že tato praxe je jazykově problematická a nejspíš zbytečná [Levin, Robinson 1997]. Navrhují naopak nahradit sousloví statistická významnost jinými slovy, jako vhodný adept se jim jeví slovo **nenáhodnost** (non-chance). Tento výraz má vyjádřit, že statisticky významný výsledek (dle alternativní hypotézy) není získán náhodně, ale je poměrně pravděpodobné, že tento výsledek platí i pro celou populaci. S tímto názorem ovšem vyslovil polemiku opět Thompson [1998b]. Na

[46] Jak správně poznamenal ve svém posudku anonymní recenzent článku, informační kritéria vybírají nejvhodnější z modelů, ale nezajišťují, že jde o model pro naše data opravdu vhodný.

citátech si ukažme problematičnost (nesprávnost) užívání sousloví statistická významnost v české vědě. „Navíc jsme dosáhli **statistické významnosti všech čtyř parametrů.“** [Trilobyte 2004]. **„Statistické významnosti bylo dosaženo**.“ [Widimský 1999]. Oba citáty nepřímo naznačují, že hlavním cílem vědy (resp. výzkumu) je dosahovat statistické významnosti. To je nutno razantně odmítnout. Statistická významnost pouze hovoří o možnosti zobecnit data z výběrů (resp. experimentů) na populaci, ale nijak neměří významnost věcnou (viz počátek článku). Uveďme s komentářem další dva citáty: „**Rozdíl nedosáhl statistické významnosti“.** [Ballantyne, M. B., A.. G. Olsson, T. J. Cook, M. F. Mercuri, T. R. Pedersen, J.Kjekshus. 2001]. „**A tento pokles (bez statistické významnosti**) nastal i ve všech sledovaných okresech“. [Státní zdravotní ústav 1999]. Pokles či rozdíl může být statisticky významný na nějaké hladině, kterou musí výzkumník explicitně uvést (a pokud možno i dopředu stanovenou), jinak je tvrzení značně nepřesné. Namístě je ovšem komentovat zejména věcnou velikost rozdílu a statistickou významnost případně pouze zmínit jako doplněk. Užívání neúplného výrazu dokumentuje citát učebnice biomedicínské statistiky: „Statistické programy nám umožňují testovat významnost parciálních korelačních koeficientů“. [Euromise] Dodejme, že statistické pakety umožňují testovat právě statistickou významnost, nikoliv jinou, slušelo by se tedy doplnit do citované věty slovo „statistickou“ .

S ohledem na uvedené citace je namístě podpořit Thompsonovo doporučení **užívat vždy plného sousloví statistická významnost a navíc jí užívat jen tam, kde je to vhodné**.

Česky by bylo možné používat místo sousloví statistická významnost v oblasti výpočtů založených na výběrech též slova **zobecnitelnost**, které nemá zavádějící konotace. Výraz nenáhodnost není pro český kontext zřejmě vhodný.

Závěr

Tento článek se pokoušel poukázat na problematičnost konceptu statistické významnosti a jeho možná zneužívání. Realisticky s ohledem na možnosti praktikujících sociologů a dostupný software lze přinejmenším doporučit používat místo konstatování statistické významnosti spíše intervaly spolehlivosti pro rozdíly, parametry či koeficienty a vyhnout se zmíněným nepřesnostem ve vyjadřování. Při plánování výzkumu je vhodné zvážit velikost výběrového souboru a velikost síly testu, aby nebyly prostředky na sběr dat vynaloženy zbytečně. V případě více si konkurujících modelů je namístě používat s opatrností jako doplněk též informační kritéria. Je nutno upozornit i na to, že někteří metodologové

statistickou významnost nesnáší natolik, že navrhují vypuštění této koncepce z vědy. Například v časopise Psychological Science (1997, vol.8) byla publikována celá sekce nazvaná „Zrušit testy statistické významnosti“. Z diskuse také vznikla celá kniha nazvaná výmluvně What if there were no significance test? [Harlow, Mulaik, Steiger 1997], Americká psychologická asociace vytvořila pracovní skupinu, která se snažila problém řešit, výsledek lze nalézt v publikačním manuálu této asociace [APA 2001]. Poměrně zajímavá je úvaha Riopelle [2000] nad tím, jak by vypadala vědecká práce, kdyby v časopisech zakázali publikování statistické významnosti. Podle Riopelle by autoři tuto ve své práci používali dále, pouze by její výsledky nepublikovali. Nelze než závěrem vyzvat k uvážlivé přípravě, sběru a analýze kvantitativních dat.

Literatura citovaná ve 3. kapitole

Akaike, H. 1972. Information theory and an extension of the maximum likelihood principle. Proceedings of 2nd international symposium. Information theory, support to probléme of control and information theory. 267-281

Anděl, J. 2003. *Statistické metody*. Praha: Matfyzpress.
Anděl, J. 2005 *Základy matematické statistiky*. Praha: Matfyzpress

Anderson, D.R., K. P., Burnham, W. L. Thompson,.2000. Null hypothesis testing: Problem, prevalence and alternative. *Journal of wildlife management*, 64(4) : 912-923.
APA. 2001. *Publication manual of the American Psychological Association*, 5th edition. Washington DC.
Arbuthnott, J. 1710. An argument for divine providence taken from the regularity in the births of both sexes. *Philosophical Transactions of the Royal Society*. 27: 186-190.
Ballantyne, M. B., A.. G. Olsson, T. J. Cook, M. F. Mercuri, T. R. Pedersen, J.Kjekshus. 2001. Vliv nízkého HDL cholesterolu a zvýšené hladiny triglyceridů na riziko kardiovaskulárních příhod a účinek terapie simvastatinem ve studii 4S. www.zdravcentra.cz/cps/rde/xchg/zc/xsl/79_1720.html
Biskin, B. H. 1998. Comment on significance testing. *Measurement and Evaluation in Counseling and Development*. 31(1): 58-62.

Borkowski, S. C., M. J. Welsh. 2001. An analysis of statistical power in gender-related research. *Accounting Enquiries*. 11(1): 83-125.

Box, G. E. P. 1976. Science and statistics. *Journal of American Statistical Association*. 71 : 791-799

Brandstätter, E. 1999. Confidence Intervals as an Alternative to Significance Testing. *Methods of Psychological Research Online*. 4 (2): 33-46.

Burnham, K. P., D. R. Anderson. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*. 33: 261-304.
Cohen, J. 1994. The earth is round ($p < .05$). *American Psychologist*, 49: 997-1003.
Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Science (2nd ed.).* Hillsdale (NJ): Erlbaum.

Cox, D. R. 1982. Statistical significance tests. *British Journal of clinical Pharmacology*. 14: 325-331.

Denis, D., J. 2003 Alternatives to Null Hypothesis Significance Testing. *Theory & Science*. 1-26

Euromise. Základy statistiky pro biomedicínské obory. http://ucebnice.euromise.cz/index.php?conn=0&section=biostat1

Fan, X., B. Thompson. 2001. Confidence Intervals for Effect Sizes: Confidence Intervals about Score Reliability Coefficients, Please: An EPM Guidelines Editorial. *Educational and Psychological Measurement*. 61(4): 517-531.

Firth, D., J. Kuha. 1999. Comments on "A Critique of the Bayesian Information Criterion for Model Selection".*Sociological Methods & Research.* 27: 398-402.

Fisher, R. A. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fisher, R., A. 1926. The arrangement of field experiments. *Journal of Ministry of agricalture of Great Britain*. 33: 503-513.

Gelman, A., D. B. Rubin. 1999. Evaluating and Using Statistical Methods in the Social Sciences: A Discussion of "A Critique of the Bayesian Information Criterion for Model Selection" *Sociological Methods & Research*. 27: 403-410.

Goldstein, R. 1989. Power and Sample Size via MS/PC-DOS Computers. *The American Statistician,* 43 (4): 253-260.

Harlow, L., L., S. A. Mulaik, M., L. Steiger. 1997. *What if there were no significance tests?* Mahwah (NJ): Erlbaum.

Hebák, P. (ed.) 2005a. *Vícerozměrné statistické metody (2).* Praha: Informatorium.

Hebák, P. (ed.) 2005b. *Vícerozměrné statistické metody (3).* Praha: Informatorium.

Hunter, J. S. 1990. Commentary. *Technometrics* 32(3): 261.

Jennions, M., D. , A., P. Miller. 2003 A survey of the statistical power of research in behavioral ecology and animal behavior. *Behavioral Ecology*. 14(3): 438–445

Jones, D. 1984. Use, misuse, and role of multiple-comparison procedures in ecological and agricultural entomology. *Environmental Entomology*. 13(3): 635-649.

Jones, D., N. Matloff. 1986. Statistical hypothesis testing in biology: a contradiction in terms. *Journal of Economic Entomology.* 79(5): 1156-1160.

Kahounová, J. 2000. *Praktikum k výuce matematické statistiky I.* Praha: Vysoká škola ekonomická

Kaiser, H. 1970. A second generation little jitty. *Psychometrika.* 35: 411-436.

King, G. 1986. How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science*. 30(3): 666-687.

Leahey, E. 2005. Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology. *Social Forces.* 84 (1): 1-24.

Loftus, G. R. 1996. Psychology will be a Much Better Science When We Change the Way We Analyze Data. *Current Directions in Psychological Science.* 1: 161-171.

Macdonald, R. 1985. *Factor analysis and related methods*. Hillsdale, NJ: Erlbaum. (český překlad Macdonald, R. 1991. Faktorová analýza a příbuzné metody v psychologii. Praha: Academia)

McCloskey, D. N. 1985. The loss Function has ben mislaid: The rhetoric of Significance tests *AEA Paper and Proceedings*.

Perry, J. N. 1986. Multiple-comparison procedures: a dissenting view. *Journal of Economic Entomology.* 79(5) : 1149-1155.

Raftery, A., E. 1986.Choosing Models for Gross-Classifications.*American Sociological Review*. 51: 145-146.

Raftery, A., E. 1995. Bayesian model selection in social research. In Mardsen, P., V. *Sociological Metodology*. MA: Cambridge. Blackwell.
Raftery, A., E. 1999. Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection". *Sociological Methods & Research.* 27: 411-427.

Riopelle, A. J. 2000. Are effect sizes and confidence levels problems for solutions to the null hypothesis test. *The Journal of General Psychology*. 127(2) : 198-216.

Robinson, D. H. 2003. An Interview with Gene V. Glass. *Educational researcher*. 33(3): 26-30.

Robinson, D., H., J., R. Levin. 1997. Reflections on statistical and substantive significance with a slice of replication. *Educational Researcher*. 26(5): 21-27.

Rosenthal, R. 1979. The „file drawer problem“ and the tolerance for null results. *Psychological bulletin*. 86: 638-641.

Rosnow, R., R. Rosenthal. 1989. Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*. 44: 1276-1284.

Rozeboom, W. W. 1960. The fallacy of the null hypothesis significance test. *Psychological Bulletin.* 57: 416-428.

Selvin, H., C. 1957 A Critique of Tests of Significance in Survey Research. *American Sociological Review*. 22 (5): 519-527.

Schwarz, G. 1978. Estimating the Dimension of a Model. *Annals of Statistics*. 6: 461-464.

Snyder, P. 2000. Guidelines for Reporting Results of Group Quantitative Investigations. *Journal of Early Intervention*. 23(3): 145–150.

Soukup, P., L. Rabušic. 2007. Několik poznámek k jedné obsesi českých sociálních věd – statistické významnosti. *Sociologický časopis*. 43(2): 379-395.

Soukup, P. 2006. Proč užívat hierarchické lineární modely. *Sociologický časopis.* 42 (5): 987-1012.

Thomas, L., Ch., J. Krebs, 1997. A review of statistical power analysis software. *Bulletin of the Ecological Society of America*. 78(2): 126-139.

Thompson, B. 1996. AERA Editorial policies regarding statistical significance tests: three suggested reforms. *Educational Researcher*. 25(2): 26-30.

Thompson, B. 1998a. Statistical significance and effect size reporting: Portrait of a possible future. *Research in the schools*. 5(2): 33-38

Thompson, B. 1998b. Five Methodology Errors in Educational Research: The Pantheon of Statistical Significance and Other Faux Pas Invited address (Divisions E, D, and C) presented at the annual meeting (session #25.66) of the American Educational Research Association, San Diego.

Thompson, B. 2001. Editor’s Note on the "Colloquium on Effect Sizes: the Roles of Editors, Textbook Authors, and the Publication Manual“. *Educational and Psychological Measurement*. 61(2): 211-212.

Thompson, B. 2002. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*. 31(3): 24-31.

Trilobyte. 2004. Stránky statistického software Trilobite (www.trilobyte.cz/qlin.html)

Tversky, A. D. Kahneman. 1971. Belief in the law of small numbers. *Psychological Bulletin*. 76(2): 105-110.

Urbánek, T. 2000. *Strukturní modelování*. Brno: Psychologický ústav.

Vacha-Haase, T., J. E Nilsson. 1998. Statistical significance reporting: Current trends and uses in MECD. *Measurement and Evaluation in Counseling and Development*. 31(1): 46-57.

Weakliem, D.L. 1999. A Critique of the Bayesian Information Criterion for Model Selection. *Sociological Methods & Research*. 27: 359-397.

Weakliem, D.L. 2004. Introduction to the Special Issue on Model Selection. *Sociological Methods & Research.* 33: 167-187

Widimský, J. 1999. Hypertenze starších osob. *Časopis české společnosti pro hypertenzi*. 2. (www.hypertension.cz/casopis/2_99/6.html)

Xie, Y. 1999. The Tension between Generality and Accuracy. *Sociological Methods & Research*. 27: 428-435.

Zvára, K., J. Štěpán. 2002. *Pravděpodobnost a matematická statistika*. Praha: Matfyzpress.

4. Věcná významnost výsledků a její možnosti měření

"Je velice špatná praxe zakládat významnost výsledků pouze na hodnotě P". (Cox).

Úvod

Při užívání kvantitativních dat často sklouzává analýza k pouhému vyhodnocení statistické významnosti výsledků. Často se zapomíná, že výsledek by měl být nejen zobecnitelný na populaci, kterou zkoumáme (tj. statisticky významný), ale též prakticky užitečný, tj., věcně významný. Cílem tohoto textu je přiblížit českému čtenáři tři čtvrtě století diskuse o věcné významnosti a způsoby, jak ji měřit, a v některých momentech tuto diskusi též prohloubit (ukázat na způsoby měření věcné významnosti v méně používaných technikách). Ke konci článku je též upozornění na další koncepty, které s obecným pojetím věcné významnosti úzce souvisejí a rozvíjejí jej. V závěru se pokoušíme o syntézu přístupů věcné i statistické významnosti a diskusi možných přínosů současného používání obou těchto postupů.

Věcná významnost

Věcná významnost je zhruba stejně stará jako statistická významnost, nicméně jde o koncept mnohem méně známý. Důvodem této skutečnosti je nejspíše absence tohoto konceptu ve výuce i odborných textech z oblasti statistiky a analýzy dat. Zřejmě první zmínkou o věcné významnosti je text Boring [1919]. V sociologii byl prvním autorem, který vyslovil varování před nekritickým užíváním statistické významnosti Selvin [1957], v psychologii Rozeboom [1960]. V posledních cca 30 letech sílí diskuse o věcné významnosti, jejím měření a zejména jejím užívání ve vědě. Mnohé časopisy a profesní asociace mění publikační standardy a vyžadují, kromě statistické významnosti, výpočty významnosti věcné [APA 2001, AERA 2006]. Diskuse o prospěšnosti (neužitečnosti) té které významnosti se objevují v prestižních časopisech zejména z oblasti psychologie, pedagogiky (více o těchto změnách píši v části nazvané Statistická nebo věcná významnost?). V české sociologii (ale ani ve světové) však tato diskuse neprobíhá a ve výuce studentů stále přetrvává pozitivistické pojetí statistické významnosti bez hlubšího pochopení. Dodejme, že česká psychologie nebo pedagogika jsou

na tom obdobně. V oblasti kinantropologie je situace lepší [Blahuš 2000, Hendl 2004] a problémy jiných než statistických významností se řeší také částečně v medicínské metodologii [Euromise]. Tento článek do jisté míry plní tuto mezeru a snaží se potřebnou diskusi vyvolat. Nejdříve je nutné vymezit, co to vlastně věcná významnost je.

Definice věcné významnosti (substantive significance)

Pro první přiblížení uveďme vymezení rozdílu mezi statistickou a věcnou významností dle Kirka [1996: 746]: „Statistická významnost zkoumá, zda je výsledek výzkumu dosažen náhodou nebo proměnlivostí výběrových dat; věcná významnost se zabývá tím, zda je výsledek užitečný v reálném světě.“ Z vymezení je zřejmé, že věcná významnost na rozdíl od statistické dokáže pomoci zhodnotit důležitost, užitečnost výsledku výzkumu. Detailnější definici věcné významnosti podali Tailor a Frideres [1972: 466]. Jejich definice je založená na myšlence, že výzkumná data slouží k prověření předpovědí plynoucích z existujících teorií. Dle nich připadají v úvahu tři případy, kdy je výsledek výzkumu věcně významný: “1) pokud jsou napozorovaná data důležitá pro dvě nebo více alternativních předpovědí plynoucích z teorie, 2) pokud data neodpovídají žádným teoretickým předpovědím (všechny teorie tedy jsou nesprávné) a 3) pokud různá míra shody mezi daty a teoretickými předpověďmi umožňuje alespoň částečně uspořádat teoretické předpovědi z hlediska správnosti a tím zprostředkovaně i seřadit teorie, z nichž byly předpovědi odvozeny“. Protože uvedené definice i přes jejich obecné přijímání, jsou poměrně nejasné, stanovme si vlastní definici věcné významnosti. Věcná významnost výsledku znamená, že naměřený rozdíl či zjištěná souvislost je důležitá pro vědecké poznání či praktické účely. Na rozdíl od statistické významnosti, která zjišťuje, zda nalezený výsledek je zobecnitelný (tj. zda není způsobený náhodou ovlivňující výběr jednotek či experimentálních podmínek), nám věcná významnost sděluje, zda o výsledku má vůbec smysl hovořit[47] a zda má praktické důsledky (vč. důsledků pro vědu samotnou). K tomu abychom zjistili, zda je výsledek věcně významný a pokud ano, pak nakolik, nám mohou pomoci určité ukazatele, míry věcné významnosti (srov. viz dále).

Kromě problémů s definicí narážíme též na nejednoznačnost jazykovou. V angličtině se nejčastěji užívá sousloví substantive significance, nicméně jak upozorňuje Blahuš [2000:58], užívají se i jiné termíny a jejich české ekvivalenty. Setkáváme se s pojmy: významnost praktická (practical significance), logická – (logical), „výsledková důležitost“ - result importace, „výsledková smysluplnost“ - result meaningfulness“ (závorky v pův. znění). Mezi

[47] Samozřejmě nezpochybňuji okřídelný výrok: „I nula je ve vědě výsledek.“

těmito výrazy jednotliví autoři zpravidla neodlišují a používají je víceméně jako synonyma. Na okraj dodávám, že se lze setkat i s výrazem vědecká významnost (scientific significance). Trošku problémem je jazykové vyjádření v českých textech, kdy autoři přejímají pro ukazatele měřící věcnou významnost, které zde označuji jako míry věcné významnosti (angl. effect size) anglický výraz, a hovoří o efektech účinku [srov. Hendl 2004]. S ohledem na skutečnost, že tato terminologie dosud není ustálena, navrhuji používat termín míra věcné významnosti, protože přímo z názvu plyne, co tato míra měří. Sousloví efekt účinku je v češtině poměrně nejasné.

Absolutní a relativní věcná významnost

Prvotní měření věcné významnosti bylo založeno na prosté a všem známé myšlence rozdílů hodnot ve dvou (či více) sledovaných skupinách. Hovoříme o **absolutní věcné významnosti rozdílů** (v původních jednotkách měření), nebo o **relativní věcné významnosti rozdílů** (v procentech) [Čelikovský a kol. 1979; Blahuš: 2000].

Vypočtené absolutní a relativní věcné významnosti jsou velmi jednoduché (každý jim rozumí), nicméně trpí jedním neduhem, a to je závislost na jednotkách měření původní veličiny (resp. v případě relativní významnosti by bylo lépe hovořit o průměrné úrovni jevu, ale ta je jen jiným vyjádřením závislosti na měřítku). Tento problém odstraňují složitější míry věcné významnosti, kterým se věnujeme dále. To ovšem nijak neznamená, že bychom měli ignorovat jednoduché ukazatele absolutní a relativní věcné významnosti [Blahuš 2000]. Naopak měli bychom tyto spočítat a poté přistoupit k výpočtu složitějších měr a poté obojí ukazatele interpretovat a poukázat na jejich smysl.

Míry věcné významnosti rozdílů a závislostí (Effect size measures)[48]

Po popisu teoretické opodstatněnosti koncepce věcné významnosti je třeba poukázat na výpočetní možnosti v rámci tohoto konceptu. K měření věcné významnosti se dnes používá již několika desítek měr, které je možné klasifikovat dle těchto kritérií:

A. Dle toho co měří [Kirk 1996]:
-míry měřící rozdíly a

-míry vyjadřující vysvětlený rozptyl.[49]

B. Dle toho, zda jsou nezkresleným odhadem hodnoty v populaci [Vacha-Haase-Thompson 2004]:

- míry, které jsou nezkresleným odhadem (unbiased) a

- míry, které jsou vychýleným odhadem (biased).

C. Dle statistické procedury namísto které (nebo se kterou) mohou být použity [Sink Stroh 2006]:

- míry pro situaci porovnání dvou skupin (t-testů),

- míry pro situaci porovnání více skupin (analýzu rozptylu),

- míry pro závislost kardinálních proměnných (regrese, analýza kovariance),

- míry pro speciální procedury (diskriminační analýzu, vícerozměrné škálování, korespondenční analýzu, víceúrovňové modely apod.).

Přehlednou klasifikaci nejčastěji používaných měr dle prvních dvou kritérií podává schéma 1.

[48] Zde již užívám svůj návrh terminologie pro anglický výraz effect size measures.
[49] Tedy míry, jejichž cílem je vystihnout sílu souvislostí.

Schéma 1 Přehled jednotlivých měr věcné významnosti

Měří rozdíl/rozptyl	Je vychýleným odhadem	Název míry
rozdíl	ano	Cohenovo d
rozdíl	ano	Hedgesovo g
rozdíl	ano	Glassovo delta
rozptyl	ne	Haysovo omega^2
rozptyl	ano	Fisherovo Eta^2
rozptyl	ano	Korelace
rozptyl	ano	Index determinace
rozptyl	ne	Upravený index determinace

V literatuře lze nalézt desítky měr, například Kirk [1996] jich nalezl 40, nicméně pro první seznámení postačí detailněji představit výše uvedených 7 měr a ostatních referovat jen okrajově. Před jejich popisem a ukázkami jejich výpočtu a vlastností si uveďme podmínky, za nichž má smysl tyto míry používat.

Prvotní záměr byl za pomoci uvedených měr **měřit v experimentech vliv sledovaného efektu** (proto obecný název přístupu a měr je v angličtině effect size, ve zkratce často jen ES). Tyto míry měly měřit rozdíly (souvislosti) mezi experimentální a kontrolní skupinou **v náhodných (randomizovaných) experimentech**. Jejich obdobou v oblasti statistcké významosti jsou běžně užívané t-testy nebo analýza rozptylu v případě více experimentálních skupin. Dodejme, že za pomoci náhodných experimentů nedochází ke zobecněním výsledků na celou populaci, ale zobecňujeme pouze vliv příslušného efektu. **Postupně ale dochází k rozšíření užívání měr** věcné významnosti rozdílů a závislostí **i do druhé oblasti tj. do oblasti náhodných výběrů**, které provádíme, abychom mohli zobecňovat na celou populaci. Představme si jednotlivé míry (rozdělené dle výše uvedeného kritéria A), ukažme způsoby jejich výpočtu a jejich vazby ke klasickým inferenčním statistikám tj. testovým kritériím.

1) Míry měřící rozdíly

Cohenovo d[50]

Tato míra věcné významnosti rozdílů a závislostí je zřejmě společně s Haysovým omega nejužívanější (zejména v psychologii a pedagogice). Je založena na rozdílu průměrů ve dvou skupinách, nicméně tento jednoduchý ukazatel standardizuje, tj. dělí směrodatnou odchylkou průměrů. Výsledkem je **bezrozměrná veličina**, která není závislá na původních jednotkách měření a umožňuje srovnání výsledků i ve výzkumech, které používali k měření stejného fenoménu různých škál. Základní verze vzorce pro Cohenovo d [Cohen 1988] má tento tvar:

$$d=(x_1-x_2)/\sqrt{s^2}, \qquad (4.1.)$$

kde x_1 a x_2 jsou průměry v první (experimentální) a druhé (kontrolní skupině) a s^2 je rozptyl společný oběma skupinám. K výpočtu společného rozptylu lze užít nejobecněji vzorce založeného na váženém průměru rozptylů v obou skupinách:

$$s^2= (n_1*s_1^2+n_2*s_2^2)/(n_1+n_2), \qquad (4.2)$$

kde s_1^2 a s_2^2 jsou rozptyly v první a druhé skupině a n_1 a n_2 velikosti prvního a druhého souboru. V případě, že jsou obě skupiny stejně velké, redukuje se vzorec na prostý aritmetický průměr dvou rozptylů:

$$s^2= (s_1^2+s_2^2)/2. \qquad (4.3)$$

Možná ještě jednodušší než provádět výpočet za pomoci výše uvedených vzorců, je využít blízkosti k hodnotě t-statistiky (tj. testového kritéria dvouvýběrového t-testu). Cohenovo d lze při znalosti hodnoty t-statistiky vypočítat:

$$d=t*(n_1+ n_2)/ \sqrt{(df*n_1*n_2)}, \qquad (4.4)$$

Kde df značí počet stupňů volnosti příslušného t-testu. Provedeme výpočet Cohenova d na příkladu z dat ICCS 2009.[51]

[50] Pro Cohenovo d volím poměrně detailní pojednání. Ostatní míry s ohledem na podobnou logiku jsou již pojednány všechny najednou. Pro všechny míry věcné významnosti používám pro srovnání příklad ze stejných dat.

[51] ICCS2009 je mezinárodní studie občanské výchovy, která byla mj. provedena i v České republice. Cílovou skupinou byli 14-letí žáci 8. ročníků (tj. žáci ze základních škol a víceletých gymnázií).

Příklad 1 Srovnání znalostí z občanské výchovy 14-letých žáků v ČR v testu ICCS 2009 pomocí Cohenova d

Tabulka 1. Popisné statistiky pro znalosti z občanské výchovy 14-žáků ZŠ a osmiletých gymnázií v ČR (2009)

Národní občanské znalosti - skóre[52]

Typ školy	Průměr	N	Směr. odchylka
Gymnázium	160,94	481	8,37
ZŠ	148,72	4131	9,37
Celkem	150,00	4613	10,00

Zdroj: ICCS 2009, N=4613

Využijeme vzorců 4.1 a 4.2 a po dosazení z tabulky 1 získáme d=1,32.

Cohenovo d může být obecně reálné číslo v intervalu od -∞ do +∞, běžně ale nabývá hodnot v řádu jednotek. Pokud vyjde hodnota kladná, znamená to, že sledovaná veličina má větší hodnotu v první skupině-experimentální (výkon žáků na gymnáziích je lepší než na základních školách) a v případě záporné hodnoty Cohenova d je naopak hodnota v první-experimentální skupině nižší. Cohen také definoval určitá rozpětí pro svou míru a přiřadil jim názvy [Cohen 1988: 25], které vypovídají o velikosti rozdílu mezi skupinami. Toto rozlišení uvádí tabulka 2.

Tabulka 2. Rozpětí absolutní hodnoty Cohenova d a jejich slovní označení

Interval	**Slovní označení**
<0,2-0,5)[53]	small
<0,5-0,8)	medium
0,8 a vyšší	large

Zdroj: Cohen [1988:25]

[52] Skóre je národně standardizováno na škále, která má průměr 150 a směrodatnou odchylku 10 a vychází z výsledků testu občanské výchovy. Více se lze dozvědět v publikacích z projektu ICCS, česky například v publikaci [Schulz a kol. 2010].

[53] Cohen nevymezil tyto intervaly, ale přiřadil slovní hodnocení konkrétním hodnotám, hodnotě 0,2 malý, 0,5 střední a hodnotě 0,8 velký. Nicméně z logiky věci plyne, že zamýšlel svá označení užít spíše pro uvedené intervaly než pro izolované hodnoty. Bohužel někteří autoři toto mechanicky přejali a hovoří tak o malých, středních či velkých efektech dle Cohena.

Uvedená tabulka může samozřejmě vyvolat oprávněné pochyby. Pokud statistici a metodologové upozorňují, že není rozumné užívat slepě Fisherovo doporučení o 5% hladině významnosti u statistických testů (srov. např. Soukup 2010), jak může být dobré užívat tyto meze pro měření věcné významnosti rozdílů. Nutno brát výše uvedená označení i intervaly jen jako jedno z možných doporučení. Mnohem vhodnější je srovnávat hodnotu d mezi jednotlivými výzkumy, případně jednotlivými zeměmi, lety apod. Takové srovnání nám může přinést daleko více než srovnání s tabulkovými hodnotami.[54]

Poměrně zajímavý je pohled na hodnotu Cohenova d skrze normální rozdělení. Hodnota Cohenova d může být interpretována jako procento osob z jedné skupiny, které převyšují průměrného člena skupiny druhé. Pro jednotlivé hodnoty Cohenova d se dá toto procento přesně stanovit za pomoci hodnot distribuční funkce normálního rozdělení (% $=\Phi^{-1}(d)$) Pro jednoduchost uveďme opět tabulku (3), která nám vše usnadní.

Tabulka 3. Tabulka hodnoty Cohenova d a příslušného procenta osob z jedné skupiny, které převyšují průměrného člena skupiny druhé

d	0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	1,1	1,2	1,3	1,4	1,5	2
%	50%	54%	58%	62%	66%	69%	73%	76%	79%	82%	84%	86%	88%	90%	92%	93%	98%

Zdroj:vlastní výpočty

V případě, že mezi skupinami není rozdíl v průměrech sledovaného znaku (d=0) je u poloviny (50%) členů první skupiny hodnota znaku vyšší než u průměrného člena druhé skupiny. V případě malého rozdílu (d=0,2), má 58 % členů první skupiny hodnotu znaku vyšší než u průměrný člen druhé skupiny, v případě středně velkého rozdílu (d=0,5) již 69 % a u velkého rozdílu (d=0,8) téměř čtyři pětiny (79 %). Při hodnotě Cohenově d=2 je již téměř vyloučen překryv obou skupin z hlediska měřené charakteristiky, 98% členů první skupiny má hodnotu vyšší než je průměr ve druhé skupině.[55] Zároveň musíme ihned vyslovit varování. Výše uvedená interpretační pomůcka je použitelná jen v případě, že rozdělení sledované proměnné v obou skupinách je normální. Tento předpoklad je ale v sociálních vědách naplněn poměrně řídce, proto užití uvedené interpretační pomůcky musíme brát spíše jako přibližné, přesnou se stává jen u normálně rozdělených veličin.

[54] Toto doporučení neplatí jen pro Cohenovo d, ale i pro všechny ostatní míry věcné významnosti. Doporučit jednu konkrétní hodnotu pro srovnání napříč vědními obory a jejich specializacemi jednoduše nelze.
[55] K měření překryvu obou skupin slouží ještě jiná charakteristika, zájemce odkazujeme na původní text [Cohen, 1988: 21-23].

Vrátíme-li se k hodnotě Cohenova d z našeho příkladu (příklad 2) můžeme rozdíly mezi výkony v občanské výchově označit jako velké. 90 % gymnazistů převyšuje svými výkony průměrného žáka na základní škole (poznamenejme, že výsledné testové skóre z testu ICCS 2009 má téměř normální rozdělení). Toto zjištění je jistě poměrně zajímavé a dává našim výsledkům nový rozměr.

Pro výzkumníka může být zajímavé srovnat hodnotu rozdílů, kterých dosáhl s hodnotami jiných výzkumníků, nicméně problémem pro většinu výzkumníků znalých postupů statistické významnosti může být otázka: Jak z tohoto bodového odhadu Cohenova d mohu něco usuzovat na situaci v základním souboru? Na tuto otázku se snažili statistici odpovědět a výsledkem jsou intervaly spolehlivosti pro Cohenovo d (a nejen pro ně). Jak lze počítat interval spolehlivosti pro Cohenovo d, nebo-li interval ve kterém s určitou pravděpodobností leží hodnota d v celém základním souboru? Nejjednodušší vzorec má tuto podobu:

$$d \pm u_{1-\&/2} * \sqrt{(n_1+n_2)/(n_1*n_2)+d^2/(2*(n_1+n_2))}, \qquad (4.5)$$

kde $u_{1-\&/2}$ je 1-α/2 procentní kvantil normovaného normálního rozdělení a význam ostatních symbolů je jako ve vzorcích 4.1. a 4.2.

Pro úplnost ukažme výpočet intervalu spolehlivosti na našich datech. Cohenovo d leží s 95% pravděpodobností mezi 1,22 a 1,42. Efekt navštěvované školy na znalosti z občanské výchovy žáků je výrazný, interval spolehlivosti je poměrně úzký, vzhledem k velikostem výběrových souborů.

Uvedený vzorec intervalu spolehlivosti je pouze přibližný a platí pro velké výběrové soubory (v řádech cca stovek či tisíců). Přesný výpočet intervalu spolehlivosti je mnohem složitější a je založen na necentrálním t-rozdělení a iteračním postupu [Cumming Finch 2001]. Naštěstí jsou k dispozici programy např. ECSI [Cumming Finch 2001] nebo předpřipravené procedury v SPSS [Smithson 2001] či jiných paketech.[56] Problému intervalů spolehlivosti pro míry věcné významnosti rozdílů a závislostí se věnuji ještě v části textu poukazující na výhody a nevýhody těchto měr.

Přehled dalších měr věcné významnosti měřících rozdíly

[56] Odkaz na stránky softwaru ESCI uvádím na svých webových stránkách http://samba.fsv.cuni.cz/~soukup/ stejně jako odkazy na mnohé další zajímavé stránky týkající se věcné a statistické významnosti.

Dalšími používanými mírami pro měření věcné významnosti rozdílů jsou zejména Hedgesovo g a Glassovo delta. Vzorce pro výpočet obsahuje tabulka č. 4.

Tabulka 4. Přehled dalších měr věcné významnosti pro rozdíly

Název míry	Vzorec	Výsledek na datech z příkladu 1
Hedgesovo g	$g=(x_1-x_2)/\sqrt{MS_w}$	1,32
Glassovo delta	$\Delta=(x1-x2)/\sqrt{(s_k^2)}$	1,30

Vysvětlivky:

MS_w - průměr vnitroskupinového součtu čtverců a
s_k^2 je rozptyl kontrolní (srovnávací) skupiny.

Hedgesovo g do jisté míry vybočuje, protože na rozdíl od ostatních měr má již v základu charakter inferenční a ne popisný. Míra je velice podobná Cohenovu d, ale rozdíl mezi průměry dělí odmocninou z průměru vnitroskupinového součtu čtverců (mean square), který se počítá v rámci analýzy rozptylu. Hodnota Hedgesova g je v našem případě velice podobná hodnotě Cohenova d a obecně platí, že ve velkých výběrových souborech jsou hodnoty d a g stejné.

Glassovo delta užívá na rozdíl od Cohenova d ve jmenovateli směrodatnou odchylku kontrolní skupiny (případně té, vůči níž chceme porovnávat skupinu jinou). Hodnota je tedy opět velice blízká Cohenovu d. Při srovnatelných velikostech rozptylů v obou sledovaných skupinách je toto samozřejmostí. Obdobně jako se užívají doporučení pro velikost Cohenova d (tabulka 2), někteří autoři doporučují shodné meze i pro Glassovo delta. Slepé následování těchto mezí, ale opět nelze doporučit.

2) Míry vyjadřující vysvětlený rozptyl

Alternativně k mírám měřícím věcnou významnost rozdílů lze užít míry měřící vysvětlený rozptyl. Pokud máme více než dvě srovnávané skupiny, pak dříve představené míry pro měření věcné významnosti rozdílů užít nelze a nezbývá než užívat míry vysvětleného rozptylu. Přehled nejběžnějších měr a jejich vzorců obsahuje tabulka č 5.

Tabulka 5. Přehled měr věcné významnosti (pro vysvětlený rozptyl)

Název míry	Vzorec	Výsledek na datech z příkladu 1
Fisherovo Eta^2	$\eta^2=SS_b/SS_T$	0,140
Haysowo ω	$\omega^2=(SS_b-(\upsilon-1)*MS_w)/(SS_T+MS_w)$	0,146
Korelační koeficient	$r=\sum\sum(x_i-x)*(y_i-y)/(s_x*s_y)$	0,382
Index determinace	$R^2=SS_R/SS_T$	0,146
Upravený index determinace	$R_{adj}^2= 1-(1-R^2)*(n-1)/(n-k-1)$,	0,146

Vysvětlivky:

SS_T - celkový součet čtverců,
SSb je meziskupinový součet čtverců,
υ - počet srovnávaných skupin,

MS_w - průměrný vnitroskupinový součet čtverců,

SS_R - regresní součet čtverců,

n – velikost výběru a

k - počet nezávisle proměnných.

Poznamenejme, že s výjimkou korelačního koeficientu nabývají všechny uvedené míry věcné významnosti zaměřené na vysvětlený rozptyl hodnot z intervalu <0,1>, korelační koeficient hodnot z intervalu <-1,1>. Z výsledků na datech z příkladu 1 je navíc zřejmé, že s výjimkou korelačního koeficientu jsou výsledky použití těchto měr obdobné.

Zřejmě nejjednodušším ukazatelem měřícím vysvětlený rozptyl je čtverec ukazatele Eta. Jeho použití u experimentálních designů navrhoval už Fisher [1925]. Eta je propojeno s analýzou rozptylu a měří se podílem meziskupinového součtu čtverců a celkového součtu čtverců. Hodnota se interpretuje (obdobně jako u dalších měr) po vynásobení stem jako procento vysvětleného rozptylu za pomoci rozdělení do skupin. Připomeňme, že Eta^2 je zkresleným odhadem charakteristiky v základním souboru. Nicméně u velkých souborů je toto zkreslení minimální, Eta^2 je tedy pro náš příklad (s mnohatisícovým výběrem) vhodnou mírou.

Problémy zkreslenosti odhadu Eta^2 řeší Haysovo omega. Jde o opět o jednu z nejstarších měr věcné významnosti rozdílů [Hays 1963]. Není překvapením, že vzhledem k velikosti našeho výběru je hodnota po zaokrouhlení na dvě desetinná místa totožná s hodnotou Eta^2. Všechny popsané míry tedy budou mít na datech z příkladu č. 1 podobnou interpretaci, tj. typ školy ovlivňuje výsledek žáka v testu z občanské výchovy cca z 15 %.

Korelační koeficient (r) a index determinace (R^2)

Zaměřme se ještě více u běžně užívaného korelačního koeficient (v případě jedné nezávislé třídící proměnné) resp. spíše jeho druhé mocniny, indexu determinace (v případě jedné i více nezávislých třídících proměnných). Na okraj poznamenejme, že pro měření věcné významnosti se nejčastěji užívá biseriálního korelačního koeficientu mezi třídící proměnnou a sledovanou vlastností, ale běžně se užívá i Pearsonova korelačního koeficientu (viz tabulka č. 5 výše). Pro interpretaci věcné významnosti bývá zvykem brát v potaz jeho absolutní hodnotu. Cohen [1988] navrhl pravidlo (obdobné jako pro Cohenovo d), aby byly vzaty v potaz meze pro hodnocení věcné významnosti (viz tabulka č. 6).

Tabulka 6. Rozpětí absolutní hodnoty korelačního koeficientu (r) a jejich slovní označení

Interval	Slovní označení
<0,1-0,3)[57]	small
<0,3-0,5)	medium
0,5 a vyšší	large

Zdroj:Cohen [1988]

Nutno dodat, že zejména pro sociologii jsou tato doporučení naprosto nevhodná a mělo by být cílem výzkumníka za pomoci komparací s výzkumem v jiných zemích a jiných letech si obdobná pravidla formulovat ad hoc, nikoliv následovat výše uvedené doporučení. Pro rozšíření souvislostí dodejme, že hodnotu korelačního koeficientu lze získat i přepočtem z hodnoty Cohenova d a z hodnoty t-kritéria z t-testu. Pro poslední uvedenou možnost užíváme vzorce:

$$r=\sqrt{(t^2/(t^2+df_w))}, \qquad (4.6)$$

kde t^2 je druhá mocnina t kritéria a df_w je počet stupňů volnosti připadající na vnitroskupinový součet čtverců (zjistitelné například z tabulky pro analýzu rozptylu nebo z dvouvýběrového t-testu ve verzi pro stejné rozptyly ve skupinách).[58]

[57] Cohen nevymezil tyto intervaly, ale přiřadil slovní hodnocení konkrétním hodnotám, hodnotě 0,1 malý, 0,3 střední a hodnotě 0,5 velký. Nicméně z logiky věci opět plyne, že zamýšlel svá označení užít spíše pro uvedené intervaly než pro izolované hodnoty.

[58] Další převodní vztahy mezi vzorci pro kritéria věcné významnosti a testovými kritérii lze nalézt na mých webových stránkách http://samba.fsv.cuni.cz/~soukup/.

Mnohem častěji než korelační koeficient, zejména díky přímé interpretaci, se užívá indexu determinace. V případě jediné třídící proměnné se vypočítá jako druhá mocnina korelačního koeficientu, v případě více třídících proměnných pak dle vzorce v tabulce č. 5.

Výhodou indexu determinace je přímá interpretovatelnost (po vynásobení stem udává procento rozptylu vysvětlené třídícími proměnnými), ale trpí jedním neduhem. Jde o zkreslený odhad, který procento vysvětleného rozptylu vždy nadhodnocuje.[59] Aby bylo tomuto problému zabráněno, odvodili mnozí statistici vzorce pro upravené indexy determinace, zřejmě nejznámější je Ezekielův vzorec (opět viz tabulka č. 5).

Jiné vzorce odvozené Wherrym, Herzbergem či Lordem lze nalézt v textu Sink a Stroh [2006:405]. I pro index determinace lze nalézt doporučené hodnoty pro zhodnocení jejich věcné významnosti, nejčastěji 0,01 (malá věcná významnost), 0,06 (střední věcná významnost) a 0,14 (velká věcná významnost). Lze uvažovat i o druhých mocninách z doporučovaných hodnot pro korelační koeficient, o hodnotách 0,01 (malá), 0,09 (střední) a 0,25 (velká), nicméně opět doporučuji být vůči těmto mezím velice obezřetný. Ve statistické literatuře se dokonce objevil i názor, že index determinace by se vůbec neměl používat [King 1985: 675-678]. Dle názoru Kinga neexistuje důvod, proč by ukazatel, který měří rozptýlení bodů kolem regresní křivky, byl vhodným ukazatelem kvality regresní analýzy. King aforisticky dodává, že když by tomu tak bylo, pak by zřejmě nejvhodnější nezávislou proměnnou byl jinak měřený ukazatel, který je závisle proměnnou [King 1985: 678]. King proto doporučuje v rámci regrese interpretovat hodnoty jednotlivých regresních parametrů a celkový F-test a index determinace užívat pouze doplňkově nebo vůbec. Domnívám se, že v mnohém je Kingova kritika indexu determinace přehnaná (i když je v mnohém sympatická) a má i nadále smysl tento ukazatel pro hodnocení věcné významnosti používat. Zejména může sloužit ke srovnání jednotlivých výzkumů se stejnou měřenou charakteristikou, nebo i pro srovnání jednotlivých analýz v rámci jednoho výzkumu při stejné měřené charakteristice a různých vysvětlujících proměnných.

Výhody a nevýhody měr věcné významnosti

Po uvedení základních měr věcné významnosti se pokusme poukázat na výhody a nevýhody významnosti věcné. Autoři, kteří doporučují užívat míry věcné významnosti (a často nadto

[59] Toto zkreslení se snižuje s velikostí výběrového souboru, v našem příkladu je minimální (rozdíl je na 4. desetinném místě).

doporučují neužívat statistickou významnost) se snaží tyto dva koncepty srovnávat a nacházet výhody prvého. Mezi výhody měr věcné významnosti dle těchto autorů patří [Thompson, 1998b]:

a) nezávislost na velikosti výběrového souboru, stejná využitelnost pro malé i velké výběry,

b) nezávislost na měřítku (srovnatelnost) a možnost využití v metaanalýze,

c) výpověď o velikosti rozdílu nebo souvislosti.

Ad a) Zatímco statistická významnost určitého rozdílu nebo závislosti je různá pro různě velké výběry a platí, že ve velkých výběrech se daří prokázat téměř všechny rozdíly (souvislosti) jako statisticky významné, míry věcné významnosti vychází stejně velké pro malé i velké výběry při stejném rozdílu či souvislosti ve výběru. Nehrozí, že bychom u velkých výběrových souborů snadno nalézali velké hodnoty měr věcné významnosti a vice versa. Nicméně tuto příjemnou vlastnost bodových odhadů měr věcné významnosti problematizují jejich intervalové odhady (více viz výše v části věnované Cohenově d a dále v závěru článku).

Ad b) Výhodou představených měr věcné významnosti oproti jednoduchým měřítkům absolutní a relativní významnosti je nezávislost na měřítku sledované veličiny. Tato vlastnost je důležitá zejména pro metaanalytické postupy, které se snaží za pomoci výsledků z mnoha studií nalézt skutečný vliv sledovaných efektů. Není náhoda, že autor Glassova delta je považován za zakladatele moderní metaanalýzy. Výhodnost měr věcné významnosti pro tyto účely ve srovnání se statistickou významností je zřejmá.

Ad c) Míry věcné významnosti jasně charakterizují velikost rozdílu nebo souvislosti a tyto lze srovnávat mezi jednotlivými výzkumy (lety, apod.). U statistické významnosti s ohledem na vazbu na velikost výběrového souboru tato srovnání činit nelze, připomeňme, že neplatí statistický významnější=důležitější.

Pro vyvážené hodnocení popis měr věcné významnosti je nutno také upozornit na jejich nedostatky. Mezi nejčastěji uváděné nedostatky měr věcné významnosti patří tyto [Onwuegbuzie, Levin, Leech. 2003]:

a) nejde o inferenční ale pouze deskriptivní charakteristiky,

b) jsou založeny na určitých parametrických předpokladech (zejména normalitě) a tyto nejsou často splněny,

c) závisí na reliabilitě měřeného ukazatele,

d) neměří významnost pro jedince ale průměrnou, proto je v některých oblastech problematicky použitelná (viz dále klinická významnost),

e) jsou výrazně ovlivněny uspořádáním (designem) výzkumu.

Ad a) Míry věcné významnosti jsou většinou pouze bodovými odhady a navíc často nadhodnocují skutečnou hodnotu v základním souboru (viz schéma 1 a popis jednotlivých měr uvedený výše). Proto jejich nekritické přijetí je nevhodné. Částečné řešení problémů nabízí intervaly spolehlivosti pro tyto míry (viz další odstavec).

Ad b) Interpretace v duchu překryvu skupin (viz tabulka 3) je závislá na normalitě sledovaného ukazatele. Toto ale není v sociálních vědách často splněno, a proto je interpretace měr i jejich užití problematické.

Ad d) Míry věcné významnosti jsou obdobně zejména díky využití popisných statistik založeny na všech pozorováních ve výběru. Proto umožňují průměrné zhodnocení a nikoliv zhodnocení na úrovni jednotlivců. Některé disciplíny ovšem s tímto nevystačí a proto je zapotřebí užívat i jiných významností, například v medicíně byla proto zavedena klinická významnost (viz dále).

Obecně je nutno mírám věcné významnosti ještě vytknout další rysy, ktteré ale spíše souvisí s jejich užíváním než s nimi samými. Míry věcné významnosti musí být vnímány jako pomůcka k věcné interpretaci výsledků, nikoli jako cíl. Pokud se tak neděje, hrozí že obsese statistickou významností (skrze prezentování P hodnot) nahradí obsese mírami věcné významnosti (skrze prezentování hodnot jako je Cohenovo d). Dále nelze mechanicky aplikovat Cohenova doporučení (či jiných autorů) ohledně interpretace velikosti měr věcné významnosti, to bychom sklouzli k období hvězdičkování statisticky významných výsledků. A v neposlední řadě je třeba pamatovat, že pro mnoho situací míry vcné významnoti chybí (lze odhadovat, že budou postupně dovozeny) a nepokrývají vždy celou věcnou šíři zkoumaného problému (ostatně toto platí pro statistiku jako celek).

Intervaly spolehlivosti pro míry věcné významnosti rozdílů a závislostí

Pro vylepšení praxe užívání měr věcné významnosti byly postupně odvozeny postupy, které umožňují odhadnout intervaly spolehlivosti pro hodnotu příslušné míry v základním souboru. Výpočet přibližného intervalu spolehlivosti pro Cohenovo d byl ukázán v části věnované této míře (vzorec 1.5), výpočet intervalu spolehlivosti korelačního koeficientu skrze Fisherovu transformaci je běžně dostupný v různých pomůckách nebo přímo implementován v software (R, STATA, SAS, STATISTICA). Myšlenka intervalů spolehlivosti pro míry věcné významnosti je shodná jako u intervalů spolehlivosti pro průměr či podíl, nicméně výpočet intervalů spolehlivosti měr věcné významnosti je bohužel obtížnější. Důvodem je zejména skutečnost, že míry věcné významnosti (resp. funkce od nich odvozené) nemají klasická pravděpodobnostní (centrální) rozdělení, jako je t, F či χ^2, ale sledují různá necentrální rozdělení. U necentrálních rozdělení je nutno znát (resp. odhadnout z výběru) kromě počtu stupňů volnosti ještě parametr necentrality (noncentrality parametr). Necentrální rozdělení nejsou symetrická a jsou posunutá právě o zmíněný parametr. Lze je dobře aproximovat centrálními rozděleními (zejména t, nebo normálním rozdělením) v případě velkých výběrových souborů a malých hodnot parametru necentrality. Problematice intervalových odhadů měr věcné významnosti bylo věnováno monotematické číslo časopisu časopisu Educational and Psychological Measurement (2001 61: No. 4), kde lze nalézt texty tento problém popisující [Fan, Thompson 2001; Cumming Finch 2001; Fidler Thompson 2001; Smithson 2001; Algina Moulder 2001]. Ve statistických paketech se lze setkat s výpočty kvantilů necentrálních rozdělení (např. SPSS umí pracovat s necentrálním rozdělením t, F, χ^2 a β), v tabulkových kalkulátorech naopak tato rozdělení chybí. Je jen otázkou času, kdy tvůrci statistických paketů zahrnou výpočty měr věcné významnosti a jejich intervalů spolehlivosti do příslušných procedur. Zatím nezbývá než využívat speciální pakety, nejlepší je v tomto ohledu již zmíněný ECSI.

Další míry věcné významnosti a jejich interpretace

Pro doplnění základních měr věcné významnosti je vhodné uvést i některé méně tradiční ukazatele, které plní tuto úlohu. Konkrétně se zmíníme o ukazatelích užívaných pro mnohorozměrné techniky: víceúrovňové modelování, mnohorozměrné škálování, diskriminační analýzu, logistickou regresi a korespondenční analýzu. Protože literatura se této oblasti téměř nevěnuje (nepodařilo se mi najít texty tomuto specifickému problému věnované), jde v této části o názory autora, které nelze opřít o žádné citace uznávaných autorit z oboru.

V případě víceúrovňového modelování je základní charakteristikou, která ukazuje na věcnou významnost (de facto důležitost použití víceúrovňového modelu) vnitrotřídní korelační koeficient (ICC, intra class correlation coefficient), který ukazuje nakolik je sledovaná charakteristika ovlivněná kontextuální proměnnou (navážeme-li na náš příklad, pak se můžeme ptát, nakolik ovlivňuje výsledek žáka z matematiky škola, do které dochází). Výpočty a interpretace této míry věcné významnosti nalezne zájemce např. v článku [Soukup 2006]. Ukazatel se běžně interpretuje po vynásobení stem v procentech (tedy analogicky jako index determinace či Fischerovo Eta^2 – viz výše). Opět nelze klást jednoznačná doporučení, ale platí, že čím je hodnota ICC větší, tím je podstatnější provádět analýzu víceúrovňově. Nadto ve víceúrovňových modelech využíváme míry analogické k indexu determinace na jednotlivých úrovních, s detaily lze opět odkázat na článek [Soukup 2006] a na literaturu v článku uvedenou.

V případě mnohorozměrného škálování se používají míry pro hodnocení kvality zobrazení mnohorozměrné konfigurace v málorozměrném prostoru (typicky ploše). Běžně se užívají dva typy ukazatelů, tzv. stresy a ukazatele založené na korelacích. Nevypovídají přímo o věcné významnosti výsledků ale o věrohodnosti zobrazení výsledků v prostorech s nižší dimenzí. U stresových charakteristik i korelačních ukazatelů se většinou setkáváme s hodnotami mezi 0 a 1. Zatímco stres měří nesoulad a žádoucí jsou nízké hodnoty (doporučení bývá pod 0,1) u korelačních měr měřících soulad (vzdáleností v mnoho a málorozměrném prostoru) jsou žádoucí hodnoty vysoké (doporučení nad 0,9). Opět platí, že hodnoty doporučené je nutno brát spíše orientačně a je nutno využívat srovnání s jinými modely a jinými daty. Více se lze o konstrukci těchto měr a mnohorozměrném škálování dočíst v knize Hebáka a kolektivu [2005] a literatuře tam uvedené.

V případě diskriminační analýzy se pro hodnocení věcné významnosti (úspěšnosti klasifikace) používá nejčastěji Wilksovo lambda (podíl vnitroskupinového rozptylu a celkového). Jde o ukazatel obdobný k Eta^2. Hodnoty jsou mezi 0 a 1, ale pro interpretaci je potřeba dopočítat doplněk Wilksova lambda do jedné a ten zpravidla interpretovat po vynásobení stem jako procento rozdílů mezi skupinami, které vysvětlují jednotlivé predikátory v diskriminační analýze. Obdobné ukazatele používané pro vyhodnocení úspěšnosti modelu diskriminační analýzy jsou: kanonická korelace a tzv. vlastní číslo (eigenvalue). Tyto hodnoty ale nemají tak snadnou interpretaci, proto jim zde nevěnujeme pozornost.

V logistické regresi se analogicky k indexům determinace používá tzv. pseudo indexů determinace, zřejmě nejužívanější je Nagelkerkovo pseudo R^2. Hodnota je mezi 0 a 1 a vyšší hodnoty opět značí lepší kvalitu modelu. Je zapotřebí upozornit, že při interpretaci těchto měr nevystačíme s dosavadními doporučeními a nelze provádět násobení stem a vyjadřovat procenta vysvětleného rozptylu. Důvodem je skutečnost, že závisle proměnná v logistické regresi není spojitá (binární, ordinální či nominální) a proto zde rozptyl měřit nelze.

Obdobně jako u logistické regrese neuspějeme s jednoduchou interpretací ani u míry věcné významnosti v modelech korespondenční analýzy, tzv. inercie. Technicky jde o charakteristiku, která je odvozena z testového kritéria chi-kvadrát měřícího souvislosti v kontingenční tabulce. Inercie (obdobně jako ukazatele u mnohorozměrného škálování – viz výše) měří věrohodnost zobrazení, jen na rozdíl od charakteristik typu stress či korelačních koeficientů se inercie běžně rozkládá na části, které vysvětlují jednotlivé dimenze (nejčastěji bývají dvě a obrázek se vykresluje v ploše). Získáme tak kromě čísla charakterizujícího celkovou vysvětlenou inercii modelem (obdoba korelací u mnohorozměrného škálování, jen s hodnotami mezi 0 a 100) hodnoty, které charakterizují přínos jednotlivých dimenzí (a de facto jejich potřebnost). Platí opět jako orientační doporučení, že hodnota celkové vysvětlené inercie by měla být minimálně 90 %. Více se opět lze dozvědět v knize Hebáka a kolektivu [2005] a literatuře tam uvedené.

Standardizované koeficienty jako míry věcné významnosti

Většina měr věcné významnosti dosud představená v zásadě slouží k vystižení věcné významnosti celých statistických modelů. Samozřejmě v případě, když je model založen pouze na dvou proměnných, měří vlastně působení jednotlivých proměnných, ale v sociálních vědách většinou užíváme modely s více proměnnými. Cílem pak je často nejen posoudit vliv všech proměnných najednou, ale i jednotlivě. Toto klasické a již představené míry věcné významnosti neumí (neznamená to, že jsou díky tomu nepoužitelné!). Pro tyto případy je nejvhodnějším nástrojem standardizovaný koeficient. Připomeňme, že běžně udává posun závisle proměnné (v jednotkách, které odpovídají její směrodatné odchylce) při navýšení nezávisle proměnné o jednu směrodatnou odchylku. V lineární regresi lze využít tzv. beta koeficienty běžně nabízené statistickým software, v ostatních složitějších technikách (logistické regresi, víceúrovňových modelech, strukturních modelech apod.) je nutno jejich obdoby počítat skrze standardizaci proměnných. Samozřejmě pro srovnávání těchto koeficientů platí, že je lze využít pro srovnání modelů se stejnými proměnnými (závisle i

nezávisle) pro data z různých regionů, let apod. Tím je samozřejmě využití omezeno (často měříme stejné fenomény různými indikátory). Přes výše uvedený nedostatek je velice vhodné kromě dříve uvedených měr věcné významnosti charakterizujících modely jako celky, používat při publikaci výsledků i tyto „dílčí“ míry věcné významnosti jednotlivých proměnných. Jak pro interpretaci konkrétních výsledků, tak pro možná následná srovnání a provádění metaanalýz (srov. dále) je tato praxe žádoucí.

Statistická nebo věcná významnost?

Po představení měr věcné významnosti se nabízí otázka, zda máme hodnotit výsledky analýz prismatem statistické, nebo věcné významnosti. Případně, zda se máme snažit oba koncepty sloučit. V literatuře věnované věcné významnosti se objevují tři typy autorů (přístupy) [Thompson, 1998a]:

1) ti, co navrhují zcela koncept statistické významnosti zcela opustit [Loftus 1993, Schmidt 1996],
2) ti, co jsou zastánci věcné významnosti, ale připouštějí i používání statistické významnosti či jiných statistických postupů [Thompson, 1998a] a
3) krajní zastánci statistické významnosti [Robinson, Levin, 1997; Onwuegbuzie, Levin, Leech 2003].

Zejména představitelé druhého a třetího přístupu učinili snahy o sloučení věcné a statistické významnosti. Jedním z velmi známých postupů je dvoustupňový postup (two step) Robinsona a Levina [1997]. Zjednodušeně ho lze popsat následovně. Nejdříve posuďte statistickou významnost, a když zjistíte, že zjištěný výsledek je statisticky významný, vypočítejte míru věcné významnosti a interpretujte ji. Právě první krok ostře odsoudil Thompson [1997] či Cahan [2000], kteří upozorňují, že díky tomuto budou výsledky statisticky nevýznamné opominuty z hlediska věcné významnosti a navíc nebudou vůbec publikovány. Thompson [1997, 1998a, 1998b, 1999, 2002a] navrhuje, aby při posuzování výsledků bylo užito tří postupů paralelně vedle sebe:

1) výpočet intervalu spolehlivosti a hodnocení výsledku prismatem statistiky,
2) výpočet některé míry věcné významnosti a její interpretace a
3) výpočet intervalu spolehlivosti míry věcné významnosti a jeho interpretace.

Tato svá doporučení díky své autoritě prosadil Thompson i do zásad, které zakotvila Americká psychologická asociace [APA 2001] a Americká asociace vzdělávacího výzkumu [AERA 2006]. V České republice nabídl doporučení pro práci s věcnou a statistickou významností několikrát Blahuš [Čelikovský a kol. 1979: 264] a jeho postup je schematicky tento:

1) nejprve posuďte věcnou významnost výsledků a
2) poté posuďte zobecnitelnost výsledku z výběru na základní soubor pomocí statistického testu.

Debaty o problémech statistické a věcné významnosti se v posledních deseti letech odehrávají na stránkách světových časopisů, za zmínku stojí diskuse Levin&Robinson vs. Thompson [Thompson 1996; Robinson, Levin, 1997; Thompson 1997; Cahan 2000] dále pak diskuse mezi Biskinem a Thompsonem [Vacha-Haase; Bruce Thompson 1998, Biskin 1998].
V časopise Psychological Science (1997, vol.8) byla celá sekce nazvaná „Zrušit testy statistické významnosti“. Z diskuse také vznikla celá kniha nazvaná výmluvně What if there were no significance test? [Harlow, Mulaik, Steiger 1997], Americká psychologická asociace vytvořila pracovní skupinu, která se snažila problém řešit [APA Task Force, 1999]. Nejvýraznější postavou diskusí je profesor Thompson, který napsal více než 30 článků a postupně opustil své radikální postoje a hájí dnes věcnou významnost a snaží se prosazovat její užívání do odborných časopisů.

A co klinická či ekonomická významnost?

Výtky proti statistické významnosti a podpora užívání věcné významnosti ovšem neznamenají konec diskusí. Zejména z medicínských oborů zaznívají výtky proti věcné významnosti a navrhuje se užívat klinickou významnost. I česká medicína již tento koncept užívá, ukázkou může být třeba Salajkův text [Salajka 2001], o klinické významnosti hovoří i učebnice medicínské statistiky [Euromise]. Stranou této diskuse nezůstal ani Thompson [2002b], detailní informace o klinické významnosti nabízí Campbell [2005]. Definici klinické významnosti podává Kazdin [1999:332]: „Klinická významnost vypovídá o praktické hodnotě nebo důležitosti dopadu intervence-tj. zda intervence má skutečné (tedy pravé, hmatatelné, praktické, zřetelné) dopady na každodenní život pacientů, nebo těch, se kterými se pacienti setkávají“. Klinická významnost tedy sleduje dopady intervencí lékařských, psychologických na jedince a zjišťuje, zda došlo ke změnám či nikoliv. Samozřejmě problematické je měření změn, nabízí se subjektivní měření pocitů pacientů nebo měření

objektivních charakteristik (např. krevní tlak, nebo sofistikované metody jako je CT či MRI[60] vyšetření apod.). Často se klinická významnost měří jako procento pacientů, u nichž došlo ke zlepšením, případně procento pacientů, kteří se po intervenci vrátili do normy (tedy výsledky jejich vyšetření jsou srovnatelné s normální zdravou populací). Samozřejmě, že definice normy je mnohdy problematická a klinická významnost má tedy též své nedostatky. V sociologii zřejmě klinickou významnost často nepotřebujeme, nicméně například v sociodiagnostice by jistě mohla najít uplatnění. O tom, že je potřebná v oborech, kde jde o intervence, a míří se přímo na jedince, není sporu.

Někteří autoři ovšem jdou dále a konstruují další typy významnosti. Pro posouzení ekonomického dopadu rozdílů mezi skupinami navrhl Levin se svými spolupracovníky užívání významnosti ekonomické [Onwuegbuzie, Levin, Leech 2003: 39]. Ta vychází z úvahy, že ani statistická ani věcná významnost nepracují s kategoriemi reálného života - penězi. Proto, aby bylo možné posoudit rozdíly mezi skupinou kontrolní a experimentální navrhli užívat vyjádření efektu v penězích. Je tedy například možné zkoumat, kolik se ušetří, když se pacienti vyléčí za pomoci určitého léku, kolik peněz ušetří společnost na sociálních dávkách, pokud zavede předškolní výukové programy pro určité žáky apod.

Důsledky následování doporučení oponentů statistické významnosti

Nutné je zamyslet se nad důsledky doporučení užívat míry věcné významnosti společně se statistickou významností. V zajímavé studii [Posavac, Sinacore 1984] autoři prokázali, že znalost konceptu věcné významnosti a jeho měr pomáhá studentům nepřeceňovat statisticky významné výsledky. V novější studii zaměřené na míry věcné významnosti a odhady velikosti výběrových souborů [Robinson, Fouladi, Williams, Bera 2002] dospěli autoři k závěru, že pokud studentům dáme informaci o míře věcné významnosti, mají často sklon přeceňovat důležitost takového výsledku ve srovnání se situací, kdy tuto informaci nemají. Mnohem horší dopady má dle autorů studie praxe, kdy dochází ke zveřejňování velikosti výběrových souborů pro získání statisticky významných výsledků. Studenti, kteří si například přečtou, že kdyby byl výběr dvakrát větší, byl by již rozdíl statisticky významný, uvažují naprosto nesprávně. Namísto toho, aby je tato infromace varovala a upozornila na problémy statistické významnosti, většina z nich reaguje radostně v duchu hesla: „Stačí navýšit výběr a

[60] CT-počítačová tomografie, MRI-magnetická rezonance.

výsledek bude statisticky významný“. Na základě uvedených skutečností je nutno zvážit, zda všechna doporučení proponentů věcné významnosti jsou v praxi vhodná.

Další zdroje k poučení

Případným zájemcům o další informace lze doporučit zejména knihy či sborníky věnované tématu. První publikace pochází již z roku 1970 [Morisson, Henkel 1970]. Dalšími tituly jsou Sense and nonsense of statistical inference [Wang, 1993], Contrasts and Effect Sizes in Behavioral Research [Rosenthal, Rosnow,. Rubin. 2000]. Z nedávné doby pochází kniha Beyond the statistical testing [Kline 2004]. Samozřejmě čerpat poznatky lze i z obrovské časopisecké literatury. V téměř každé disciplíně existuje alespoň úvodní článek upozorňující na věcnou nebo klinickou významnost [Deal, Anderson 1995, Anderson, Burnham, Thompson 2000, Fan 2001, Ives 2003, Meline, Wang 2004, Buhi 2005, Campbell 2005, Sink, Stroh 2006 Watkins, Revers, Rowel, Green, Revers 2006].

Shrnutí a doporučení

Cílem tohoto textu je komplexněji představit českému čtenáři koncepci věcné význmnosti, měr věcné významnosti a popsat diskusi o užívání věcné a statistické významnosti. Je škoda, že v českých poměrech se toto téma nediskutuje a i ve výuce je toto téma opomíjeno. Další snahy je tedy třeba zaměřit na tyto cílové skupiny [Kirk 2001: 216]:

1) Učitele metodologických a statistických předmětů,
2) autory metodologických a statistických textů,
3) redakční rady časopisů a na
4) tvůrce publikačních manuálů.

Všechny tyto osoby mohou vylepšit stávající problematickou práci se statistickou významností a zajistit častější užívání věcné významnosti. Klíčové je z mého pohledu zejména působení učitelů, kteří vychovávají další vědecké generace a také autorů učebních textů. Zatím máme ze statistické oblasti v ČR pouze vhodný text Hendlův [2004] a z metodologické oblasti překlad textu Ferjenčíka [2000]. Je ovšem otázkou nakolik jsou tyto texty užívány pro vzdělávání sociologů, psychologů či pedagogů. Na nedostatky učebnic a výuky v zahraničí již upozornily mnohé analýzy [Kliner, Leech, Morgan. 2002, Halley, Krauss. 2002, Finch, Cumming, Thomason 2001, Robinson, Fouladi, Williams, Bera. 2002]. V České republice na kritické a analytické zhodnocení výuky a učebních textů teprve čekáme.

Dále je třeba zahrnout praktiky správného užívání věcné a statistické významnosti do publikačních manuálů jednotlivých profesních asociací a časopisů a prosazovat je v redakčních radách časopisů. Jako minimum lze doporučit následující:

1) Vypočtení, publikace a interpretace výsledků statistické významnosti.
2) Vypočtení, publikace a interpretace měr věcné významnosti.

Je třeba nejen počítat různé míry, ale snažit se je interpretovat a posoudit praktický dopad výsledků. Rozhodně je nutno odsoudit automatizovaný způsob analýzy dat, který je založen na produkci tabulek bez jejich hlubší interpretace a pokusu o porozumění. Osobně se přimlouvám za citlivé posouzení publikačních náležitostí v jednotlivých redakčních radách, radikální boj může mnohdy přinést více problémů než užitku. Obdobně i ve výuce metodologie a analýzy dat ve společenských vědách je nutné zmínit problematiku věcné významnosti a jejího měření. Ještě mnohem potřebnější je vysvětlit správně koncept významnosti statistické a zejména poukázat na jeho omezení. Jen tak je možné bránit špatným interpretacím a zneužívání. Samozřejmě nejplodnější strategií je srovnávat výsledky s výsledky jiných autorů v rámci ČR, ale zejména mezinárodně. Vhodné by bylo vytvořit databázi, kam by se výsledky jednotlivých studií zaznamenávaly a bylo možné s nimi dále metaanalyticky pracovat. Zatím máme pouze databáze dat, ale dohledat jednotlivé výsledky z nich vypočítané jednoduše nelze.

Realistické je ale očekávat, že ke změnám nedojde hned, psychologické kořeny tohoto podává opakovaně Thompson [1998a, 1999, 2002a]. Výraznější změny lze čekat až od další generace vědců, která bude podrobena upravené výuce. Hlavní výzvou pro stávající generaci vědců proto je vyhrát boj s vlastní pohodlností (nic nového se neučit, nad výsledky nepřemýšlet a pracovat postaru) a zkusit změnit vědecké praktiky u sebe sama v duchu nejnovějších poznatků. Je to obtížné, ale efektivní. Zkusme to, výsledky za to stojí.

Literatura citovaná v kapitole 4

AERA. 2006. *Standards for Reporting on Empirical Social Science Research in AERA Publications.* American Educational Research Association.

Algina, J. , B. C. Moulder. 2001. Sample Sizes for Confidence Intervals on the Increase in the Squared Multiple Correlation Coefficient. *Educational and Psychological Measurement*. 61 (4): 633-649.

Anderson, D.R., K. P., Burnham, W. L. Thompson,.2000. Null hypothesis testing: Problem, prevalence and alternative. *Journal of wildlife management*. 64(4): 912-923.

APA. 2001. *Publication manual of the American Psychological Association, 5th edition*. Washington DC.

APA Task Force in Statistical Inference. 1999. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*. 54: 594-604.

Biskin, B. H. 1998. Comment on significance testing. *Measurement and Evaluation in Counseling and Development*. 31(1): 58-62.

Blahuš, P. 2000. Statistická významnost proti vědecké průkaznosti výsledků výzkumu. *Česká kinantropologie* 4(2): 53-72.

Boring, E., G. 1919 Mathematical versus statistical significance. *Psychological Bulletin*. 15: 335-338.

Buhi, E., R. 2005. The Insignificance of "Significance" Tests: Three Recommendations for Health education research. *American Journal of Health Education.* .36(2): 109-112.

Cahan, S. 2000. Statistical significance is not a „kosher certificate“ for observed effects: A critical analysis of the two-step approach to the evaluation of emprirical results. *Educational Researcher*. 29 (1): 31-34.

Campbell, T. C. 2005 An Introduction to Clinical Significance: An Alternative Index of Intervention Effect for Group Experimental Designs. *Journal of Early Intervention*. 27 (3): 210–227.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Science (2nd ed.).* Hillsdale (NJ): Erlbaum.

Cox, D. R. 1982. Statistical significance tests. *British Journal of clinical Pharmacology*. 14 : 325-331.

Cumming, G., S. Finch. 2001. A Primer on the Understanding, Use, and Calculation of Confidence Intervals that are Based on Central and Noncentral Distributions. *Educational and Psychological Measurement*. 61(4): 532-574.

Čelikovský, S. a kol. 1979. *Antropomotorika.* Praha : Státní pedagogické nakladatelství.

Deal, J., E., E.R. Anderson. 1995. Reporting and Interpreting results in family research *Journal of Marriage and Family*. Vol. 57: 1040-1048.

Euromise. Základy statistiky pro biomedicínské obory. http://ucebnice.euromise.cz/index.php?conn=0&section=biostat1

Fan, X. 2001. Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*. 94 (5): 275-282.

Fan, X., B. Thompson. 2001. Confidence Intervals for Effect Sizes: Confidence Intervals about Score Reliability Coefficients, Please: An EPM Guidelines Editorial. Educational and *Psychological Measurement.* 61(4): 517-531.

Ferjenčík, J. 2000. *Úvod do metodologie psychologického výzkumu*. Praha: Portál

Fidler, F., B. Thompson B. 2001. Computing Correct Confidence Intervals for Anova Fixed- and Random-Effects Effect Sizes. *Educational and Psychological Measurement*. 61(4): 575-604.

Finch, S., G. Cumming, N. Thomason. 2001. Colloquium on Effect Sizes: the Roles of Editors, Textbook Authors, and the Publication Manual: Reporting of Statistical Inference in the Journal of Applied Psychology: Little Evidence of Reform**.** *Educational and Psychological Measurement*. 61(2): 181-210.

Fisher, R. A. 1925. *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Halley, 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.

Hays, W. L. 1963. *Statistics for psychologists*. New York: Holt, Rinehart & Winston.

Hebák, P. (ed.) 2005. *Vícerozměrné statistické metody (3).* Praha: Informatorium.

Hendl, J. 2004. *Přehled statistických metod zpracování dat: analýza a metaanalýza dat*. Praha: Portál.

Ives , B. 2003. Effect size use in studies of learning disabilities. *Journal of Learning Disabilities.* 36 (6): 490-504.

Kazdin, A.E.. 1999. The meanings and measurement of clinical significance. *Journal of consulting and clinical psychology*. 67: 332-339.

King, G. 1986. How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science*. 30(3): 666-687.

Kirk, R. 1996. Practical significance: A concept whose time has come. *Educational and Psychological Measurement.* 56(5): 746-759.

Kirk, R., E.. 2001. Promoting Good Statistical Practices: Some Suggestions. *Educational and Psychological Measurement.* 61(2): 213-218.

Kline, R. B. 2004. *Beyond the statistical testing. Reforming data analysis methods in behavioral research.* Washington, DC: American Psychological Association.

Kliner, J.A., N. L Leech, G. Morgan. 2002. Problems with Null Hypothesis Significance Testing (NHST): What do the textbooks say. *The Journal of Experimental Education*. 71(1): 83-92.

Loftus, G. R. 1996. Psychology will be a Much Better Science When We Change the Way We Analyze Data. *Current Directions in Psychological Science.* 1: 161-171.

Meline, T.;, B. Wang. .2004. Effect-Size Reporting Practices in AJSLP and Other ASHA Journals, 1999-2003. *American Journal of Speech - Language Patology*. 13(3): 202-207

Morisson, D. E., R., E. Henkel. 1970. *The significance test controversy – a reader*.Chicago: Aldine

Onwuegbuzie, Anthony J., J.R. Levin, N. L. Leech. 2003. Do Effect-Size Measures Measure Up?: A Brief Assessment. Learning Disabilities: *A Contemporary Journal.* 1(1): 37-40.

Posavac, E. J.; Sinacore, J. M. 1984. Improving the Understanding of Statistical Significance: Reporting Effect Size. *Knowledge*. 5(4): 503-508.

Robinson, D. H., R.T. Fouladi, N. J. Williams, S. J. Bera. 2002. Some effects of including effect size and "what if" information. *The Journal of Experimental Education.* 70(4): 365-382.

Robinson, D., H., J., R. Levin. 1997. Reflections on statistical and substantive significance with a slice of replication. *Educational Researcher*. 26(5): 21-27.

Rosenthal, R., R. L. Rosnow, D. B. Rubin. 2000. *Contrasts and Effect Sizes in Behavioral Research: A Correlational Approach*. Cambridge University Press.

Rozeboom, W. W. 1960. The fallacy of the null hypothesis significance test. *Psychological Bulletin*. 57: 416-428.

Schmidt, F. 1996. Statistical significance testing: implications for the training of researchers. *Psychological Methods*, 1(2): 115-129.

Selvin, H., C. 1957 A Critique of Tests of Significance in Survey Research American *Sociological Review*. 22 (5): 519-527.

Salajka, F. 2001. Bronchiální astma a kvalita života nemocných. *Alergie* 2(2): 68-70.

Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010). *Prvotní zjištění z Mezinárodní studie občanské výchovy*. Praha: ÚIV: International Association for the Evaluation of Educational Achievement (IEA).

Sink, Ch.,A.,H., R. Stroh. 2006. Practical Significance: The Use of Effect Sizes in School *Counseling Research Professional School Counseling.* 9(5): 401-411.

Smithson, M. 2001. Correct Confidence Intervals for Various Regression Effect Sizes and Parameters: The Importance of Noncentral Distributions in Computing Intervals. *Educational and Psychological Measurement.* 61(4): 605-632.

Soukup, P. 2006. Proč užívat hierarchické lineární modely. *Sociologický časopis* 42(5):987-1012.

Tailor, K. W., J. Frideres. 1972. Issues Versus Controversies: Substantive and Statistical Significance. *American Sociological Review*. 37(4): 464-472.

Thompson, B. 1996. AERA Editorial policies regarding statistical significance tests: three suggested reforms. *Educational Researcher*. 25(2): 26-30.

Thompson, B.1997. Editorial policies regarding statistical significance tests: further comments. *Educational Researcher*. 26(5): 29-32.

Thompson, B. 1998a. Statistical significance and effect size reporting: Portrait of a possible future. *Research in the schools*. 5(2): 33-38

Thompson, B. 1998b. Five Methodology Errors in Educational Research: The Pantheon of Statistical Significance and Other Faux Pas Invited address (Divisions E, D, and C) presented at the annual meeting (session #25.66) of the American Educational Research Association, San Diego.

Thompson, B. 1999. Why "encouraging" effect size reporting is not working: The etiology of research. *The Journal of Psychology.* 2: 133-139.

Thompson, B. 2002a. What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*. 31(3): 24-31.

Thompson, B. 2002b. "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider. *Journal of Counseling and Development*. 80 (1): 64-71.

Vacha-Haase, T, B. Thompson. 1998. Further comments on statistical significance tests. *Measurement and Evaluation in Counseling and Development*. 31(1): 63-68.

Vacha-Haase, T.,, B. Thompson 2004. How to estimate and interpret various effect sizes. *Journal of Counseling Psychology*. 51(4): 473-481.

Wang, Ch. 1993. *Sense and Nonsense of Statistical inference*. Dekker 1993.

Watkins, D., C, D. Revers, K., L , Rowell, B., L.. Green, B. Revers. 2006. A Closer Look at Effect Sizes and Their Relevance to Health Education. *American Journal of Health Education*. 37(2): 103-108.

5. P a d (Používání statistické a věcné významnosti v českých sociálních vědách)[61]

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 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 zaměřen (2008-2009) a i tématické zaměření je poměrně úzké (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 3 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?*

[61] Název P a d 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. *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 primárně zaměřená na produkci Sociologického časopisu, nicméně pro zohlednění kontextu v ČR je provedeno též srovnání 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 závěru jsou ještě 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 mnohočetného 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 20. letech 20. století se objevují texty, které kritizuje 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 50. let a v 60.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 80. 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 Harlow, Mulaik a 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 zaměřena sekce časopisu Psychological Science (1997, vol.8) nazvaná „Zrušit 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 pokud možno se 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 byl ostatně zakotven 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 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 6

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. Návdavkem 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, resp. 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 Ziliak[62] a McCloskey [2008].

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 zdena počátku popíšeme dvě zřejmě neslavnější studie, jednu zaměřenou na American Sociological review [Morrison, Henkel, 1969] a další zaměřenou na American Econnomic Review [Zliak, McKloskey, 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 na výzkumnickou praxi vliv. 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 nejpíše první studií tohoto typu a je nadto ukázkou monotématicky zaměřené studie, tj.

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

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 McKloskey 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 McKloskey 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 McKloskey [2008: 67-73]). S ohledem na detailnost jejich zkoumání (i mírně jiné oborové zaměření) zde zmíníme jen některé výsledky. První obecný závěr Zliaka a McKloskey 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 pozonost 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 závěru 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 Bernardi a kol. [2017], která se zaměřuje na texty publikované v European Sociological Review (mezi lety 2000–2004 resp. 2010–2014) a které využívají regresní metodologii. Cílem textu je zmapovat užívání statistické významnosti v těchto textech (celkem 356 článků), autoři použili 5 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 resp. 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 velice 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 prezentován v tomto článku práci. Za pozornost stojí práce Oakese [1986] a Hallera s Krause [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, resp. publikace po rozšíření výzkumu, bude námětem samostatného článku).

Další zajímavou studií byl výzkum členů AERA prezentovaný v článku Mittag a Thompson [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 postavila za užívání mechanických postupů statistického testování reprezentovaného stepwise postupem výběru proměnných. Poslední zde uvedenou studií je článek DeVaney [2001], který je založen na výzkumu šéfredaktorů časopisů z oblasti pedagogiky, psychologie a sociologie. Výzkum byl proveden na širokém souboru 311 časopisů. Dle zjištění DeVaney 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. Většina šéfredaktorů 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 nesprávně statistické testy, 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.

Metodologie provedeného výzkumu

S ohledem na výše uvedené příklady empirických studií se předložený text snaží komplexněji zmapovat využívání koncepcí statistické a věcné významnosti (po vzoru Zliaka a McCloskey). Nebylo možné automaticky přijmout kritéria užívaná Zliakem a McCloskey [2008], byť sloužila pro inspiraci. Analytické kategorie navazují přímo na předchozí v Česku publikované texty Soukupa a Rabušice [2010] a Soukupa [2010, 2013]. Výsledky je tak možné částečně srovnat s výsledky Cuberek a Frömela [2011]. Autor článku provedl obsahovou analýzu produkce tří časopisů 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 mechanickou práci s daty a výsledky analýz (druhá dílčí výzkumná otázka),
III. věcnou interpretaci výsledků, používání měr věcné významnosti a jejich interpretaci (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 15 náhodně vybraných textů.

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

a) z censu
b) ze záměrného výběru (kvótního, dostupného apod.)

c) z malého výběru[63]
d) z výběru s extrémně velkým počtem vybraných jednotek či daty spojenými z různých datových souborů[64]

Ve všech výše uvedených případech je užívání statistických testů (resp. intervalů spolehlivosti) bez dalšího (srov. dále uvedené exkurzy) 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, že v této oblasti ovšem probíhají výrazné diskuse a o jejich stručné zachycení se pokoušejí dva dále 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 článku, tak i jednotlivým analýzám[65]).

ad II. U nesprávného užívání statistické významnosti (ad II.) 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),

b) záměně 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, resp. interpretace v rozporu s těmito výsledky

[63] Pro různé testy je kritérium malého výběru obecně různé, nicméně platí, že výběry do cca 30-50 jednotek bývají považovány za malé a vyžadují pro případ, že rozdělení zkoumaných veličin není normální, užívání neparametrické statistiky místo parametrické.
[64] 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.
[65] Některé prezentované výsledky mají jako analytickou jednotku článek (platí pro oblast I), některé pak používají též jako analytickou jednotku 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).

Pro obtížnost, respektive častou nemožnost, rozlišování případů pod písmeny b) a c) zejména v případě Československé psychologie[66] 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 16 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 v příloze 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 jeich interpretace je sledováno, zda autor nějak věcně výsledky komentuje).[67]

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

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

jen na posledních deseti letech, tj. 2005-2014[68]). Výběr těchto časopisů ke srovnání byl veden těmito kritérii[69]:

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.[70]
   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 působící 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 WoS).
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 světové sociologii zejména v 60. letech, v pedagogice a psychologii pak v posledních třiceti letech). Připomeňme, že studie DeVaney [2001] citovaná v přehledu obdobných studií měla totožné zaměření (tj. časopisy z oblasti sociologie, psychologie a pedagogiky). Nezanedbatelná je oborová blízkost, připomeňme, že

[68] 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.

[69] Díky popsaným kritériím je zřejmé, že nebylo možné najít v zásadě jiné časopisy. Politologické časopisy téměř kvantitativní analýzy 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.

[70] 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 ozebírat metodiku hodnocení české vědy (částečně je jí věnována pasáž na konic článku), 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).

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 body I-III výše uvedené). S ohledem na maximální rozsah jednoho článku není zcela jistě předložená analýza vyčerpávající, data a téma bude možné rozvíjet v dalších textech (získaná data budou veřejně dostupná k další analýze). Doplňkově jsou na konci články provedeny analýzy zaměřeny 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. [71]

Dodejme, že pro potřeby orientačního sledování změn v čase u Sociologického časopisu bylo období 2005-2014 rozděleno na čtyři pětiletky[72], 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

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

[72] Toto členění poměrně dobře odděluje produkci z 90. 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.

členěny dle použité statistické techniky a umožňují tak konkrétní poučení, jaké problémy se u jednotlivých statistických technik objevují.

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

Exkurz o nadpopulacích[73] (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 data za populaci lze pak považovat za výběr z této nadpopulace vzniklý náhodnými procesy ve společnosti.[74] 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 utoři 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

[73] Tento i následující exkurz vznikl na základě podnětu jednoho z anonymních recenzentů, za což mu patří dík. Za případné nedostatky obou exkurzů odpovídá samozřejmě autor článku.
[74] Tato úvaha se blíží teoriím alternativních historií, z nichž my pozorujeme pouze jedinou, a je pro mnoho vědců zcela nepřijatelná.

bayesovského přístupu má být preferováno před klasickými testy pro celopopulační data, přináší dle nich 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 další 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 (tj. 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 tuto kapitolu. 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 některými názory preferována bayesovská statistika není (bohužel?) v českých sociálně vědních časopisech užívána, bude analýza článků sledovat, zda autoři užívají pro celopopulační data klasické statistické testy a bude tento postup považovat za nevhodný (tedy postup bude souladný s textem Soukupa a Rabušice, 2007). Pro úplnost dodejme, že oblíbené víceúrovňové modely, které jako druhou úroveň využívají země, vychází 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 v této oblasti, 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; Lévy, Lemeshow, 2008). Tento přístup by se dal označit jako klasický či tradiční. Zhruba od 70. 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 možností, 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 vybrání (prvním přehledovým textem srovnávajícím oba přístupy je zřejmě text Sarndal, 1978). Zatímco problémem pro designový přístup je 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:

715-716]. Proto bude užívání statistických testů pro nenáhodné výběry hodnoceno jako nevhodná praxe (tedy postup bude opět souladný s textem Soukupa a Rabušice, 2007).

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

Během uplynulých dvaceti let bylo na stránkách Sociologického časopisu publikováno 162 textů, které využívají postupy kvantitativní analýzy.[75] 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 deseti letech otištěno těchto textů výrazně více než v Sociologickém časopise (155, 10 z nich byly psychometrické analýzy a ty jsou z následujících výsledků vyloučeny[76]).

Detailní přehled počtu článků zaměřených na kvantitativní analýzu v Sociologickém časopise uvádí graf č. 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 2 texty). Samozřejmě část uvedených výkyvů má 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.

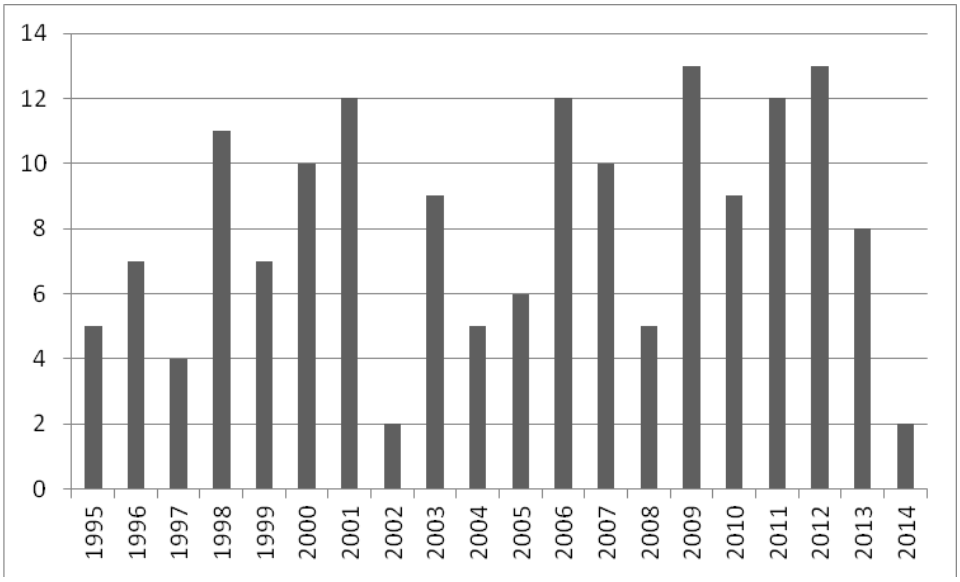
Při detailnější analýze se lze zaměřit na jednotlivá čísla a nalézt tak vydání Sociologického časopisu s nejvyšším počtem kvantitativně laděných textů. Vítězem této pomyslné soutěže bylo první číslo vydané v roce 2011, kde bylo námi sledovaných textů celkem devět (šlo o monotématické číslo zaměřené na 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 č.2 vývoj počtu

[75] 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 zcela srozumitelná každému 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 seznámit se detailně s kompletní kvantitativní produkcí všech tří časopisů.

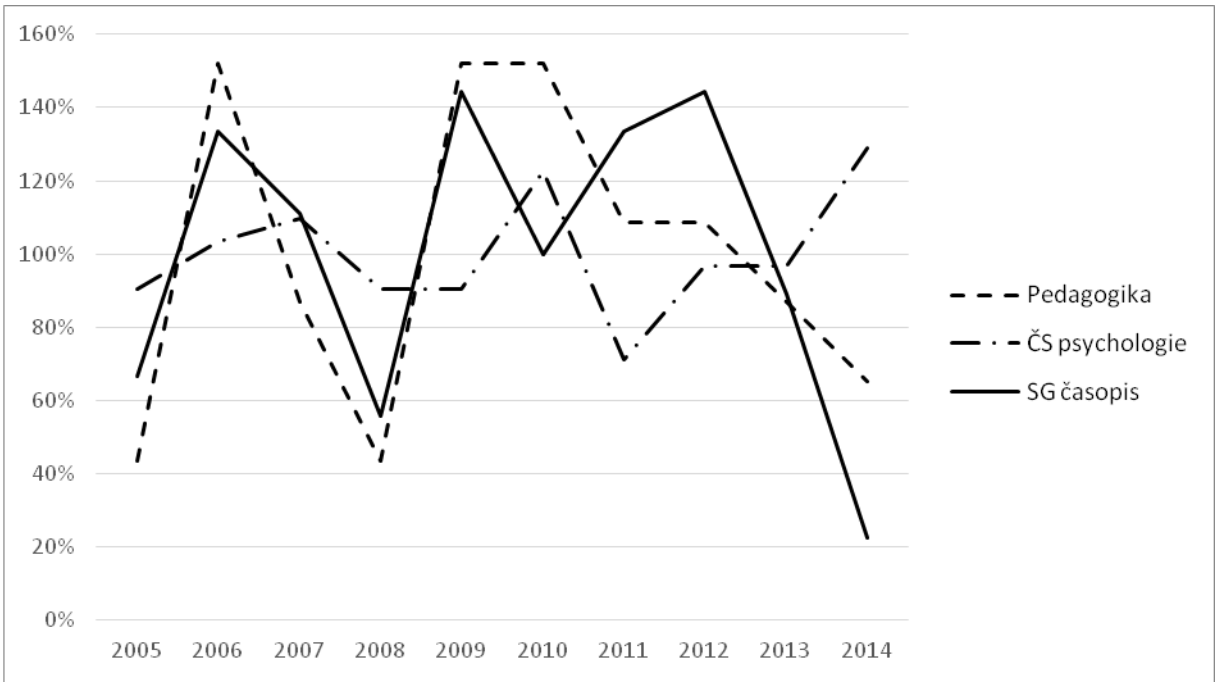
[76] Psychometrické analýzy ověřují běžně v českém kontextu fungování měřících škál. Typicky (a nejde zde o chybu) jsou prováděny na záměrných výběrech a jejich zařazení by tak neférově zhoršovalo výsledky zjištěné pro Československou psychologii. Dodejme, že některé zahraniční časopisy však vyžadují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

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.

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

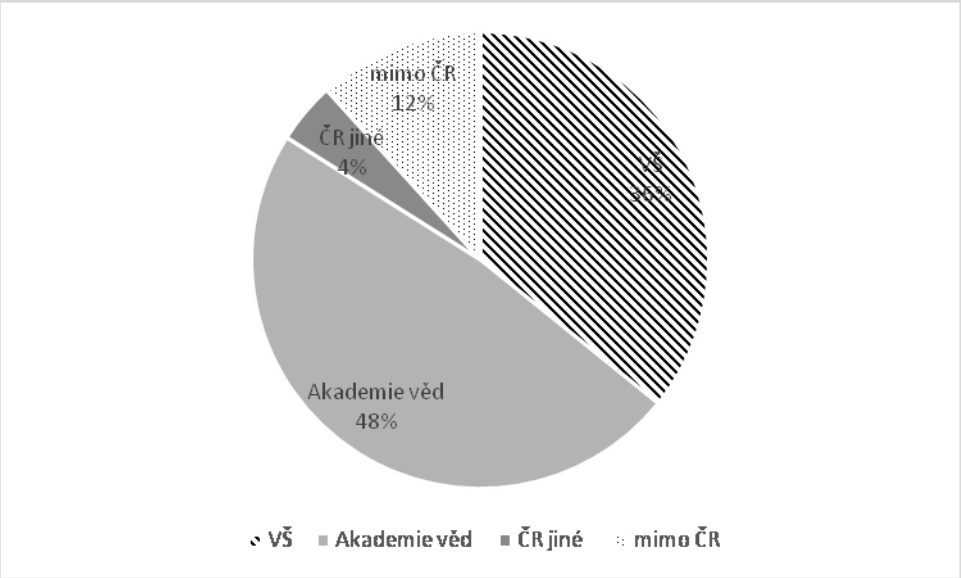


Graf 2. Vývoj počtu článků zaměřených na kvantitativní analýzu (100%=průměr sledovaného období) publikovaných v letech 2005-2014 (dle časopisu)



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

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



Téměř polovina textů v Sociologickém časopisu 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.[77]

Předposlední popisnou informací o analyzovaném souboru je podíl analyzovaných textů, které byly publikovány česky resp. anglicky. Podíl anglicky psaných kvantitativně laděných textů v Sociologickém časopise je poměrně překvapivě vysoký (41 %)[78] a s ohledem na předchozí rozbor institucionálního zázemí, je nutné konstatovat, že většinu těchto textů publikovali autoři působící v Česku (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

[77] 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 má zcela specifické postavení a publikační aktivita jeho pracovníků na stránkách časopisu, který ústav sám vydává, je enormní.
[78] V Pedagogice nebyl za celé sledované období publikován anglicky jediný text, v Československé psychologii celkem 8 článků (tj. 6 % kvantitativní produkce).

je zajímavé též zjištění, že zatímco u českých textů je podíl autorství mezi českými VŠ a Akademií věd téměř vyrovnán (42 resp. 47 textů) u textů v angličtině je výrazný nepoměr (16 vs. 31 textů). Tyto rozdíly lze též číst jako potvrzení většího tlaku na mezinárodní ohlas publikací na půdě Akademie věd, resp. 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 žádoucí, aby autoři uváděli v textech 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 dvaceti letech 3,5 roku (směrodatná odchylka činila necelé dva roky)[79]. Nutno zároveň dodat, že tato doba postupně narůstá, protože mezi lety 1995-1999 to bylo průměrně jen 2 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 žádoucí, aby k publikaci docházelo co nejrychleji od získání výzkumných dat, aby byly publikované výsledky ještě platné pro zkoumanou společnost. 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.

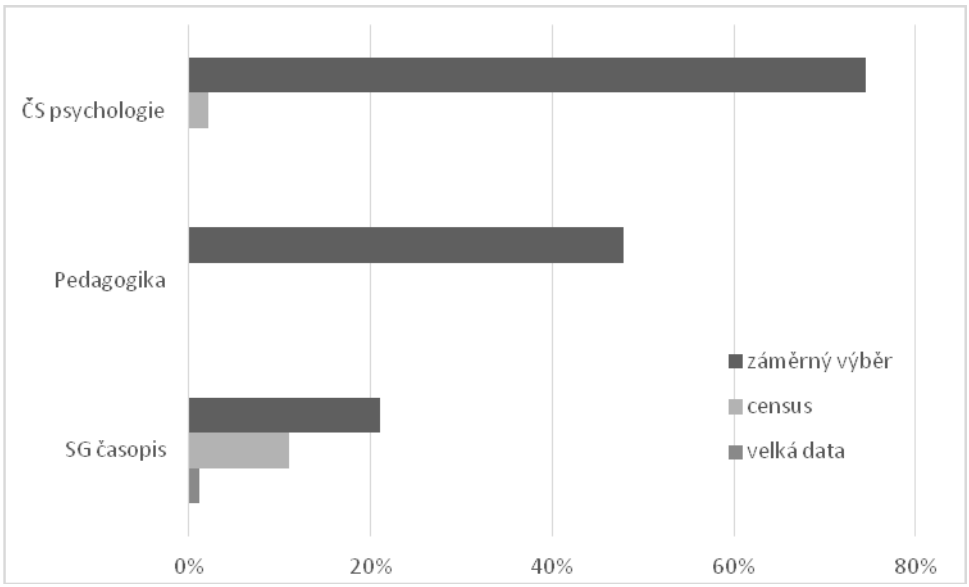
Nesprávné používání statistických testů

a) Srovnání časopisů (2005-2014)

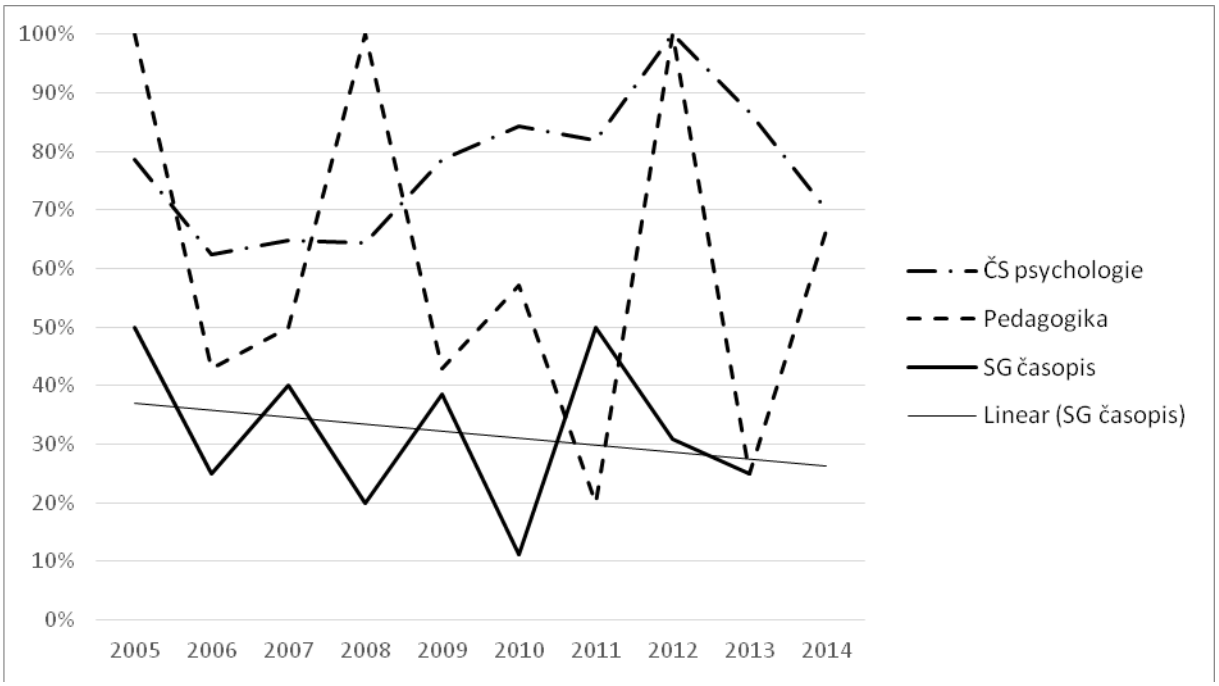
Nyní se zaměříme na první dílčí výzkumnou otázku, tj. budeme sledovat, zda se v analyzovaných článcích používají statistické testy i v případech, kdy to analyzovaná data neumožňují, resp. to není vhodné (srov. výše část věnovanou metodologii). Základní přehled situace ve všech sledovaných časopisech v posledních deseti letech podávají grafy č. 4 a 5.

[79] Maximální doba mezi sběrem dat a publikací činila 12 let, pětina textů pak vycházela z dat, která byla maximálně 1 rok stará.

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



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

V případě Sociologického časopisu autoři zhruba v pětině případů užívali statistické testy pro data z kvótních výběrů a cca v desetině v případů pro data, která byla získána pro celou zkoumanou populaci. *Celková míra nesprávného užívání statistických testů tedy činila v případě Sociologického časopisu cca jednu třetinu.* 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 77 %![80] 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 lepší, než byla v American Sociological Review v 50. a 60. letech minulého století, obdobně jsou výsledky v zásadě příznové při srovnání se studií 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).

b) Detailnější analýza Sociologického časopisu (1995-2014)

Nahlédneme-li na dvacetiletí Sociologického časopisu, pak se míra chybné aplikace statistických testů ještě sníží (klesne na 23 %). Z toho logicky plyne, že v posledním desetiletí se situace zhoršila. Pro detailnější analytický náhled po pětiletkách využijeme tabulku č. 1.

Tabulka 1. Podíly článků v Sociologickém časopise zaměřených na kvantitativní analýzu (1995-2014), kde se nesprávně 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 32 % (tj. více než trojnásobek!).

[80] Cuberek a Frőmel [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.

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)[81]. Na okraj dodejme, že při srovnání článků psaných česky a anglicky jsou výsledky v zásadě totožné.

Nesprávná interpretace statistické významnosti a mechanické využívání statistiky

a) 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 č. 6.

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

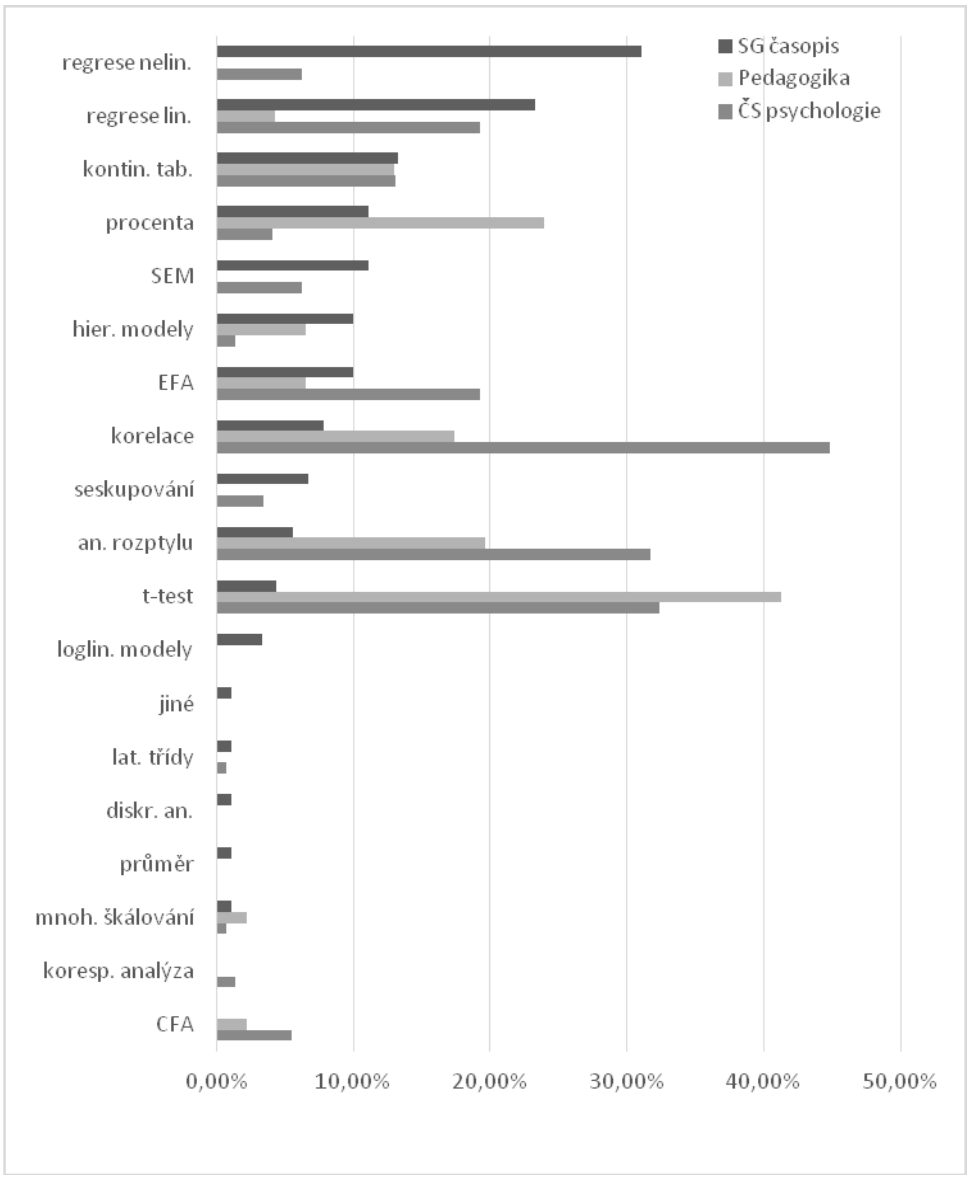
[81] Zde je ale nutno poznamenat, že tito autoři byli prvními autory jen u sedmi článků a v jediném bylo využito statistického testování pro nevhodná data (tvořila celý základní soubor). Nelze z toho tedy dovozovat žádné závěry.

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[82]). 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.
V Sociologickém časopise byla nadto třikrát za posledních deset let užita loglineární analýza, která se v dalších dvou časopisech nevyskytuje (nadto v předchozím desetiletí byly hojně užívány i velice příbuzné logitové modely, srov. další popis). 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).

Po základním popisu technik užitých v článcích publikovaných ve všech třech sledovaných časopisech v posledních deseti letech ještě proveďme podrobnější rozbor technik užívaných v Sociologickém časopise za 20 let.
Nejdříve pro zajímavost (ocení to jistě historici české sociologie) v tabulce č. 2 uvádíme, kdy poprvé bylo v Sociologickém časopise použito jednotlivých složitějších statistických technik v rámci sledovaného období posledních dvaceti let.

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

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



Poznámka:
EFA=explorační faktorová analýza
CFA=konfirmační faktorová analýza
SEM=strukturní modelování

Tabulka 2. Rok, kdy byla poprvé v Sociologickém časopise použita statistická technika v období 1995-2014

Technika:	**rok prvního užití**
nelineární regrese (zejm. logistická)	1995
konfirmační faktorová analýza	1995
strukturní modely	1995
loglineární modely	1997
logitové modely	1998
korespondenční analýza	1998
explorační faktorová analýza	1999
diskriminační analýza	1999
seskupovací analýza	2001
mnohorozměrné škálování	2004
hierarchické lineární modely	2006
analýza latentních tříd	2007

Z přehledu je zřejmé, že většina složitějších statistických techniky (např. strukturní modely a logistická regrese) byla používána již v 90. letech.[83] Relativní novinkou je užívání hierarchických lineárních modelů (2006) a analýzy latentních tříd (2007). Následující tabulka č. 3 pak sleduje, k jakým změnám v užívání složitějších statistických technik docházelo na stránkách Sociologického časopisu v čase.

Pokud se zaměříme na nejpoužívanější techniky, pak lze vysledovat určité tendence. Používání korelací v posledních pěti letech vymizelo (obdobná je situace u analýzy rozptylu a procent), lineární regrese se užívá zhruba stále ve stejné míře. Nelineární regresní modely (převážně binární logistická regrese) se v posledním desetiletí užívá častěji než v předchozím desetiletí.
U strukturních modelů lze vypozorovat jistou renesanci v užívání (poměrně často se užívaly v druhé polovině devadesátých let a objevují se častěji i v poslední pětiletce). Poslední zajímavostí je vymizení logitových modelů v posledních deseti letech, což lze snadno vysvětlit ukončením publikační aktivity Blanky Řehákové, která tuto techniku v české sociologii jako jediná užívala. Celkově nelze konstatovat, že by docházelo k nárůstu nebo

[83] V mnoha případech ještě před rokem 1995, ostatně například první metodologický text věnovaný strukturnímu modelování otiskl Sociologický časopis již v roce 1989 [Matějů, 1989].

poklesu v užívání složitějších statistických technik a lze hovořit spíše o stagnaci. Lze tak uzavřít, že se v průběhu posledních dvaceti let mnoho nového v oblasti technik používaných v Sociologickém časopise neobjevilo.

Tabulka 3. Vývoj užívanosti složitějších statistických technik v Sociologickém časopise (1995-2014) dle pětiletých období (procento užití technik v kvantitativně orientovaných článcích)

Technika:	**1995-1999**	**2000-2004**	**2005-2009**	**2010-2014**
Korelace	20,6%	13,2%	15,2%	0,0%
regrese lineární	23,5%	23,7%	23,9%	22,7%
regrese nelineární	17,6%	7,9%	34,8%	27,3%
EFA	5,9%	10,5%	13,0%	6,8%
CFA	5,9%	0,0%	0,0%	0,0%
SEM	14,7%	2,6%	8,7%	13,6%
mnohor. škálování	0,0%	2,6%	2,2%	0,0%
Procenta	38,2%	18,4%	6,5%	15,9%
Průměr	2,9%	0,0%	2,2%	0,0%
analýza rozptylu	11,8%	7,9%	8,7%	2,3%
t-test	8,8%	10,5%	0,0%	9,1%
loglineární modely	2,9%	0,0%	2,2%	4,5%
kontin. Tabulky	8,8%	21,1%	10,9%	15,9%
diskr.. analýza	2,9%	2,6%	2,2%	0,0%
logitové modely	5,9%	15,8%	0,0%	0,0%
koresp. Analýza	2,9%	2,6%	0,0%	0,0%
Seskupování	0,0%	5,3%	0,0%	13,6%
analýza přežití	0,0%	2,6%	0,0%	0,0%
hier. Modely	0,0%	0,0%	8,7%	11,4%
latentní třídy	0,0%	0,0%	2,2%	0,0%
Jiná	0,0%	0,0%	0,0%	2,3%

Poznámka:
EFA=explorační faktorová analýza
CFA=konfirmační faktorová analýza
SEM=strukturní modelování

b) 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, resp. jejich mechanické aplikace (druhá dílčí výzkumná otázka) Lze očekávat zejména výskyt

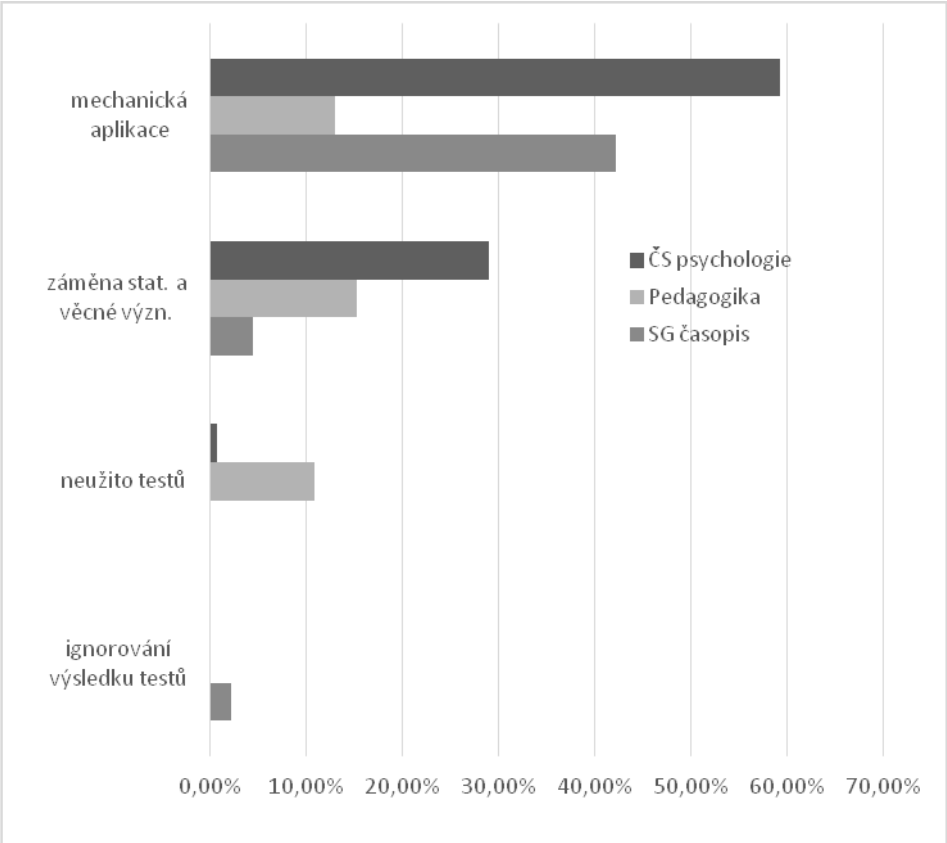
mechanických aplikací (nejčastěji zřejmě hvězdiček ve výstupech statistických procedur)[84] a nesprávnou interpretaci výsledků, typicky záměnu věcné a statistické interpretace, tj. označování statisticky významných výsledků za výsledky „významné“, „signifikantní“, důležité“. Dodejme, že problém těchto 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.[85]

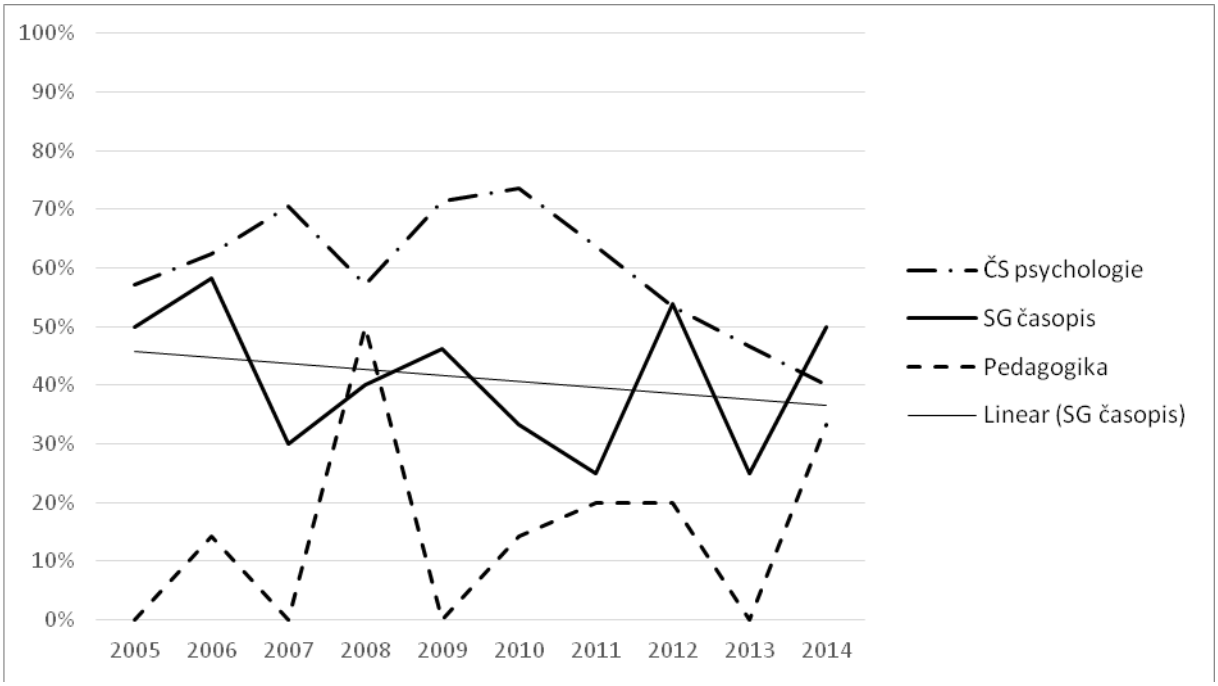
Základní popis výskytu mechanické aplikace statistiky a její případné nevhodné interpretace ve sledovaných časopisech zachycují grafy č. 7-9.

[84] Toto očekávání plyne z běžné praxe, kdy autoři kopírují do článků a beze změny publikují tabulku produkované statistickým software. Tento software pro pohodlí uživatelů většinou hvězdičky do tabulek zobrazuje.
[85] 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 výsledek a 4) výsledek statisticky i věcně významný.

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

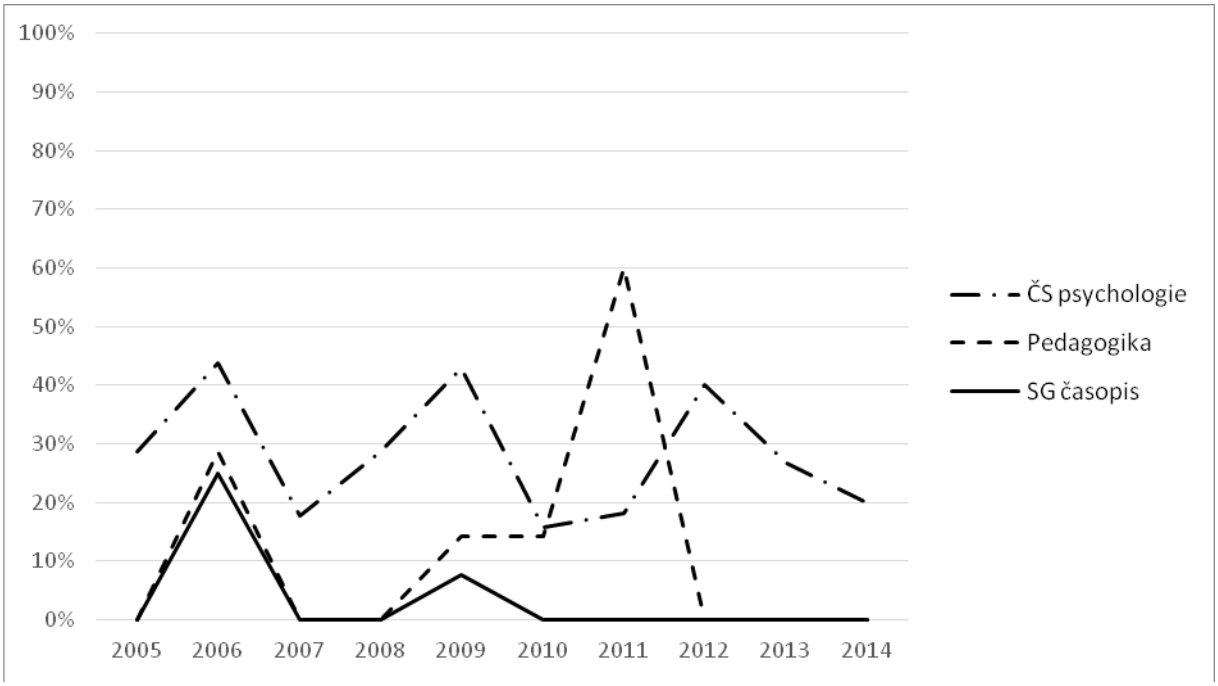


Graf 8. Vývoj podílu mechanického užívání statistických testů v článcích publikovaných v letech 2005-2014 (dle časopisu)



Poznámka: Pro Sociologický časopis byl doplněn lineární trend

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



Zanedbatelným problémem je ignorování výsledků statistických textů, resp. 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 (resp. mechanické označování výsledků) a nesprávná interpretace je naopak poměrně častá. *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).[86] 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 jednoprocentní a tři pro 0,01%) 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 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

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

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*.

c) Detailnější analýza Sociologického časopisu (1995-2014)

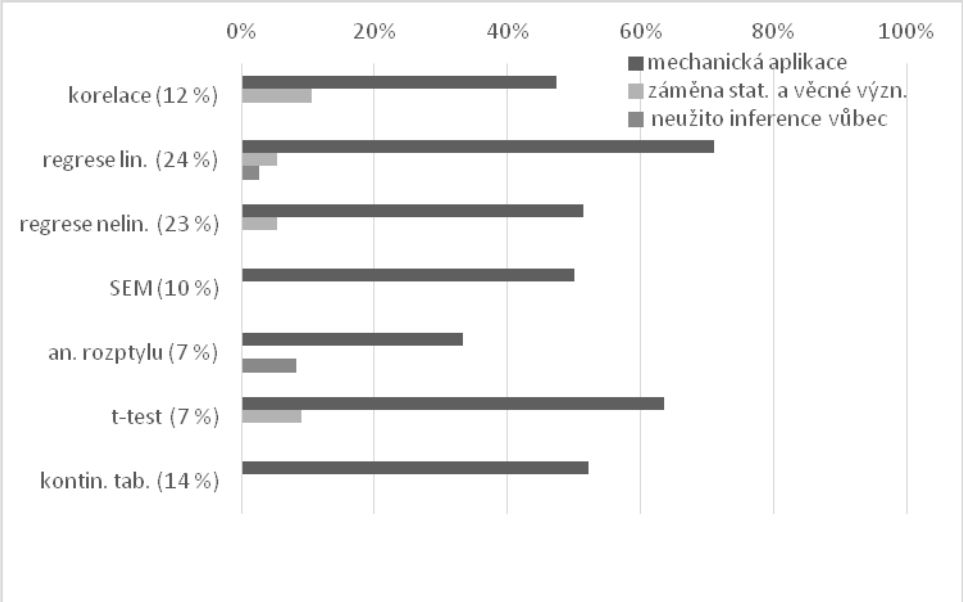
Po obecnějším popisu a porovnání situace tří sledovaných časopisů, opět obraťme pozornost na detailnější analýzu Sociologického časopisu. S ohledem na poměrně příznivé výsledky (plynoucí z předchozího srovnání časopisů) je provedeno pouze jednoduché srovnání mechanických aplikací v čase a stručný rozbor mechanických aplikací u nejčastějších statistických technik. Vývoj mechanického užívání statistiky v čase zobrazuje tabulka č. 4.

Tabulka 4. Podíly článků v Sociologickém časopise zaměřených na kvantitativní analýzu (1995-2014), kde se mechanicky používá statistika (pro pětiletá období)

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

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 č. 8).* Zajímavější vhled nabízí analýza chybné interpretace či mechanické aplikace dle jednotlivých, nejčastěji užívaných statistických technik (srov. graf č. 10).

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



Poznámka: Procento za technikou udává, v kolika procentech kvantitativních článků je technika v daném období v Sociologickém časopise použita

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).

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.

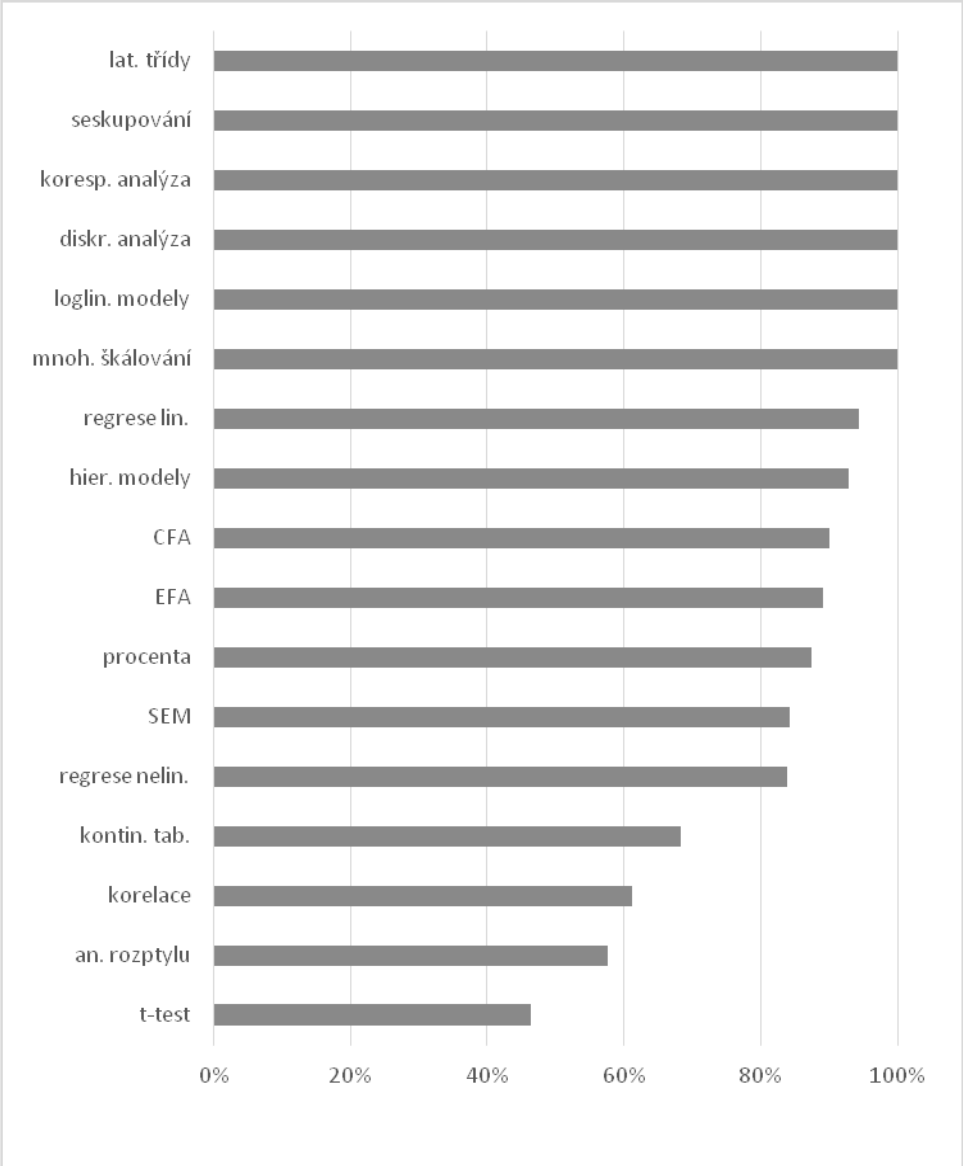
S výjimkou Československé psychologie, kde někteří autoři publikují intervaly spolehlivosti (průměrů a regresních koeficientů) a publikují též výsledky síly použitých testů (méně časté), nebyly výše popsané alternativy zachyceny (resp. z 363 analyzovaných textů obsahoval jediný resampling a jediný bayesovský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

Používání měr věcné významnosti a věcná interpretace výsledků

a) 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 věcné 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 č. 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 č. 11 zobrazuje míru věcné interpretace dle jednotlivých použitých statistických technik napříč třemi sledovanými časopisy.

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)



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.

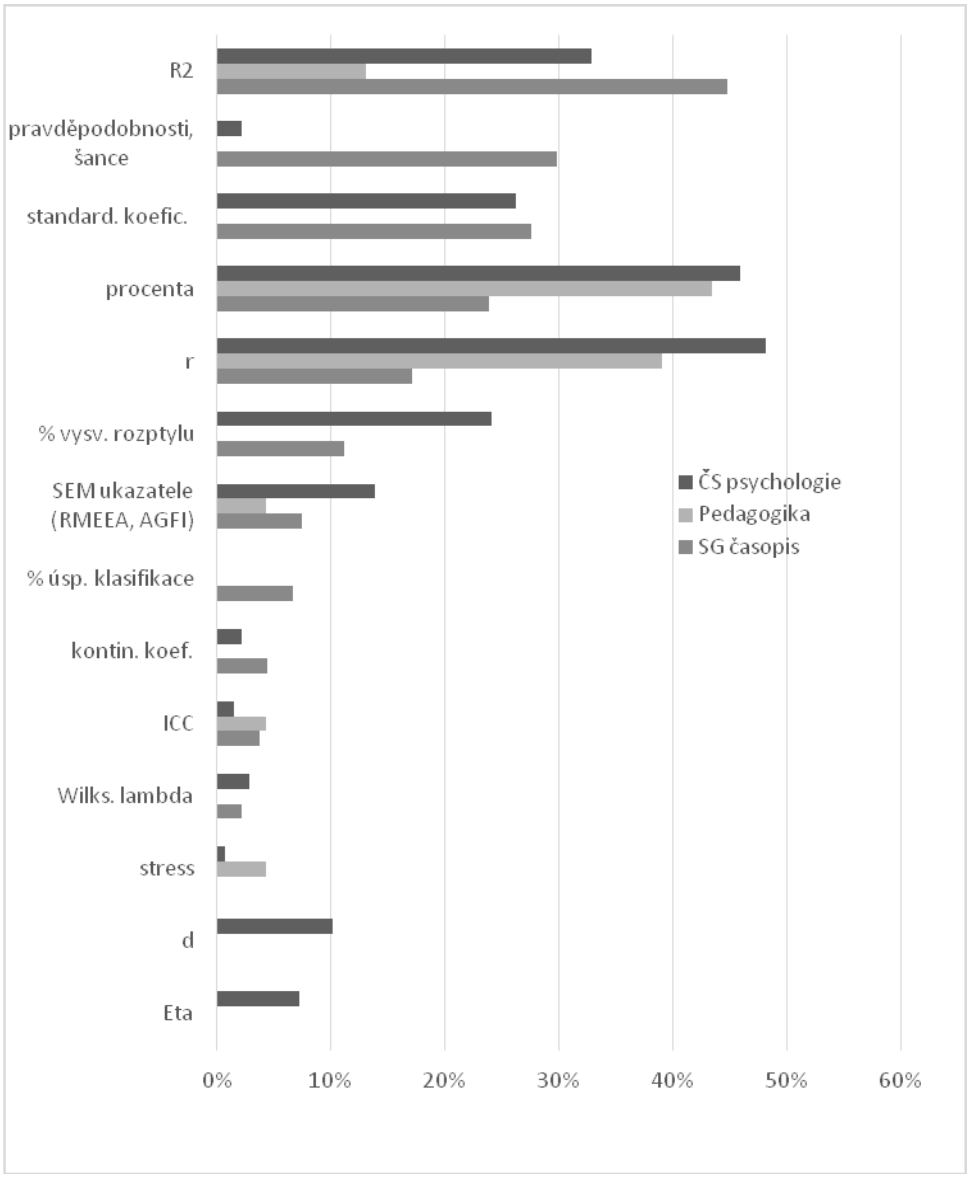
b) Srovnání frekvence užívání měr věcné významnosti a jejich interpretace výsledků v jednotlivých časopisech (2005-2014)

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 č. 12, v následujícím grafu č. 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[87] (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 %, resp. 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, resp. Eta^2) a pro t-testy (Cohenovo d[88]). 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 %, resp. 88 %). Vývoj užívání měr věcné významnosti ve sledovaném desetiletí je ve třech sledovaných časopisech výrazně odlišný (srov. graf č. 13). V případě Československé psychologie dochází ke zlepšení (srov. dále diskusi o institucionálním kontextu vědy), podíl textů, kde míry věcné významnosti chybí 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.

[87] Do této analytické kategorie byl kódován jak ukazatel pro lineární regresní analýzu, tak jeho obdoby pro nelineární regrese.
[88] 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.

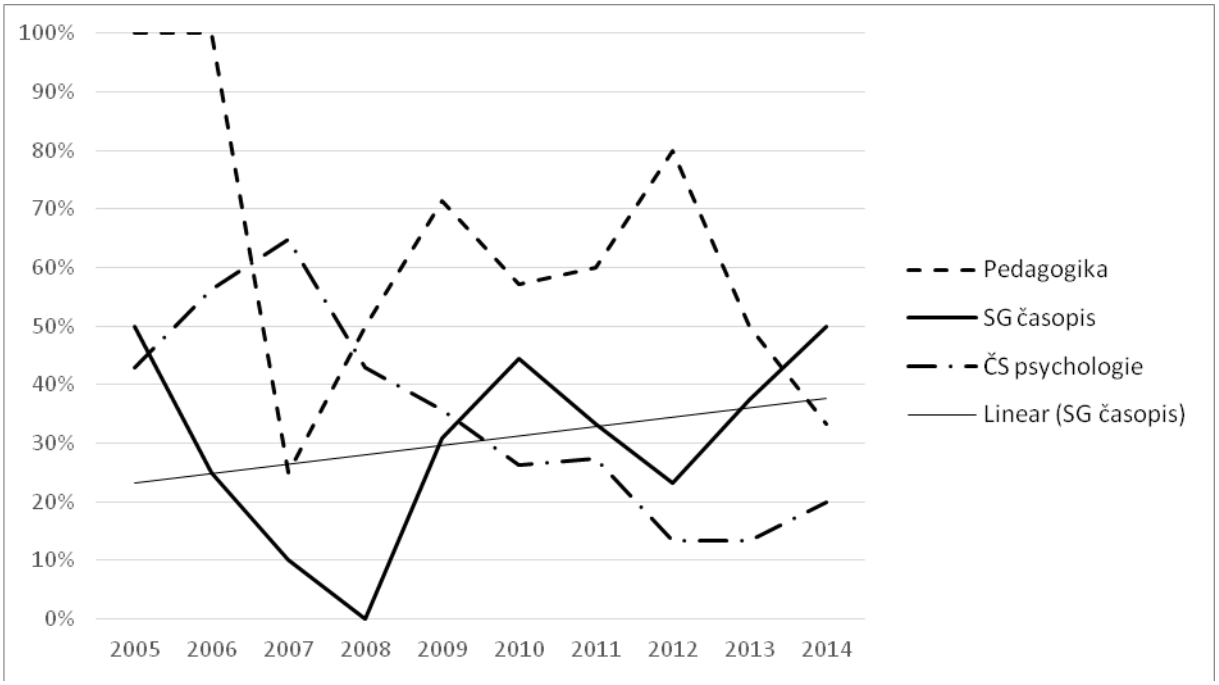
Graf 12. Užívání měr věcné významnosti v letech 2005-2014 (dle časopisu)



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.

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

c) 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 č. 5 jsou uvedeny po pětiletých cyklech podíly článků, ve kterých bylo minimálně jednou užito míry věcné významnosti.

Tabulka 5. Podíly článků v Sociologickém časopise zaměřených na kvantitativní analýzu (1995-2014), kde se je užito minimálně jedenkrát míry věcné významnosti (pro pětiletá období)

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

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 a ti 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 č. 6 shrnuje typicky užívané míry věcné významnosti pro nejčastěji užívané statistické techniky v Sociologickém časopise.

Tabulka 6. 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, v kolika procentech článků 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 dvaceti 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.

Institucionální kontext české vědy a možné dopady na změny kvality kvantitativních analýz

V rámci vlivu institucionálního kontextu české vědy má smysl se zaměřit nad dvě oblasti. První je výrazná změna, která nastala ve financování vědy zejména s ukončením výzkumnných záměrů a zavedením financování dle bodů přidělených vědeckým výsledkům zavedeným do databáze RIV (srov. popis dále). S ohledem na zvýšený tlak na publikování skrze tento systém financování by bylo možné očekávat výrazný nárůst vědecké produkce a snížení její kvality.

Druhým sledovaným fenoménem bude vliv publikačních standardů, které nastavují oborové asociace, resp. jednotlivé časopisy. V oblasti našeho zkoumání (Československá psychologie, Pedagogika a Sociologický časopis) nedošlo ve sledovaném období k žádné aktitvitě příslušných českých (československých) oborových asociací. Psychologové mohli respektovat doporučení APA (5. manuál z roku 2001, resp. 6. manuál z roku 2010), pedagogové pak doporučení AERA (z roku 2006). Kromě toho vyšly v časopise Československá psychologie dva články Urbánka [2007,2008] a posléze tento časopis zveřejnil též metodologické požadavky na výzkumné studie, které na zmíněné dva články obsahově navazují. Bylo by tedy možné předpokládat, že zejména v Československé psychologii bude cca roku 2009 docházet ke zlepšování kvality kvantitativní produkce.

Před prezentací závěrů stručně připomeňme historii hodnocení a financování vědy v ČR. Nejvýraznější změnu přinesl dosud platný zákon č. 130/2002 Sb., o podpoře výzkumu, experimentálního vývoje a inovací z veřejných prostředků a o změně některých souvisejících zákonů (zákon o podpoře výzkumu, experimentálního vývoje a inovací), který v návaznosti na Rejstřík informací o výsledcích (dále jen „RIV“) zakotvil základní principy hodnocení a financování české vědy. Instituce vykazují své vědecké a výzkumné výsledky skrze RIV a dle

těchto výsledků (zpravidla několika let zpětně) získávají finanční prostředky. Metodika se zpravidla mezi jednotlivými roky proměňovala, ale princip byl v zásadě stejný. Každý výsledek vědy a výzkumu byl obodován a dle počtu získaných bodů získala příslušná instituce finanční podporu. Detailnější parametry nastavilo Usnesení vlády ze dne 23. června 2004 č. 644 k hodnocení výzkumu a vývoje a dále dokument nazvaný Reforma systému výzkumu, vývoje a inovací v ČR, schválený usnesením vlády pod č. 287, dne 26. března 2008. Obecně platí, že zhruba od roku 2008 začaly výzkumné a vědní instituce v ČR výrazně sledovat své vědní a výzkumné výsledky a začal mj. sílit tlak na publikování. Dodejme, že dostupná analýza výsledků vědy a výzkumu v ČR zpracovaný Součkem [2012] konstatuje, že nejvíce bodů bylo získáno za články v časopisech s impakt faktorem (cca 60 % v letech 2006-2010), nezanedbatelný byl i podíl bodů za články v recenzovaných časopisech (cca 15 %). Pro úplnost doplňme, že podíl knih či kapitol v knize na celkovém počtu hodnocených výsledků byl lehce pod 10 %. S ohledem na skutečnost, že v oblasti sociálních a příbuzných věd jsou jiné než časopisecké či knižní výstupy velice nepravděpodobným výstupem vědy a výzkumu, lze cum grano salis konstatovat, že v případě sociálních a příbuzných věd, směřoval po zavedení popsaného systému hodnocení největší tlak právě na publikování v časopisech, zejména těch s impakt faktorem, protože zde se zpravidla získávalo více bodů. Dále je nutno poznamenat, že mnohé instituce převzaly systém hodnocení institucí na základě získaných bodů interně a začaly dle těchto kritérií hodnotit jednotlivá pracoviště, případně i jednotlivé pracovníky. Tím logicky tlak na publikování ještě zesílil. Lze si položit otázku, zda tento tlak mohl mít negativní dopady na publikování, které začalo být méně kvalitní. Samozřejmě otázkou je též, zda tyto skutečnosti lze prokázat na provedeném výzkumu tří předních českých časopisů. Cílem tohoto exkurzu bude tedy sledovat vývoj kvantitativně laděné produkce v Československé psychologii, Pedagogice a Sociologickém časopise v čase z hledisek, které byly rozebrány v předchozích analýzách (využívání statistických testů pro nevhodná data, mechanické či jiné nevhodné interpretace získaných výsledků, věcná interpretace výsledků a užívání měr věcné významnosti). Je nutno na počátku otevřeně přiznat, že tato analýza rozhodně nepodá komplexní odpověď na otázku dopadů systému hodnocení a financování vědy, ale může poukázat na některé efekty.

Pokud by mělo platit, že změna financování české vědy započatá v roce 2009 změnila strategie publikování vědců, pak by měl ve třech sledovaných časopisech být přinejmenším patrný nárůst produkce (1) a zhoršení kvality článků (2).

Ad 1. Z grafu č. 2 plyne, že minimálně v oblasti kvatitativně orientovaných článků se od roku 2009 žádný očekávaný boom neodehrál, naopak v případě Československé psychologie je patrný setrvalý pokles počtu článků od roku 2010. Samozřejmě lze nalézt i alternativní vysvětlení, které nahrává tezi o vlivu nového systému financování české vědy. Toto vysvětlení by říkalo, že schopnější autoři začali více publikovat v zahraničí a přestali publikovat v českých časopisech, protože zahraniční publikace byly více bodované (na základě sledování produkce tří českých časopisů je tato teze neověřitelná).

Ad 2. Zajímavější je pohled na proměnu kvality kvantitativní produkce ve třech sledovaných časopisech. Grafy č. 5, 8, 9 a 13 ukazují vývoj nesprávného používání statistických testů, mechanické aplikace statistických testů, záměny statistické a věcné významnosti a používání měr věcné významnosti. V žádné ze zkoumaných oblastí, nelze od roku 2009 (ve srovnání s předchozím pětiletým obdobím) vysledovat systematické zhoršení a nelze tedy konstatovat, že by změna systému financování vědy v ČR měla přímý a uvedenými nástroji měřitelný negativní dopad na kvalitu kvantitativní vědecké produkce publikované ve třech zkoumaných časopisech.

Naopak patrný je pokles mechanické aplikace statistických testů a zejména pak nižší míra absence měr věcné významnosti v Československé psychologii. Samozřejmě je možné se přít, zda je to způsobeno publikovanými manuály APA, nebo články Urbánka [2007, 2008] a doprovodnými metodologickými požadavky na výzkumné studie. Ve srovnání s dalšími dvěmi časopisy (v zásadě beze změn) je ovšem zřejmé, že v časopise Československá psychologie dochází ke zlepšování v dodržování standardů kvantitativní produkce, chtělo by se konstatovat, novému systému financování vědy navzdory.

Vliv autorství na kvalitu kvantitativních analýz

Kromě pokusu zmapovat dopad kontext české vědy na užívání statistické a věcné významnosti v kvantitativní produkci v českých sociálních a příbuzných vědách, je možné na základě provedeného výzkumu časopisů zaměřit pozornost na vliv jedinců (autorů). Pro možnost detailnějšího vhledu se zde omezíme jen na výsledky publikované v Sociologickém časopise. Možnosti výzkumných otázek jsou zde mnohé, nicméně omezíme se jen na dvě, s tím, že první je inspirována výzkumným zjištěním Ziliak a McCloskey [2008] a další pak plyne z výsledku již dříve v této práci prezentovaných analýz. Konkrétní znění výzkumných otázek je následující:

1) Ovlivňuje počet autorů (spoluautorů) kvalitu zpracování kvantitativního výzkumu a kvalitu prezentace a interpretace výsledků (z pohledu využití koncepcí statistické a věcné významnosti)?

2) Liší se kvalita zpracování kvantitativního výzkumu a kvalita prezentace a interpretace výsledků (z pohledu využití koncepcí statistické a věcné významnosti) dle toho, zda autor v oboru začíná (juniorní pracovník), nebo zda jde o již zkušeného autora (seniorní pracovník).

K nabídnutým výzkumným otázkám je vhodné doplnit podrobnější vysvětlení.

Ad 1) Ziliak a McCloskey [2008] na základě výzkumu článků publikovaných v American Economic Review ukázali, že texty, které mají více autorů jsou z hlediska statistického zpracování horší, tj. autoři se dopouštějí více chyb. Jako nejméně problematické tak byly texty, které měly autora jediného. Analogicky tedy bude provedeno srovnání kvality statistického zpracování článků publikovaných v Sociologickém časopise v posledních dvaceti letech dle počtu autorů.

Ad 2) Odlišnost publikační práce juniorních a seniorních pracovníků by mohla pomoci vysvětlit nárůst počtu nesprávných či nevhodných používání statistické významnosti v průběhu sledovaného období (viz popisy výše). Lze očekávat, že juniorní pracovník, který vstřebal poučky kvantitativní metodologie a statistiky na počátku své vědecké kariéry, ještě postrádá potřebný odstup a uplatňuje tak striktně postup dle učebnicových pouček, které ovšem nejsou zpravidla určeny pro přímé vědecké využití. Teprve v průběhu kariéry s nárůstem zkušeností (zejména skrze četbu odborných textů a jejich oponování v recenzních řízeních) dochází k získání potřebného nadhledu a zažití postupů de lege artis. Pro účely analýzy rozdílu mezi juniorními a seniorními pracovníky byla určena mez 35 let věku (mez vychází z hranice, která je běžně užívána v oblasti vědy v ČR). S ohledem na trvání recenzního řízení tak byl výzkumník do 36. roku svého věku v rok publikace článku označen jako juniorní a poté již jako seniorní.

Začněme s první zkoumanou otázkou, tj. zda ovlivňuje počet (spolu)autorů kvalitu zpracování kvantitativního výzkumu a kvalitu prezentace a interpretace výsledků (z pohledu využití koncepcí statistické a věcné významnosti). Pro tyto účely byly zjednodušeny dříve uvedené kategorie a bylo zkoumáno:

a) zda data, která autor užívá umožňují zpracování skrze statistické testy,

b) zda autor využívá v článku minimálně v jedné analýze mechanické aplikace statistických testů

c) zda autor v článku minimálně v jedné analýze zaměňuje slovně či fakticky věcnou a statistickou významnost výsledků,

d) zda autor neopomněl věcně interpretovat minimálně u jedné analýzy věcně získané výsledky a

e) zda autor neopomněl minimálně u jedné analýzy použít některou z měr věcné významnosti.

Ze 162 analyzovaných článků Sociologického časopisu vydaných v letech 1995-2014 měla celá polovina jednoho autora, 37 % dva spoluautory a zbývajících 13 % tři a více autorů (více než tři autory jen 3,5 %, maximální počet byl 6 spoluautorů u jediného textu). S ohledem na toto rozložení jsou použity pro analýzu dat tři kategorie: jeden autor, dva spoluautoři a tři a více spoluautorů. Toto rozložení zajišťuje, že i v nejméně početné kategorii článků se třemi a více spoluautory je k dispozici dostatečný výzkumný soubor (21 článků). Odpověď na otázku č. 1 nabízí souhrnně tabulka č. 7.

Tabulka 7. Problémy při využívání statistické a věcné významnosti dle počtu (spolu)autorů v Sociologickém časopise v letech 1995-2014

Počet (spolu)autorů/fenomén	**1**	**2**	**3 a více**
Použití st. testů pro nevhodná data	19,8 %	23,3 %	33,3 %
Mechanická aplikace st. testů (min 1x v článku)	49,4%	36,7%	23,8%
Záměna věcné a statistické významnosti (min 1x v článku)	8,6%	5,0%	0 %
Chybějící věcná interpretace (min 1x v článku)	17,3%	5,0%	0 %
Chybějící míra věcné významnosti (min 1x v článku)	38,3%	40,0%	19,0%

Připomeňme, že dle výsledků Ziliak a McCloskey [2008] v případě, že publikuje více autorů, je pravděpodobnější chybné využití koncepce statistické či věcné významnosti. Ziliak a McCloskey to odůvodňují tím, že pokud publikuje více autorů společně, pak každý spoléhá na druhého a nepřijímá dostatek vlastní odpovědnosti (jde o ex post facto vysvětlení, které nemá žádnou oporu v jejich empirickém materiálu). Když nahlédneme na výsledky platné pro posledních dvacet let v Sociologickém časopise, lze závěry Ziliak a McCloskey zpochybnit, resp. mnohé výsledky v tabulce č. 7 jim přímo protiřečí. Jediný fenomén, kde platí, že publikuje-li více autorů, pak je větší riziko nesprávného užití statistické významnosti je patrné u používání statistických testů pro data, která toto neumožňují. U textů se třemi a více

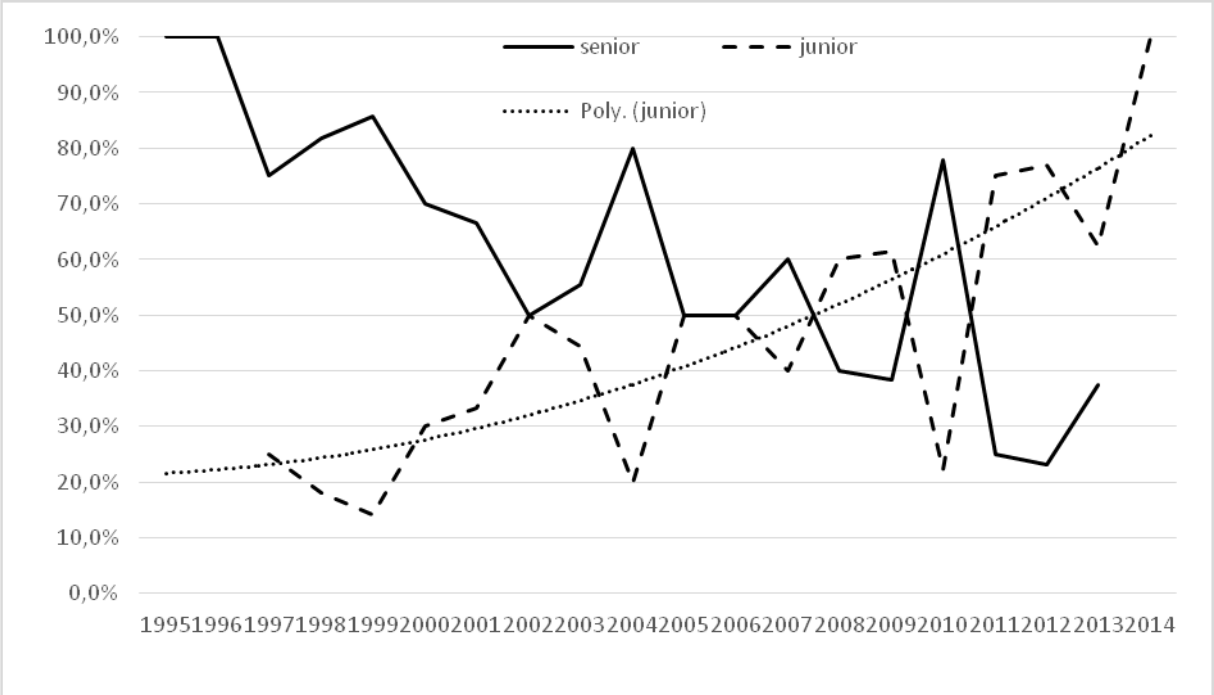
spoluautory je tento fenomén přítomen ve třetině případů, pokud byl autor jediný, pak jen v pětině. Nutno ale poznamenat, že při detailnějším vhledu (není zobrazen v tabulce č. 7) platí, že bez ohledu na počet autorů je podíl textů, kde je vyžito testů pro zpracování dat z nenáhodných výběrů v zásadě vyrovnaný. U více spoluautorů je pouze častější tendence užívat statistické testy pro data, která pocházejí z censu. S ohledem na diskusi uvedenou výše v příslušném exkurzu je tato praxe některými statistiky a metodology doporučována a nelze jí zcela odmítnout. U všech ostatních sledovaných fenoménů platí, že v textech, kde je více autorů (zejména tři a více) se s nimi setkáváme méně často. Nejvýraznější rozdíly jsou u mechanické aplikace statistických testů a u chybějících měr věcné významnosti. Oproti tezi Ziliak a McCloskey o neodpovědnosti spoluautorů, můžeme tedy postavit konkurenční tezi o pozitivním efektu spoluautorství z hlediska užívání statistické a věcné významnosti (lidová moudrost by velela užít přísloví: „Více očí více vidí.“). Samozřejmě definitivní rozhodnutí, která z tezí o vlivu spoluautorství je pravdivá, by rozhodl obsáhlý reprezentativní výzkum vědních publikací napříč vědními obory. Je docela dobře možné, že v některých vědních oborech může být platná teze o pozitivním efektu a jinde o negativním efektu vícečetného autorství.

Analogicky k analýze směřované vlivu počtu autorů bude provedena analýza rozdílů mezi juniorními a seniorními výzkumníky. Budou použity stejné analytické kategorie, není tedy třeba tyto opakovat. Je zajímavé nahlédnout podíl kvantitativní produkce, která byla napsána juniorními, resp. seniorními pracovníky (u textů s více autory bylo využito zařazení dle hlavního autora článku) a zejména vývoj tohoto podílu v čase. Základní přehled nabízí graf č. 14. Pro lepší nahlédnutí trendů byla přidána trendová křivka (kvadratický trend) pro podíl textů juniorních výzkumníků. Když odhlédneme od extrémů (v roce 1995 nebyl žádný text od juniorního výzkumníka a v roce 2014, kdy bylo textů velice málo, nebyl žádný od seniorního výzkumníka) je patrný poměrně jasný trend nárůstu podílu textů od juniorních výzkumníků a logický pokles podílu seniorních výzkumníků. Vysvětlení (bez datové opory) jsou v zásadě dvě. Jednak během sledovaného období dochází v české sociologii ke střídání generací a střední věk téměř chybí (po roce 1990 zažila česká sociologie poměrně výrazný odliv, kdy pracovníci nejrůznějších státních výzkumných institucí buď zakládali soukromé výzkumné agentury, nebo do nich odcházeli pracovat). Druhé možné vysvětlení je internacionalizace české sociologie, díky níž seniorní pracovníci začali publikovat v zahraničích časopisech a ponechali prostor v Sociologickém časopise mladším.

Meritem analýzy ovšem je skutečnost, zda je rozdíl mezi užíváním statistické a věcné významnosti seniorních a juniorních výzkumníků. Odpověď nabízí tabulka č. 8.

Na první pohled nejvýraznější rozdíl nalezneme u používání statistických testů pro nevhodná data, ve třetině textů juniorních výzkumníků byly takto testy použity, u seniorních výzkumníků činil podíl jen 17 %. Za pozornost stojí ještě rozdíl u mechanické aplikace statistických testů, která je také častější v textech juniorních výzkumníků Lze tedy vyslovit tezi o okouzlení statistickými testy u juniorních výzkumníků. Ostatní rozdíly již nejsou věcně tak výrazné (maximálně 5 procentních bodů). Zajímavostí je (zde lze usuzovat na pozitivní vliv nedávného vzdělávání juniorních výzkumníků), že u v textech juniorních výzkumníků méně často absentují míry věcné významnosti pro vypočtené výsledky. Téma rozdílů mezi výzkumníky a vlivu vzdělání (jak během studia, tak poté) bude námětem dalšího výzkumu autora práce. Pozornost bude zaměřena na znalosti výzkumníků z oblasti statistické významnosti a na proces přípravy budoucích výzkumníků na vysokých školách (kurikula, učebnice a vyučující).

Graf 14. Vývoj podílu seniorních a juniorních hlavních autorů kvantitativních článků v Sociologickém časopise v letech 1995-2014



Tabulka 8. Problémy při využívání statistické a věcné významnosti dle věku hlavního autora v Sociologickém časopise v letech 1995-2014

Hlavní autor/fenomén	juniorní	seniorní
Použití st. testů pro nevhodná data	31,9 %	17,2 %
Mechanická aplikace st. testů (min 1x v článku)	46,4%	37,6%
Záměna věcné a statistické významnosti (min 1x v článku)	7,2%	5,4%
Chybějící věcná interpretace (min 1x v článku)	13,0%	8,6%
Chybějící míra věcné významnosti (min 1x v článku)	33,3%	38,7%

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 třetinu 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 [Bernardi a kol. 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 studie zaměřená na European Sociological Review se 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. Bylo by tedy možné konstatovat, 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 výsledky jednoznačně platící pro odborné komunity (lze očekávat, že nároky zahraničních kvalitních časopisů budou vyšší a díky tomu výsledky pro texty publikované v zahraničí by byly příznivější).[89] Nicméně i stávající výsledky umožňují debatu a případné úpravy publikačních standardů jednotlivých časopisů po vzoru zahraničních asociací, zejména APA [2010].[90] Výsledky 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 výsledky statistických testů nadto přebrané bez dalšího ve formě nabízené statistickým software. 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. [91]

Bylo by určitě vhodné provést obdobnou analýzu pro více časopisů působících 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

[89] Autor textu hodlá tuto hypotézu v budoucnu dalším výzkumem v oblasti sociologické produkce prověřit.
[90] Lze doporučit stručné přetlumočení klíčových požadavků z APA manuálu ve vztahu ke statistické analýze, které podal ve svém článku Cumming a kol. [2012].
[91] I zde hodlá autor textu dalším výzkumem navázat.

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, ale spíše podnětem pro její zlepšování.

Literatura citovaná v 5. kapitole

AERA.2006. Standards for Reporting on Empirical Social Science Research in AERA Publications. *Educational Researcher*. 35(6): 33–40

APA. 2010. *Publication Manual of the American Psychological Association*. 6th edition. Washington, DC : American Psychological Association.

Bernardi, 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.

Berk R.A., B. Western, R.E. Weiss. 1995a. Statistical Inference for Apparent Populations. *Sociological Methodology*. 25: 421-458.

Berk R.A., B. Western, R.E. Weiss. 1995b. Reply to Bollen, Firebaugh, and Rubin. *Sociological Methodology*. 25: 481-485.

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.

Bryan, M.L. S. P. Jenkins. 2015. Multilevel Modelling of Country Effects: A Cautionary Tale. *European Sociological Review*, 0(0): 1–20. doi: 10.1093/esr/jcv059

Cuberek, R., Frömel, K. 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. doi: 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.

Firebaugh, G.1995.. Will Bayesian Inference Help? A Skeptical View. *Sociological Methodology*. 25: 469-472.

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*. Califormia: SAGE.

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.

Leahey, E. 2005. Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology. *Social Forces*, 84(1): 1-24.
Levy, P.S., S. Lemeshow 2008. *Sampling of Populations: Methods and Applications* (4th edition). Wiley
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. doi: 10.1198/016214504000000467
Matějů, P. 1989. „Metoda strukturního modelování. Přehled základních problémů.“ *Sociologický časopis.* 25: 399-418.

Mittag, K. C, B. Thompson. 2000. A national survey of AERA members' perceptions of statis tical 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.

Rubin, D.R.1995. Bayes, Neyman, and Calibration. *Sociological Methodology*. 25: 473-479.

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.

Souček M. 2012. Analýza vědy a výzkumu na základě dat z databáze RIV. Ikaros [online]. 16(12). Dostupné z: http://ikaros.cz/node/14018

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. doi: 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. doi - 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. doi:10.1080/00273170903333574

Trafimow, D. 2014. Editorial. *Basic and Applied Social Psychology*. 36(1): 1–2.

Trafimow, D., M.Marks (2015) Editorial. *Basic and Applied Social Psychology*, 37(1): 1-2.

Urbánek, T. 2007. K prezentaci výsledků statistických analýz - 1. část. *Československá psychologie*. 51(6): 601-609.

Urbánek, T. 2008. K prezentaci výsledků statistických analýz - 2. část. *Československá psychologie.* 52(1): 70-79.

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.

Ziliak, S. T., D. M. McKloskey. 2008. *The Cult of Statistical Significance (How the Standard Error Costs Us jobs, Justice, and Lives),* The University of Michigan Press.

Příloha: Kódovací schéma pro obsahovou analýzu článků (s komentářem)

Následující popis je zaměřen na kódování při obsahové analýze článků. Pro část I byl analytickou jednotkou článek, pro části II a III pak jednotlivé analýzy (jednotlivé statistické techniky v článcích použité).

Kromě toho byly zachyceny tyto informace o publikovaných článcích:

1. ***Časopis*** - 1 = Sociologický časopis, 2 = Pedagogika, 3=Československá psychologie
2. ***Rok vydání časopisu*** – číselná proměnná, roszah 1995-2014
3. ***Číslo časopisu*** – číselná proměnná, rozsah 1-6
4. ***Jazyk článku*** – 1= čeština nebo slovenština, 2=angličtina
5. ***Autor/autoři článku*** - textová proměnná, výpis všech autorů článku
6. ***Institucionální příslušnost prvního autora*** – nominální proměnná 1 = VŠ ČR, 2 = Akademie věd, 3= Jiné pracoviště v ČR, 4= mimo ČR
7. ***Rok sběru dat*** - číselná proměnná

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

e) z censu (kód 1),
f) ze záměrného výběru (kvótního, dostupného apod.) (kód 2),
g) z malého výběru (kód 3) a
h) z výběru s extrémně velkým počtem vybraných jednotek či daty spojenými z různých datových souborů (kód 4, resp. 5).

Bylo použito trojice proměnných, aby bylo možné zaznamenat více problematických rysů použitých dat, nicméně vždy postačila jen jediná. Na rozdíl od dále uvedených charakteristik, jde o charakteristiku, která přináleží článku. Problematické bylo, že ne vždy je v článku charakter analyzovaných dat popsán. Na rozdíl od postupu Cubereka a Frömela [2011], kteří v takovém případě považovali data za data ze záměrného výběru, byla informace o datech hledána na webových stránkách příslušných vědeckých projektů. Pokud článek analyzoval dva datové soubory (jde o okrajový jev), pak by byla připsána charakteristika příslušné analýze, nicméně nikdy nenastal případ, že by dva v článku použité soubory byly kódovány odlišně.

II. **V oblasti nesprávného užívání statistické významnosti** bylo sledováno trojicí proměnných, 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) (kód 1),

b) záměně statistické a věcné významnosti (kód 2)

c) slovní popis „významné, signifikantní“ pro statisticky významné výsledky (kód 5)

d) ignorování výsledku testů statistické významnosti, resp. interpretace v rozporu s těmito výsledky (kód 6)

Původní kódovací schéma využívalo ještě kódu 3 (opomíjení statisticky nevýznamnných výsledků v abstraktu či závěru) a kódu 4 (neužito statistické inference vůbec), nicméně pilotáž ukázala, že kód 4 je zcela okrajový a kód 3 je naopak velice rozšířen a bude smysluplné mu věnovat samostatnou studii (nadto nelze zjistit kolik statisticky nevýznamných výsledků bylo autorem odstraněno již před publikací textu, a tak by šlo o zkreslené výsledky). Pro úplnost je nutno dodat, že mechanická aplikace statistické významnosti ve formě zvýrazňování výsledků (hvězdičky, tučné písmo, kurzíva apod.) byla hodnocena jako přítomná bez ohledu na to, zda se nacházela přímo v textu či doprovodných tabulkách. Pro zaznamenání výskytu postačoval jediný výskyt mechanické aplikace u příslušné statistické analýzy. Různé variace tohoto fenoménu jsou popsány v analytických výsledcích.

Pro obtížnost, respektive častou nemožnost rozlišování případů pod písmeny b) a c) zejména v případě Československé psychologie 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. U jednotlivých analýz (statistických technik) je vždy zaznamenáno (ve formě mnohonásobné proměnné) jaké formy nesprávného užívání statistické významnosti se vyskytují.

Kromě nesprávného užívání statistické významnosti bylo též u každé analýzy (statistické techniky zaznamenáno), ***jaká statistická technika je používána***, využit byl následující seznam, resp. kódy:

1-korelace (bez rozlišení toho, jaký korelační koeficient se počítá)

2-regrese lineární

3-regrese nelineární

4-explorační faktorová analýza

5- konfirmační faktorová analýza

6-strukturní modelování

7-mnohorozměrné škálování

8-procenta

9-průměry

10-analýza rozptylu (bez rozlišení zda je jedno či vícefaktorová, nebo vícerozměrná) včetně neparametrických obdob

11- t-test (a případně neparametrické obdoby)

12-loglineární modely (bez logitových modelů, srov. kód 15)

13-kontingenční tabulky

14-diskriminační analýza

15-logitové modely

16-korespondenční analýza

17-shluková (seskupovací) analýza

18-analýza událostí

19-víceúrovňové modely

20-analýza latentních tříd

29-jiná technika

III. **V oblasti užívání měr věcné významnosti a věcné interpretace** bylo (opětovně na úrovni jednotlivých analýz) sledováno následující:

1. ***Použité míry věcné významnosti*** (max. 3):

1 – korelační koeficient (bez rozlišení typu resp. vzorce, s výjimkou ICC, viz kód 13)

2 - R^2 (vč. pseudo R^2 pro nelineární regrese)

3 – standardizované koeficienty

4- Eta, resp. Eta^2

5- Cohenovo d

6 - procenta

7 - pravděpodobnosti, šance

8 - Wiksovo lambda

9 - procento vysvětleného rozptylu

10 - kontingenční koeficienty (např. Cramerovo V)

11 - ukazatele vhodnosti modelů SEM (např. RMSEA, AGFI)

12- míra úspěšnosti klasifikace

13- vnitrotřídní korelační koeficient (ICC)

14- jedinečnost

15 – adjustovaná residua

Dále bylo zaznamenáno ve formě dichotomické proměnné (kódované 0,1) ***zda je provedena interpretace minimálně jedné použité míry věcné významnosti*** (1=ano).

Poslední zaznamenanou charakteristikou bylo, ***zda je provedena věcná interpretace analýzy***. Bylo opět využito jednoduché dichotomické proměnné (kódované 0,1). Kódu 1 (přítomnost věcné interpretace) bylo využito, pokud autor článku u příslušné analýzy kromě zhodnocení statistické významnosti interpretoval (byť částečně, nebo jen část) věcně výsledky, například tedy uvedl, že posuzovaný vliv je pozitivní či negativní, velký nebo malý. Oproti studii Barnardi a kol. [2017] jde tedy o minimalistický přístup, proto nelze výsledky obou studií přímo srovnávat.

6. Závěr

V této části jsou shrnuty základní poznatky obsažené ve čtyřech článcích tvořících disertaci. Nadto s ohledem na lehce netradiční strukturu práce (kapitoly 2-5 tvoří publikované resp. k publikaci připravené články) jsou ještě provedena doplnění (dovysvětlení), aby byla jasnější vazba mezi jednotlivými částmi práce a případně doplněny některé koncepce. Dále jsou zde precizněji a souhrnně formulována doporučení pro publikační a výukovou praxi, která jsou obsažena v dílčích závěrech jednotlivých kapitol (2-5).

Omezení při používání statistických testů

Ve druhé kapitole je pojednáno, kdy je a kdy není správné používání statistických testů. Výsledkem je zjištění, že klasické testy (například t-testy, analýza rozptylu) jsou vhodné pro data získaná náhodným výběrem z velké (ideálně nekonečné populace) a slouží pro zobecnění výsledku z našeho výběru na populaci. Dále jsou pak zkoumány případy, kdy testy buď použít nelze, nebo je to zbytečné, případně lze je užít s určitými úpravami, případně jen při dodržení striktních požadavků. Konkrétně jsou zmíněny tyto situace:

a) data pochází z censu
b) data pochází ze záměrného výběru (zejména kvótního)
c) data pochází z malého náhodného výběru
d) data pochází z náhodného výběru s extrémně velkým počtem vybraných jednotek či daty spojenými z různých datových souborů

Doplňkově jsou pak ještě pojednány základní problémy při práci s vahami, které se často vyskytují v sociálně vědních datech, a popsána je také situace, kdy je vybíráno z malé populace větší procento jednotek (cca více než jedna desetina).

Pokud se zaměříme na čtyři situace (a-d) výše popsané pak platí:

1) Pro data z malých výběrů (ad c) musíme místo běžných parametrických testů nejčastěji používat jejich neparametrické alternativy a tím je problém vyřešen.
2) Pro data z extrémně velkých výběrů (a jiné obdobně velké soubory, cca od tisíce jednotek) není třeba statistické testy používat, protože rozdíly či souvislosti na první pohled viditelné budou statisticky průkazné (výběrová chyba je v těchto případech minimální).

3) Poněkud obtížnější je situace, kdy naše data pochází z nenáhodných výběrů (ad b)), případně analyzujeme data populační (ad a)). V těchto situacích se názory statistiků a metodologů stále ještě rozcházejí. Základem této kontroverze je koncepce superpopulace (tj. jakési teoretické nadpopulace stojící nad našimi běžně zkoumanými populacemi, kde se předpokládá, že získaná populační data jsou náhodně generovaná z této superpopulace). Dále se pak vychází z rozdělení statistických postupů na ty, které jsou založeny na designovém přístupu a ty, které jsou založeny na modelovém přístupu. Detailnější diskusi obsahuje počátek čtvrtého (empirického) článku v disertaci), proto zde jen shrneme základní poznatky. Pokud přijmeme předpoklad existence hypotetické nadpopulace (superpopulation) nad běžně dostupnými populacemi, lze pro populační data používat postupy pro zobecnění, nicméně autoři doporučují používat spíše bayesovské přístupy než klasické statistické testování [Berk, Western, Weiss, 1995]. Jiní autoři ale doporučují i zde setrvat u klasického pojetí a používat pro populační data pouze prostředků popisné statistiky. Právě toto doporučení (jednoduchá a přitom zcela funkční) je využito v celé disertaci včetně samotné analýzy získaných výzkumných dat, která mají celopopulační charakter. V případně nenáhodných výběrů pak s využitím koncepce nadpopulace a modelového přístupu zdůrazňují někteří autoři [např. Little, 2004] že skrze dobře formulovaný model lze užívat statistické testy. Reálným problémem je však zajistit dobře formulovaný model, zřejmě nejlépe se to daří v ekonometrické praxi. S ohledem na skutečnost, že v české sociologii, pedagogice i psychologii se složitější modely, které by zohledňovaly u nenáhodného výběru fenomén stratifikace, shlukování případů a nestejných pravděpodobnostní vybrání, nevyskytují, je případné používání statistických testů pro nenáhodné výběry hodnoceno jako problematické.

Problematické rysy koncepce statistických testů a jejich alternativy

Třetí kapitola poukázala na problematičnost konceptu statistické významnosti a jeho možná zneužívání. Konkrétní popsané problémy jsou tyto:

a) nedostatečná výpověď o základním souboru (skrze P hodnotu měříme pravděpodobnost získání dat při platnosti nulové hypotézy, nikoliv pravděpodobnost této hypotézy samotné),

b) nereálnost nulových hypotéz (většina nulových hypotéz v základním tvaru konstatuje, že neexistují žádné rozdíly či souvislosti, to ovšem ani v praxi nepředpokládáme, a proto dochází k jejich zamítání),

c) mechanická práce s klasickou 5% hladinou statistické významnosti (hvězdičky ve výstupech, stepwise postupy vybírání proměnných, nejlepší modely apod.),

d) statisticky významné neznamená důležité (autoři slovně záměňují mezi statistickou a věcnou významností),

e) nepublikování statisticky nevýznamných výsledků (autoři tak činí často z obavy, že text nebude publikován, díky čemuž dochází ke zkreslování při používání metaanalytických postupů).

S ohledem na možnosti praktikujících sociologů a příbuzných vědců je doporučeno používat místo statistického testu (a konstatování statistické významnosti) spíše intervaly spolehlivosti pro rozdíly, parametry či koeficienty a vyhnout se zmíněným nepřesnostem ve vyjadřování. Při plánování výzkumu je vhodné zvážit velikost výběrového souboru a velikost síly testu, aby nebyly prostředky na sběr dat vynaloženy zbytečně. V případě více si konkurujících modelů je namístě používat s opatrností jako doplněk též informační kritéria.

Věcná významnost a způsoby jejího měření

Čtvrtá kapitola si klade za cíle představit českému publiku koncepci věcné významnosti a měr, které slouží k jejímu zachycení. Zatímco statistická významnost slouží zejména k možnosti zachytit výběrovou chybu a zobecnit výsledek na námi zkoumanou populaci, věcná významnost zkoumá, zda je naměřený rozdíl či zjištěná souvislost je důležitá pro vědecké poznání či praktické účely (zpravidla prizmatem příslušné vědecké disciplíny a předchozích zjištění). Pro zachycení velikosti věcné významnosti a možnosti jejího vyhodnocení byly vyvinuty desítky ukazatelů, které můžeme souhrnně označit jako míry věcné významnosti (termín samotný je návrhem autora disertace, dosud byla terminologie nejednoznačná, nejčastěji se využívalo pojmu efekt účinku, který byl téměř doslovným překladem anglického výrazu effect size). Čtvrtá kapitola seznamuje čtenáře s nejčastějšími mírami, konkrétně Cohenovým d pro srovnání průměrů ve dvou výběrech (doplňkově pojednává i Hedgesovo g a Glassovo delta pro stejnou situaci). Dále je ukázáno Haysovo

omega^2 a Fisherovo Eta^2 (míry vysvětlení sledované charakteristiky třídícím faktorem v analýze rozptylu). Pojednány jsou i běžně známé charakteristiky pro souvislosti proměnných, tj. korelační koeficient, index determinace a jeho upravená verze.

Výsledkem je též popis předností a nedostatků měr věcné významnosti. K přednostem patří zejména nezávislost na velikosti výběrového souboru (tj. stejná využitelnost pro malé i velké výběry), nezávislost na měřítku sledované proměnné (srovnatelnost) a možnost využití v metaanalýze. Naopak nedostatkem je, že míry věcné významnosti nejsou inferenční ale pouze popisné charakteristiky (a pro zobecnění je třeba počítat intervaly spolehlivosti, což nebývá zcela triviální). Dalším nedostatkem některých měr je skutečnost, že jsou založeny na určitých parametrických předpokladech (zejména normalitě) a tyto nejsou často splněny. Mírám je též vytýkáno, že neměří významnost pro jedince, ale průměrnou, proto jsou míry v některých oblastech problematicky použitelné (například lékařství používá spíše tzv. klinickou významnost). Kromě základních měr autor na základě vlastního bádání v oblasti věcné významnosti popisuje i některé speciální míry pro vícerozměrné statistické techniky (v literatuře věnované tématu nezmiňované), konkrétně popisuje míry pro víceúrovňové modelování, mnohorozměrné škálování, diskriminační analýzu, logistickou regresi a korespondenční analýzu.

Výsledky výzkumu časopisů z hlediska používání statistické a věcné významnosti

Výsledky představené v páté kapitole (empirické části) ukazují na skutečnost, že v české sociálněvědní a příbuzné kvantitativní produkci (skrze obsahovou analýzu kvantitativních textů publikovaných v Československé psychologii, Pedagogice a Sociologickém časopise) zdaleka není vše v pořádku. Poměrně často dochází k využívání statistických testů pro data, která toto neumožňují (v Sociologickém časopise se jedná cca o třetinu případů, v Československé psychologii pak o více než tři čtvrtiny). V případě Sociologického časopisu má tento fenomén v posledních dvaceti letech spíše narůstající tendenci. Kromě toho dochází poměrně často k mechanické aplikaci statistiky či nesprávné interpretaci (ve více než 40% článků Sociologického časopisu za posledních 10 let) a absentuje věcná interpretace výsledků. Míry věcné významnosti jsou používány poměrně často, nicméně v jejich užívání jsou také mezery, nadto poměrně málo dochází k jejich interpretaci. Pro korektní interpretaci je nutno dodat, že výsledky obdobných studií zahraničí (čaasopisů jako je American Sociological Review, American Journal of Sociology nebo European Sociological Review) jsou srovnatelné či ještě horší (v případě ASR a AJS jde ovšem o starší studie).

Doplňkově byl zkoumán vliv změny financování české vědy na dodržování standardů v oblasti statistické a věcné významnosti. Lze konstatovat, že na produkci třech sledovaných časopisů nelze prokázat žádné negativní vlivy systému založeného na evidování výsledků v RIV a přepočtu bodů za publikace na finanční prostředky. Naopak skrze pozorované snížení míry užívání mechanických aplikací statistických testů a nárůst užívání měr věcné významnosti v časopise Československá psychologie lze usuzovat, že doporučení (od asociace APA[2010] či v publikovaných článcích Urbánka[2007,2008]) měla pozitivní efekt na dodržování standardů.

V závěru analýzy byl ještě posuzován vliv autorství na dodržování standardů v oblasti statistické a věcné významnosti v Sociologickém časopise. Díky provedené obsahové analýze se neprokázala teze Ziliak a McCloskey [2008], že texty s více autory častěji nedodržují standardy. Naopak výsledky spíše naznačují, že texty více autorů standardy častěji dodržují. Analogicky byla provedena analýza rozdílů mezi juniorními a seniorními výzkumníky. Nebyly nalezeny výraznější rozdíly, nicméně platí, že používání statistických testů pro nevhodná data je častější v textech juniorních výzkumníků, a tito též častěji užívají mechanické aplikace statistických testů.

Pro korektnost závěrů je třeba upozornit na jasná omezení předloženého výzkumu. Výzkum se zaměřuje na tři přední české časopisy za posledních deset let (2005-2014) S ohledem na možný překryv autorů publikujících v časopisech, nelze automaticky výsledky považovat za výsledky platné pro tři odborné komunity působící v ČR (tj. pedagogy, psychology a sociology). Nadto se kromě českých článků vyskytují v Československé psychologii (zřídka) a v Sociologickém časopise (jedna třetina produkce) též texty anglické, jejichž autoři bývají ze zahraničí. V Pedagogice a Československé psychologii nejsou výjimkou texty slovenských autorů. Cum grano salis je možné považovat výsledky výzkumu za výsledky, které vypovídají o československé komunitě. Je samozřejmě nutné upozornit, že zejména v posledním desetiletí dochází k častému publikování mimo ČR (zejména mezi psychology a sociology) a díky tomu nelze výsledky v této studii brát jako výsledky jednoznačně platící pro odborné komunity (lze očekávat, že nároky zahraničních kvalitních časopisů budou vyšší a díky tomu výsledky pro texty publikované v zahraničí by byly příznivější).

Náměty pro další výzkum

Na tomto místě je vhodné též formulovat náměty pro další zkoumání předmětné oblasti. Určitě lze pokračovat v obsahové analýze publikovaných textů, další výzkum by se měl zaměřit na více časopisů působících v České i Slovenské republice v sociálněvědní oblasti (určitě by bylo vhodné provést výzkum slovenské Sociológie) pro srovnání by bylo vhodné zaměřit pozornost i na přírodovědecké, lékařské či technické časopisy, Obdobně bude žádoucí provést i komparativní výzkum zahraničních sociologických časopisů. Nicméně lze zvolit i další výzkumné strategie. Pro pochopení toho, proč se při užívání statistické a věcné významnosti nedodržují doporučené postupy, by bylo vhodné provést testování studentů a výzkumníků, které může pomoci odhalit, v čem spočívají nejběžnější nepochopení těchto koncepcí. Vzorem zde může být test Oakese [1986], který používá šesti otázek, které zjišťují, zda testovaný jedinec rozumí konpcepci statistické významnosti. Dále by bylo vhodné provést hloubkové rozhovory s učiteli statistických a metodologických předmětů a provést obsahovou analýzu příslušných učebnic.

Doporučení pro praxi

Kromě klasického shrnutí poznatků je na tomto místě vhodné shrnout a systematizovat doporučení pro publikační a vzdělávací praxi, která se na mnoha místech roztříštěně nacházejí. Následující doporučení lze považovat za minimum, které je v této oblasti vhodné učinit. Doporučení jsou vedena zvlášť pro publikační praxi a zvlášť pro výukovou praxi.

A) Publikační praxe

   V oblasti publikační praxe lze vyjít z doporučení předních světových asociací, zejména APA [2010], AERA [2006] a ASA [2016]. Vhodné je zohlednit i doporučení spojené s požadavky na tzv. reproducibility research (viz dále uvedené možnosti e,f). Minimální požadavky pro kvantitativní analýzu jsou tedy tyto:

   a) Kromě (namísto výsledků statistických testů) publikovat a interpretovat intervaly spolehlivosti pro vypočtené charakteristiky.
   b) Pro menší datové soubory počítat, publikovat a okomentovat sílu testu.
   c) Zvážit použití alternativních postupů (bayesovská statistika, resamplingové přístupy, srovnávání modelů skrze informační kritéria či bayesovský faktor).
   d) Počítat, publikovat a interpretovat míry věcné významnosti výsledků.
   e) Zveřejňovat data použitá pro analýzu.
   f) Zveřejňovat sadu příkazů, které byly použity pro úpravy dat a jejich analýzu.

B) Vzdělávání budoucích výzkumníků

   a) Výuku musí vést ti, kdo sledují nové trendy v používání technik pro kvantitativní analýzu a nejsou jim cizí doporučení uvedená výše.
   b) Výuka musí být nadto vedena dle učebních textů, které zohledňují novinky a to je nejen v ČR poměrně problém. Ze všech učebnic bude vyhovující snad jen Hendlova [2015].[92]
   c) Ve výuce musí být kladen důraz na praktické postupy (místo teoretických pouček) a musí se ideálně kombinovat statistický rozměr problému s věcným rozměrem. Pokud se toto nedodrží, odchází studenti z hodin

[92] V této oblasti bude třeba provést další výzkum, tj. zjistit, jak reálně probíhá výuka (jednak provést výzkum sylabů, dále i testování znalostí studentů) a jaké učební texty se používají.

s pocitem, že analýza je hotová tím, že se vypočte a vyhodnotí statistický test, opomíjí pak ale věcnou stránku problému.

d) Ve výuce je žádoucí používat reálné datové soubory a komplexnější analýzy (podmínky pro obojí jsou splněny, protože česká i světová data jsou dostupná v nejrůznějších datových archivech a učební texty pro komplexnější analýzy jsou též již poměrně bohatě zastoupeny).

Seznam literatury citované v závěru

AERA.2006. Standards for Reporting on Empirical Social Science Research in AERA Publications. *Educational Researcher*,. 35(6): 33–40

APA. 2010. *Publication Manual of the American Psychological Association*. 6th edition. Washington, DC : American Psychological Association.

Berk R.A., B. Western, R.E. Weiss. 1995. Statistical Inference for Apparent Populations. *Sociological Methodology*. 25: 421-458.

Hendl, J. 2015. *Přehled statistických metod: analýza a metaanalýza dat. Páté, rozšířené vydání.* Praha: Portál.

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. doi: 10.1198/016214504000000467

Oakes, M. 1986. *Statistical inference*. Wiley.

Urbánek, T. 2007. K prezentaci výsledků statistických analýz - 1. část. *Československá psychologie*. 51(6): 601-609.

Urbánek, T. 2008. K prezentaci výsledků statistických analýz - 2. část. *Československá psychologie.* 52(1): 70-79.

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.

Ziliak, S. T., D. M. McKloskey. 2008. *The Cult of Statistical Significance (How the Standard Error Costs Us jobs, Justice, and Lives),* The University of Michigan Press.

Resume

The thesis is focused on the usage of statistical and substantive significance in the Czech social science. The thesis consists of four articles (3 published, 1 unpublished yet), introduction and conclusion including practical recommendations.

The aim of the thesis is theoretically describe the current " standards" in the use of statistical

and substantive significance and through research (content analysis) of articles published in the three leading Czech journals (Czechoslovak Psychology , Pedagogy and Czech Sociological Journal) empirically assess whether the published articles adhere to the " standards" in the field of statistics and the substantive significance of the results . Basic research question was: How are current " standards" followed in the use of statistical and substantive significance in the Czech social science production? In addition to the basic research questions were additional (auxiliary) research questions:

What are the basic limits of statistical significance, i.e. in which cases statistical tests are not appropriate to use?

What are the shortcomings of statistical significance itself, and what are the most common problems in the practical use of it by researchers?

What are the alternatives to the concept of statistical significance?

How is it possible to assess the substantive significance of the results, which effect sizes are recommended?

For all the articles of the three mentioned journals published in the last 10 years[93] (363 articles) quantitative content analysis was conducted. The main focus of the content analysis was then directed to evaluate the correct use of statistical methodology (whether for the analyzed data statistical tests can be used), interpretation of results (if they are substantively interpreted or their statistical significance is only discussed).

Specifically, the attention was focused on three areas:

I. The appropriate use of statistical tests

[93] For detailed insight into Czech sociolology articles published in the last 20 years were analyzed.

II. Incorrect use of statistical significance, especially interpretive errors and mechanical work with statistical procedures

III. Substantive interpretation of the results, using effect sizes and their interpretation.

The first part of the thesis explains the various errors that occur in the use of the concept of statistical significance. It points to the problem of census, nonprobability sampling, sampling of small populations and small samples. Another topic is the use of statistical methods on aggregated data files, especially from international research, and on weighted data. The author points out that in many cases the use of statistical significance is not appropriate, and warns against the incorrect use of traditional statistical methods.

The main goal of the second part of the thesis is to show the limits of using statistical significance as a sole means of making inferences; and to present alternative statistical fit indicators readily available within frequentist approach to statistics: confidence intervals, minimum sample size and power analysis. Multiple working hypotheses are also explored together with two well-known information criteria – AIC and BIC. In the third part it is argued that one of the main problems in data analysis is an over-emphasis on statistical rather than substantive significance. Statistical significance reports the improbability of specific outcomes from sample data using a null hypothesis. In contrast, substantive significance is concerned with the real-world meaning of data modelling results for a population, regardless of p value, where an effect size estimator is used for evaluation. The argument presented in the third part begins with a consideration of how substantive significance may be defined. Thereafter, there is a summary of the literature on substantive significance and its measurement using a variety of effect size estimators, many of which are little known to Czech researchers. This part also examines the topics of economic and clinical significance. In the conclusion of this part there is first proposal to synthesize statistical and substantive significances.

The results presented in the empirical part point to the fact that the Czech social sciences and related quantitative production (through quantitative content analysis of texts published in the Czechoslovak psychology, Pedagogy and Czech sociological journal) do not conform fully to contemporary standards. Authors quite often use statistical tests for data that do not allow this (in the Czech Sociological Journal it is about one-third of cases, in the Czechoslovak psychology more than three quarters). In the case of Sociological Review this phenomenon has increased in the last twenty years. Authors often mechanically apply statistics (asterisks,

stepwise approaches) or misinterpret results of statistical tests (more than 40% of articles Sociological Review in the past 10 years) and some articles lack a substantive interpretation of results. Authors use quite often effect sizes, but there are also gaps in their usage (some of these are not applied e.g. Cohen's d), moreover, quite a few authors interpret these measurements.