Věcná významnost výsledků a její možnosti měření*

Petr Soukup**
Fakulta sociálních věd, Univerzita Karlova v Praze

Substantive significance and it´s measures

Abstract: *In this article 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 this article 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 researchers. This article also examines the topics of economic and clinical significance. In the conclusion, this study discusses attempts to synthesise different concepts of substantive significance and recommends some practical usage of these concepts.*

Key words: *substantive significance, effect size, economic significance, clinical significance*

Data a výzkum - SDA Info 2013, Vol. 7, No. 2: 125-148.
DOI: http://dx.doi.org/10.13060/23362391.2013.127.2.41

* Vznik tohoto článku byl podpořen z projektu GA ČR P404/11/0949 (Klasické výzkumné projekty jako inspirace současného sociologického výzkumu).

** Veškerou korespondenci posílejte na adresu: Petr Soukup, Institut sociologických studií, Fakulta sociálních věd UK, U Kříže 8, 158 00 Praha-Jinonice; e-mail: soukup@fsv.cuni.cz.

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

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 defi-

nici 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 jsou i přes jejich obecné přijímání 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[1] a zda má praktické důsledky (vč. důsledků pro vědu samotnou). K tomu, abychom zjistili, zda je výsledek věcně významný, a pokud ano, pak nakolik, je třeba mít 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 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). Menším 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].

1 Samozřejmě nezpochybňuji okřídlený výrok: „*I nula je ve vědě výsledek*."

Vypočtené absolutní a relativní věcné významnosti jsou velmi jednoduché (každý jim rozumí), nicméně trpí jedním neduhem, kterým 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, poté přistoupit k výpočtu složitějších měr a následně oba ukazatele interpretovat a poukázat na jejich smysl.

Míry věcné významnosti rozdílů a závislostí (effect size measures)[2]

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.[3]

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 (regresi, analýzu 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. V literatuře lze nalézt desítky měr, například Kirk [1996] jich nalezl 40, nicméně pro první seznámení postačí detailněji představit výše uvedených sedm měr a o 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

2 Zde již užívám svůj návrh terminologie pro anglický výraz effect size measures.

3 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 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 statistické významnosti jsou běžně užívané t-testy nebo analýza rozptylu v případě více experimentálních skupin. Dodejme, že za pomoci náhodných experimentů nedochází k 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 [4]

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 používaly k měření stejného fenoménu různých škál. Základní verze vzorce pro Cohenovo d [Cohen 1988] má tento tvar:

$$d = (x_1 - x_2) / \sqrt{s^2}, \qquad (1.1.)$$

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

4 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.

lze užít nejobecněji vzorce založeného na váženém průměru rozptylů v obou skupinách:

$$s^2 = (n_1 * s_1^2 + n_2 * s_2^2) / (n_1 + n_2), \quad (1.2)$$

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:

$$s^2 = (s_1^2 + s_2^2) / 2 \quad (1.3)$$

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:

$$d = t * (n_1 + n_2) / \sqrt{(df * n_1 * n_2)}, \quad (1.4)$$

kde df značí počet stupňů volnosti příslušného t-testu. Provedeme výpočet Cohenova d na příkladu z dat ICCS 2009[5] (viz Příklad 1).

Cohenovo d může být obecně reálné číslo v intervalu od $-\infty$ do $+\infty$, běžně ale nabývá hodnot v řádu jednotek. Pokud vyjde hodnota kladná, znamená to, že sledovaná veličina má větší hodnotu v první, experimentální skupině (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.

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? Je tedy 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.[6]

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 (% = Φ^{-1}(d)). Pro jednoduchost uveďme opět tabulku (3), která nám vše usnadní.

5 ICCS 2009 je mezinárodní studie občanské výchovy, která byla mj. provedena i v České republice. Cílovou skupinou byli 14letí žáci 8. ročníků (tj. žáci ze základních škol a víceletých gymnázií).

6 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.

Příklad 1:

Srovnání znalostí z občanské výchovy 14letých žáků v ČR v testu ICCS 2009 pomocí Cohenova d. Využijeme vzorců 1.1 a 1.2 a po dosazení z tabulky 1 získáme d = 1,32.

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

Národní občanské znalosti – skóre*			
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.

Poznámka: * 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].

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

Interval	Slovní označení
< (0,2–0,5)*	small
< (0,5–0,8)	medium
0,8 a vyšší	large

Zdroj: Cohen [1988: 25].

Poznámka: * 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.

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
%	50 %	54 %	58 %	62 %	66 %	69 %	73 %	76 %	79 %
d	0,9	1	1,1	1,2	1,3	1,4	1,5	2	-
%	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 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.[7] 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, přesným se stává jen u normálně rozdělených veličin.

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 cokoliv usuzovat o 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 neboli interval, ve kterém s určitou pravděpodobností leží hodnota d v celém základním souboru? Nejjednodušší vzorec má tuto podobu:

$$d \pm u_{1-\&/2} * \sqrt{(n_1 + n_2) / (n_1 * n_2) + d^2 / (2 * (n_1 + n_2))}, \qquad (1.5)$$

kde $u_{1-\&/2}$ je 1-α/2procentní kvantil normovaného normálního rozdělení a význam ostatních symbolů je jako ve vzorcích 1.1 a 1.2.

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

7 K měření překryvu obou skupin slouží ještě jiná charakteristika, zájemce odkazujeme na původní text [Cohen 1988: 21–23].

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	$\Delta = (x_1 - x_2) / \sqrt{(s_k^2)}$	1,30

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

v SPSS [Smithson 2001] či jiných paketech.[8] 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

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

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

8 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.

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

Název míry	Vzorec	Výsledek na datech z příkladu 1
Fisherovo eta²	$\eta^2 = SS_b / SS_T$	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,
SS_b 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.

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 eta² je zkresleným odhadem charakteristiky v základním souboru. Nicméně u velkých souborů je toto zkreslení minimální, eta² je tedy pro náš příklad (s mnohatisícovým výběrem) vhodnou mírou.

Problémy zkreslenosti odhadu eta² 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 eta². 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 koeficientu (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).

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

Interval	Slovní označení
< (0,1–0,3)*	small
< (0,3–0,5)	medium
0,5 a vyšší	large

Zdroj: Cohen [1988].

Poznámka: * 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.

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).

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 v t-testu. Pro poslední uvedenou možnost užíváme vzorce:

$$r = \sqrt{(t^2 / (t^2 + df_w))}, \qquad (1.11)$$

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).[9]

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.[10] Aby bylo tomuto problému zabráněno, odvodili mnozí statistici vzorce pro upravené indexy determinace, zřejmě nejznámější je Ezekielův vzorec (opět viz tabulka č. 5).

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

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

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 měl být ukazatel, který měří rozptýlení bodů kolem regresní křivky, 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 i nadále má 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 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 jednodu-

chým 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 popisu 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) závisejí na reliabilitě měřeného ukazatele,
d) neměří významnost pro jedince, ale průměrnou významnost, proto jsou v některých oblastech problematicky použitelné (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 je jejich nekritické přijetí 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).

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 s intervaly 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 parameter). 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 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. Domnívám se, že je jen otázka 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, intraclass 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 Fisherovo 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 [autor 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 a celkového rozptylu). 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 chí-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 neumějí (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 softwarem, 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, kteří navrhují koncept statistické významnosti zcela opustit [Loftus 1993, Schmidt 1996],
2) ti, kteří 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 statisticky nevýznamné výsledky 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 nazvána „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 a 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[11] vyšetření apod.). Často se klinická významnost měří jako procento pacientů, u nichž došlo ke zlepšení, 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 jdou ovšem 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

11 CT - počítačová tomografie, MRI - magnetická rezonance.

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 informace varovala a upozornila na problémy statistické významnosti, většina z nich reaguje radostně v duchu hesla: „Stačí navýšit výběr a 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 významnosti, 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
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 otázka, 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.

PhDr. Petr Soukup *je vyučujícím na katedře sociologie FSV UK. Soustředí se na výuku a aplikace statistických metod, zejména na multivariační analýzu dat, regresní přístupy a analýzu kategoriálních dat. Z věcného hlediska se zaměřuje na problematiku sociologie vzdělání a environmentální sociologii. V poslední době publikoval články o aplikacích statistické významnosti ve speciálních případech a několik kapitol v monografiích zaměřených na problematiku nerovností přechodu na vysokou školu zejména v ČR.*

Literatura

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

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. DOI: 10.1177/0013164012197140.

Anderson, D. R., K. P., Burnham, W. L. Thompson. 2000. „Null hypothesis testing: Pro-

blem, 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 signifikance". *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. DOI: 10.1080/19325037.2005.10608167.

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. DOI: 10.1177/105381510502700307.

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. DOI: 10.1177/0013164401614002.

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* 57 (4): 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. DOI: 10.1080/00220670109598763.

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. DOI: http://dx.doi.org/10.1177/0013164401614001.

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. DOI: 10.1177/001316440161400.

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. DOI: 10.1177/0013164401612001.

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. DOI: 10.1177/00222194030360060101.

Kazdin, A. E. 1999. „The meanings and measurement of clinical signifikance". *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* 6 (5): 746–759. DOI: 10.1177/0013164496056005002.

Kirk, R., E. 2001. „Promoting Good Statistical Practices: Some Suggestions". *Educational and Psychological Measurement* 61 (2): 213–218. DOI: 10.1177/00131640121971185.

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. DOI: 10.1080/00220970209602058.

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. DOI: 10.1111/1467-8721.ep11512376.

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 1970*. 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. DOI: 10.1177/107554708400500404.

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. DOI: 10.1080/00220970209599513.

Robinson, D., H., J., R. Levin. 1997. „Reflections on statistical and substantive significance with a slice of replication.“ Educational Researcher 26 (5): 21–27. DOI: 10.3102/0013189X026005021.

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. DOI: 10.1177/00131640121971392.

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. DOI: 10.3102/0013189X025002026.

Thompson, B.1997. „Editorial policies regarding statistical significance tests: further comments.“ Educational Researcher 26 (5): 29–32. DOI: 10.3102/0013189X026005029.

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. DOI: 10.1080/00223989909599728.

Thompson, B. 2002a. „What future quantitative social science research could look like: Confidence intervals for effect sizes.“ *Educational Researcher*. Vol. 31 (3): 24–31. DOI: 10.3102/0013189X031003025.

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. DOI: 10.1002/j.1556-6678.2002.tb00167.x.

Vacha-Haase, T, B. Thompson. 1998. „Further comments on statistical significance tests." *Measurement & Evaluation in Counseling & Development* 31 (1): 61-63.

Vacha-Haase, T., B. Thompson 2004. „How to estimate and interpret various effect sizes." *Journal of Counseling Psychology* 51, (4): 473–481. DOI: 10.1037/0022-0167.51.4.473.

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

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. DOI: 10.1080/19325037.2006.10598886.