Studia paedagogica
vol. 26, n. 3, 2021
www.studiapaedagogica.cz
https://doi.org/10.5817/SP2021-3-6

MÍRY VĚCNÉ VÝZNAMNOSTI S INTERVALY SPOLEHLIVOSTI A UKÁZKY JEJICH VYUŽITÍ V PEDAGOGICKÉ PRAXI

EFFECT SIZES AND THEIR CONFIDENCE INTERVALS: EXAMPLES OF THEIR USE IN EDUCATION

PETR SOUKUP,
PETR TRAHORSCH,
VLASTIMIL CHYTRÝ

Abstrakt

Předložený článek je teoreticko-metodologického charakteru a pojednává o možnostech využití měr věcné významnosti a jejich intervalů spolehlivosti na příkladu didaktických testů jako jedné z klíčových metod pedagogického výzkumu. Obsahově je rozdělen na tři části: V první je blíže popsána a komparována problematika používání statistické a věcné významnosti; následně se text v obecné rovině věnuje základním mírám věcné významnosti a v třetí části otázce jejich praktického využití včetně intervalů spolehlivosti. V závěru autoři prezentované informace kriticky zhodnocují a syntetizují. Za cíl si kladou přiblížit tuto problematiku čtenářům natolik, aby byli schopni se dané otázce sami věnovat a využívat zmiňované analýzy ve svých výzkumech.

Klíčová slova

míra věcné významnosti, interval spolehlivosti, kvantitativní analýza dat, didaktický test

Abstract

The presented article has a theoretical and methodological character and discusses the possibilities of using measures of effect size, including their confidence intervals, on the example of tests in education. The article is divided into three parts. The first part describes and compares the issues of using statistical significance and measures of effect size. The text then deals with the basic measures of effect size and with the questions of their use. In conclusion, the authors try to critically evaluate and synthesize the presented information. The authors aim to bring this issue closer to the readers so that they are able to address the issue themselves and use the mentioned analyses in their research. The entire presentation is performed on the basis of real data.

Keywords

effect size, confidence interval, quantitative data analysis, didactic test

Úvod

Kvantitativní analýza dat získaných v rámci výzkumného šetření patří k jedněm ze základních metodologických postupů pedagogického výzkumu (Hendl et al., 2014). Jak upozorňuje řada metodologů vědy (Chráska, 2016; Shourki & Edge, 1996; Soukup & Kočvarová, 2016), kvantitativní analýza dat musí respektovat základní pravidla použití různých početních procedur. Při nerespektování těchto pravidel může dojít k zavádějícím výsledkům. V rámci různě zaměřených studií (např. metaanalýza, empirické studie, přehledové studie) bývá upozaděna otázka věcné významnosti namísto významnosti statistické. Právě problematika věcné významnosti je v odborné literatuře velmi často přehlížena (Soukup, 2013).

Míry věcné významnosti mají potenciál obohatit interpretaci dat o informace týkající se využitelnosti dosažených výsledků v praxi, a zvýšit tak relevantnost dosažených výsledků ve vztahu k řešenému tématu (Blahuš, 2000; Cohen, 1988, 1992). Například Hedges et al. (2012) uvádějí, že měření věcné významnosti se často používá jako doplněk k testování hypotéz, dle Wilkinsona (1999) jako způsob kvantitativního shrnutí výsledků studie. Měření věcné významnosti má také důležitou funkci při srovnání výsledků různých studií (viz např. Hedges, 2008; Valentine & Cooper, 2003). Bohužel použití měr věcné významnosti není v pedagogickém výzkumu dostatečně rozšířeno (Soukup, 2016), a to i přesto, že jejich používání a interpretování je již několik let doporučováno světovými pedagogickými i psychologickými organizacemi (AERA, 2006; APA 2010, 2020). Ve výzkumném prostředí zcela absentují intervaly spolehlivosti pro tyto míry, poměrně častá je i praxe, kdy je výsledek míry věcné významnosti sice obsažen v tabulkách, ale není interpretován. Dodejme, že provedením malé výzkumné sondy na kvantitativně orientovaných článcích publikovaných v časopise Studia Paedagogica v letech 2018–2019 jsme zjistili, že i tento přední pedagogický časopis vykazuje popsané rysy: chybí výpočty Cohenova *d* či Cramerova *V* (viz níže), absentují intervaly spolehlivosti pro další míry (korelační koeficient, R^2, Eta^2). Cílem tohoto článku není poukazovat na jednotlivé publikované texty, ale upozornit na obecná východiska měr věcné významnosti a s nimi spojených intervalů spolehlivosti a poukázat na vhodné praktiky při zpracování kvantitativních dat.

Aplikace výpočtů a interpretací měr věcné významnosti do pedagogického výzkumu nepochybně naráží na nedostatečné množství metodologických studií, které by je blíže teoreticky a zároveň i exemplárně představovaly. Dosavadní studie jsou zaměřeny velmi obecně a nepracují s konkrétními daty z výzkumu (Ialongo, 2016; Lakens, 2013; Tomczak & Tomczak, 2014). Dle Soukupa (2013, 2016) časté přehlížení konceptu věcné významnosti

v empirických studiích souvisí i s absencí této problematiky ve studijních plánech českých univerzit; pokud tato problematika nebude implementována do studijních plánů (nejen) pedagogických oborů, nelze očekávat, že situace se v nejbližší době změní. O její absenci svědčí i studie zaměřená na sylaby kvantitativně orientovaných předmětů (Kočvarová & Soukup, 2018). Ostatně tato studie jasně poukazuje i na skutečnost, že převážná většina vyučujících těchto kurzů nejsou statistici, nýbrž pedagogové, a i vlivem toho je zřejmé, že sledování moderních trendů v oblasti kvantitativního zpracování dat může být pro tyto vyučující obtížnější. Posledním důvodem nepříliš častého užívání měr věcné významnosti v oblasti pedagogiky může být též jejich nedostupnost v statistických softwarech. Například Cohenovo *d* (viz níže) jako velmi užívaná míra věcné významnosti byla do SPSS implementována teprve v poslední verzi (27), která je distribuována od roku 2020. Výjimkou je v tomto ohledu volně šiřitelné *R*, pro něž existuje několik balíčků pro tyto výpočty. Je korektní poznamenat, že i grafické platformy využívající *R* (Jamovi a JASP) tyto výpočty již v několika verzích umožňují, trpí ovšem jinými nedostatky, které je zatím činí pro zpracování pedagogických výzkumů spíše neupotřebitelnými (omezená možnost přípravy dat, nemožnost pracovat s vahami apod.).

Předkládaná studie má ambici syntetizovat poznatky týkající se nejpoužívanějších měr věcné významnosti včetně výpočtů jejich intervalů spolehlivosti. Předpokládáme, že tyto poznatky (teoretické i praktické) mohou pomoci studentům i pedagogům při studiu kvantitativních metod pedagogického výzkumu stejně, jako jim může dát určitý návod k aplikaci těchto měr do statistických výpočtů. Klademe si za cíl demonstrovat obecné poznatky o pravidlech užití a interpretaci měr věcné významnosti a jejich intervalů spolehlivosti na vhodných příkladech z výzkumů tak, abychom představili výhody a možné limity výpočtu a interpretace vybraných měr věcné významnosti. Data, se kterými budeme v tomto textu pracovat, pocházejí převážně z didaktických testů jako z často používané metody pedagogického kvantitativně orientovaného výzkumu. Snahou autorů bude demonstrovat aplikaci a interpretaci měr věcné významnosti společně s intervaly spolehlivosti natolik, aby čtenáři byli schopni míry věcné významnosti vhodně používat ve svých výzkumech. Vzhledem k povaze a rozsahu článku není možné na příkladech demonstrovat všechny míry významnosti a jejich intervaly spolehlivosti; zaměříme se tedy na ty nejpoužívanější míry, které je možné při statistickém zpracování výsledků didaktických testů použít. Již na tomto místě musíme s odkazem na autory Ma (2006) i Beretvase a Chunga (2008) uvést, že mezi výzkumníky není absolutní shoda v použití konkrétních metod pro výpočet velikosti efektů v konkrétních případech.

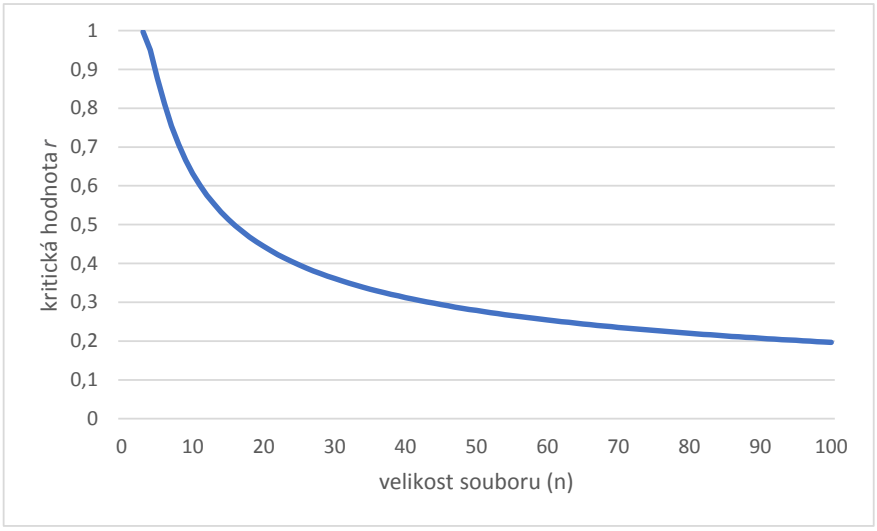
Statistická nebo věcná významnost – jaká je správná volba?

Mnohem častěji než věcná významnost se v kvantitativně orientovaných výzkumech používá **statistická významnost**, kterou Sigmundová a Sigmund (2012, s. 56) definují jako pravděpodobnost, s jakou bychom mohli při opakovaném zjišťování výsledků pomocí stejné metody obdržet data stejně, či ještě více odporující nulové hypotéze za předpokladu, že je nulová hypotéza pravdivá. Primárním účelem statistické významnosti je zobecnění dosažených výsledků. Výzkumník na základě testového kritéria, resp. *p*-hodnoty, může deklarovat, zda rozdíly, souvislosti či závislosti mezi proměnnými jsou statisticky významné a lze je zobecnit na celý základní soubor, nebo statisticky významný vztah mezi proměnnými neexistuje. Pro zamítnutí nebo nevyvrácení nulové hypotézy deklarující neexistenci vztahu mezi proměnnými se v pedagogických výzkumech používá hladina významnosti mající typicky hodnotu $\alpha = 0{,}05$ (Cohen et al., 2007; Sullivan & Feinn, 2012) a s touto předem stanovenou hladinou se srovnává právě vypočtená *p*-hodnota. Otázka 5% hladiny významnosti je však do jisté míry diskutabilní. Často bývá považována za „dogma" a automaticky využívána bez jasného zdůvodnění. Mějme na mysli, že volba hladiny významnosti je závislá na oboru a zkoumané problematice. K její volbě by tak měl výzkumník přikročit až na základě dostatečného teoretického ukotvení tématu. Ostatně nedávno autorský kolektiv 72 světových statistiků doporučil obecně používat 0,5% hladinu významnosti (Benjamin et al., 2018) s ohledem na skutečnost, že běžně užívaná 5% hladina statistické významnosti je příliš benevolentní a nedaří se výsledky mnoha studií replikovat. Mnozí statistici pak doporučují vůbec hranici nepoužívat a hlavně publikovat *p*-hodnoty, nikoli jen vyjádření o statistické významnosti. Ostatně výraz statistická významnost je vnímám spíše v negativní konotaci (srov. např. Wasserstein et al., 2019; McShane et al., 2019).

Na okraj diskuze o *p*-hodnotách, statistické významnosti, resp. statistických testech, dodejme, že existují přinejmenším dvě teoretické koncepce pro tyto postupy: historicky starší Fisherova a mladší založená na lematu Neymana a Pearsona. Jakkoli by detailní popis rozdílů zabral mnoho desítek stran (zejména Fisher svou koncepci celý život dotvářel), stručně uveďme, že Fisher pracuje s jedinou hypotézou, kterou se snaží na základě výpočtu zamítnout či nezamítnout skrze vypočtenou *p*-hodnotu. Pro Neymana s Pearsonem je pak typická dvojice hypotéz (nulová a alternativní), stejně jako dopředu stanovená hladina statistické významnosti (ovšem bez fixace na pevné úrovni typu 5 %) a také koncepce síly testu (resp. jejího doplňku označovaného jako β). V tomto textu se držíme výrazněji Neyman-Pearsonovy koncepce, pracujeme s dvojicí konkurenčních hypotéz a s předem stanovenou hladinou statistické významnosti.

Značnou nevýhodou *p*-hodnoty (míry statistické významnosti výsledku) je její silná závislost na velikosti výzkumného souboru. V případě, že výzkumný soubor je dostatečně velký, není problém získat nízkou *p*-hodnotu a „dokázat" tak statistickou významnost téměř u jakýchkoli proměnných. Exemplárním případem je závislost kritických hodnot Pearsonova korelačního koeficientu na velikosti souboru (graf 1). Je zřejmé, že v případě vyššího rozsahu výzkumného souboru stačí k prokázání statisticky významné závislosti proměnných relativně nízká hodnota koeficientu *r*. Praktickým důsledkem tohoto limitu ve výzkumu je přeceňování důležitosti statistické významnosti ve výsledcích výzkumu. Autoři nekriticky zobecňují dosažené výsledky na celý základní soubor. V některých případech se může stát, že mezi dvěma porovnávanými skupinami je jedno- až dvouprocentní rozdíl, avšak výzkumník dochází k závěru, že rozdíl je signifikantní (tj. statisticky významný), což je přesně dáno právě velikostí výběrového souboru.

Graf 1
Souvislost statisticky významných hodnot Pearsonova korelačního koeficientu a velikosti souboru na hladině významnosti $\alpha = 0{,}05$



Věcná významnost přistupuje k analýze dat odlišně. Míry věcné významnosti udávají, zda je výsledek prakticky užitečný, resp. důležitý v reálném světě. Výsledek výpočtu nám říká, zda má smysl o výsledku hovořit jako o důležitém a relevantním v praxi, tj. zda jsou rozdíly či souvislosti dostatečně velké, aby se jimi mělo smysl zabývat. Na rozdíl od *p*-hodnoty, u měr věcné významnosti velikost výzkumného souboru roli nehraje.

V angličtině jsou míry věcné významnosti často označované jako *effect size*, česky se proto užívá pojem velikost účinku. Jelikož chceme do jisté míry poukázat na skutečnost, že kromě statistické významnosti (prezentované typicky skrze *p*-hodnotu) je vhodné vyhodnotit též věcnou významnost (skrze dále představené míry a jejich intervaly spolehlivosti), budeme v textu preferovat pojem míry věcné významnosti. Tyto míry jsou v zásadě popisnou statistikou výběrového souboru. V případě, že je ambicí autorů této studie výsledky zobecňovat, je nutné tomu přizpůsobit celý design výzkumu včetně vhodného výběru respondentů. Abychom mohli provést zobecnění velikosti míry věcné významnosti na populaci, je třeba buď provést její statistické testování (to se běžně provádí například u korelací či koeficientů determinace v regresní analýze, srov. obrázek 1 na konci článku), nebo počítat intervaly spolehlivosti pro míry věcné významnosti. U měr věcné významnosti se prosazuje spíše druhý přístup. Důvody pro tuto skutečnost jsou poměrně jednoduché. Zastánci měr věcné významnosti poukazují často na problematičnost statistického testování (*p*-hodnot) a mj. propagují užívání intervalů spolehlivosti jako alternativy ke statistickým testům. Proto tento postup doporučují i pro samotné míry věcné významnosti (srov. např. Steiger & Fouladi, 1997; Vacha-Haase & Thompson, 2004).

Interpretace hodnot měr věcné významnosti je založena na dvou přístupech: (1) na doporučených tabulkových hodnotách nebo (2) na tabulkových hodnotách s možným přepočtem na procenta.

(1) Interpretace některých měr věcné významnosti je založena pouze na doporučených intervalech hodnot, podle kterých lze rozhodnout, zda je výsledek nedůležitý, středně důležitý, velmi důležitý apod. Tento přístup kritizuje Soukup (2013) za jeho formalismus; tabulkové hodnoty by měly být spíše pouhým vodítkem v interpretaci výsledků, nikoliv dogmatem (je nutné k hodnotám přistupovat kriticky). Více o tomto problému pojednáme níže.

(2) Jiné míry samozřejmě též využívají intervaly definované v doporučujících tabulkách (k tomu viz Cohen, 1988, a nověji Kline, 2013), avšak tyto hodnoty lze přepočítat na procenta, která ukazují míru vysvětlení dané proměnné. Například na základě koeficientu determinace (R^2) lze určit, z kolika procent je rozptyl jedné proměnné vysvětlen druhou proměnnou. Tento přístup však závisí na povaze dat, konkrétně je možné ho aplikovat pouze na data mající povahu intervalové stupnice. Například pro data mající nominální povahu nebo pro data s nespojitou stupnicí, jejichž výsledek je hodnota z intervalu <0; 1>, nelze interpretaci založit na procentu vysvětlujícím jednu proměnnou tou druhou.

Pro použití měr věcné významnosti ve výzkumech je výhodou i srovnatelnost jejich výsledků napříč různými výzkumy v tzv. metaanalýze. Jelikož výpoč-

ty statistických testů jsou založeny typicky na testovém kritériu (užívaném pro výpočet statistické významnosti, resp. *p*-hodnoty), mají různá statistická rozdělení (např. rozdělení *U*, *F*, χ^2, *t* apod.), míry věcné významnosti tyto různé formáty integrují a dávají výzkumníkovi lepší podklad pro srovnání různých výsledků (Ferjenčík, 2010; Hedges & Olkin, 1985). Některé online kalkulátory také nabízejí přepočet z hodnot testového kritéria (nikoliv z hodnoty *p*) na danou míru věcné významnosti (např. Lenhard & Lenhard, 2016).

Obecně však lze nahlížet na problematiku statistické a věcné významnosti jako na dvě strany jedné mince. Tyto dvě možnosti zpracování kvantitativních dat nestojí proti sobě, nýbrž je možné je označit za komplementární způsoby zpracování dat. I proto je v posledních letech doporučováno prezentované výsledky výzkumu doplnit výpočtem statistické i věcné významnosti (APA, 2020; Fan, 2001). Ostatně i v tomto článku propagované užívání měr věcné významnosti s jejich intervaly spolehlivosti je plně v souladu s těmito doporučeními.

Stručný exkurz o pokusech sloučit statistickou a věcnou významnost

Stejně jako se statistici a kvantitativní metodologové již více než 100 let snaží vyvíjet různé statistické testy a k nim přináležející míry věcné významnosti (viz příklady dále), tak roste i snaha o integraci statistické a věcné významnosti do jednoho nástroje i snaha o vylepšení nástrojů statistického testování, resp. snaha o náhradu *p*-hodnoty jiným ukazatelem, který by v sobě nesl i informaci o věcné významnosti výsledku. Připomeňme, že *p*-hodnota je ovlivněna mj. velikostí výzkumného souboru (srov. graf 1 a diskuzi k němu uvedenou), tj. při větší velikosti souboru a při stejné věcné významnosti hodnota *p* klesá. Pokusů, jak vylepšit *p*-hodnotu, byla v dějinách statistiky celá řada a lze odhadovat, že jim stále není konec. Jako jeden z prvních, na které navazují nyní používané a doporučované přístupy, byl tzv. tříhodnotový přístup, který v 60. letech výrazně prosazoval Kaiser (1960); nicméně tento přístup se nijak výrazně neujal. Ani opakované snahy pokračovatelů Kaisera (Harris, 1997; Hurlbert & Lombardi, 2009) nevedly k rozšíření tohoto postupu. Stručně uveďme, že Kaiser a jeho pokračovatelé důrazně odmítali logiku klasického testování zejména skrze jeden oboustranný test, místo nich navrhovali použít sadu testů s tím, že nebyla formulována jedna alternativní hypotéza, ale tyto hypotézy byly dvě (odtud název tříhodnotový přístup, protože kromě jedné hodnoty specifikované nulovou hypotézou postup nabízel ještě dvě hodnoty, resp. jejich intervaly specifikované dvěma alternativními hypotézami). Stručnou ukázku pro dvouvýběrový test je možné převzít z textu Harrise (1997, s. 152):

Hypotézy pro rovnost či nerovnost středních hodnot ve dvou skupinách jsou formulovány takto:

$$H_0: \mu_1 = \mu_2$$
$$H_1: \mu_1 < \mu_2$$
$$H_2: \mu_1 > \mu_2$$

Přičemž μ_1, μ_2 jsou střední hodnoty pro 1. a 2. skupinu, které srovnáváme.

Následně se prvním testem zjišťuje, zda je v první skupině větší střední hodnota než v druhé, a pokud tomu tak je, přijímá se první alternativní hypotéza H_1. Pokud tomu tak není, v druhém kroku se dalším testem zjišťuje, zda je ve druhé skupině větší střední hodnota než v první, a pokud tomu tak je, přijímá se druhá alternativní hypotéza H_2. A pokud ani jedna z předchozích situací nenastane, nezamítá se nulová hypotéza (H_0).

I když se výše uvedený postup neujal (důvodem bylo zřejmě to, že již v 60. letech se v oblasti sociálních věd užíval cca 20 let klasický *t*-test a nebyla ochota ke změně), došlo cca od 80. let k vývoji některých návazných postupů. Z těch v současnosti nejslibnějších se sluší zmínit postup ekvivalenčního testování (TOST) a dále druhou generaci *p*-hodnot (SGPV). Hlavním propagátorem ekvivalenčního testování je Daniel Lakens, který kromě autorství několika článků (např. Lakens, 2017) vytvořil též balíček pro *R*, který umí tuto proceduru vypočítat. Tento balíček byl nadto využit i pro dva freewarové produkty, konkrétně JASP a Jamovi (oba jsou uživatelsky velmi podobné SPSS, mají přehledné nabídky), a díky tomu lze očekávat rozšíření tohoto postupu. Detailní informaci o postupu TOST a příkladech využití najde český čtenář v nedávno vydaném článku Fica (2020).

Druhým postupem je tzv. druhá generace *p*-hodnoty, anglickou zkratkou označovaná jako SGPV. Hlavním autorem tohoto pokusu o sloučení statistické a věcné významnosti je Jeffrey Blume, nicméně na vývoji se podílel celý tým biostatistiků z Vanderbiltovy univerzity v Nashville. Základním textem je detailní článek vydaný v PlosOne (Blume et al., 2018), dalším stručnějším textem pak text publikovaný v American Statistician (Blume et al., 2020). Kromě textů vytvořil tým vedený Blumem i názornou a graficky zdařilou online aplikaci k výpočtu SGPV, kterou zájemce najde na adrese www.lucy.shinyapps.io/sgpvalue/, případně lze využít i rozcestník všech informací k *p*-hodnotě na www.statisticalevidence.com/second-generation-p-values.

Zkusme stručně představit logiku SGPV. Základem výpočtu je nulová hypotéza, která není formulována bodově (pro jednovýběrový *t*-test např. H_0: $\mu = 1\,000$), tedy nestanoví určitou konkrétní hodnotu, ale je formulována jako interval (např. H_0: $800 < \mu < 1\,200$). Právě toto stanovení intervalu v sobě obsahuje logiku věcné významnosti. Typicky rozpětí nulové hypotézy zahrnuje hodnoty, které jsou věcně nezajímavé, pro alternativní hypotézu pak „zbývají" hodnoty věcně významné (pro naši ukázku by tak alternativní hy-

potéza sdělovala, že střední hodnota je maximálně 800, nebo více než 1 200). Na základě takto modifikované nulové hypotézy se SGPV počítá jako podíl hodnot z oblasti definované nulovou hypotézou, které jsou daty podporovány (matematicky tedy nejde o pravděpodobnost). Pokud je SGPV nulová či blízká nule, podporují data alternativní hypotézu, naopak pokud jsou *p*-hodnoty vysoké (blízké 1), podporují data nulovou hypotézu. Pokud by *p*-hodnota byla okolo 0,5, tak data nestraní ani nulové, ani alternativní hypotéze. Na tomto místě dodejme, že tato logika je velmi podobná logice bayesovské analýze dat, zejména pak ukazateli, kterému říkáme bayesův faktor (BF).

Míry věcné významnosti a pravidla jejich užití

Po krátkém exkurzu o pokusech o sloučení měr věcné a statistické významnosti se vrátíme opět k mírám věcné významnosti. V dnešní době jde již o téměř standardní pomůcku, u které je zejména v oblasti psychologie vyžadováno uvádění a interpretace těchto měr. S ohledem na jejich stále poměrně slabé využívání v Česku a na Slovensku v této části článku ukážeme užití jednotlivých měr věcné významnosti, které doplníme příklady jejich možné interpretace. Pro přehled jsou všechny uvedené míry věcné významnosti uvedeny i v příloze článku, jež je doplněna i o vzorec výpočtu, oporu k interpretaci výsledku a pravidla užití dané míry věcné významnosti.

Soukup (2013) na základě zahraničních klasifikací dělí míry věcné významnosti do dvou skupin: míry měřící rozdíl a míry měřící rozptyl; poněkud specifickou kategorií je korelační koeficient a koeficient determinace, který však řadí do skupiny měr měřící rozptyl. Sigmundová a Sigmund (2012) dodávají, že existují i míry věcné významnosti určené pro testové kritérium chí-kvadrát (například testy pro čtyřpolní a kontingenční tabulku). V další části textu se budeme odděleně věnovat všem zmíněným oblastem.

Míry měřící rozdíl

V praxi se velmi často srovnává průměr ve dvou skupinách, pro tuto situaci v rámci testování statistické významnosti užíváme typicky dvouvýběrový *t*-test. Proto není překvapivé, že pravděpodobně nejpoužívanější mírou věcné významnosti je Cohenovo *d,* které měří právě rozdíl mezi dvěma průměry. Formálně je Cohenovo *d* založeno na rozdílu průměrů ve dvou skupinách, který se ještě dělí směrodatnou odchylkou proměnné. Výsledkem je bezrozměrná veličina, která zpravidla nabývá hodnot v řádu jednotek a po výpočtu ji lze se srovnat s doporučenými tabulkovými hodnotami významnosti výsledku například dle Cohena (1988). Jelikož je tato míra určena pro dvě nezávislé skupiny, lze ji využít v souvislosti s *t*-testem pro nezávislé skupiny, užívá se též skrze přepočet i pro jeho neparametrickou obdobu

(Mann-Whitneyův test, viz dále). Pozor však na skutečnost, že *t*-testy pracují primárně s průměry, naopak Mann-Whitneyův test s pořadími. Podrobně se této otázce věnují Rice a Harris (2005). Doporučené meze k posouzení hodnoty Cohenova *d* jsou pak: **(1)** 0,2–0,5 = malý efekt; **(2)** 0,5–0,8 = střední efekt; **(3)** 0,8 a více = velký efekt (Cohen, 1988).

K výpočtu Cohenova *d* lze využít obecného vzorce podle Cohena (1988), viz vzorec 1.1:

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2}}, \tag{1.1}$$

kde $\bar{x}_1$ a $\bar{x}_2$ jsou průměry první a druhé skupiny a s^2 je rozptyl společný oběma skupinám, který se vypočte dle vzorce:

$$s^2 = \frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2},$$

kde n_1 a n_2 je velikost prvního a druhého souboru; s_1^2 a s_2^2 jsou rozptyly v první a druhé skupině.

Pro úplnost lze ještě doplnit, že lze využívat další míry věcné významnosti měřící rozdíl u dvou skupin. Jedná se o Hedgesovo *g* a Glassovo Δ. Obě míry lze znovu použít pro dva nezávislé soubory, přičemž rozdíly mezi nimi při vyšším počtu respondentů jsou minimální (Lakens, 2013). Obě tyto proměnné jsou počitatelné dle následujících vzorců (1.2 a 1.3):

$$H_g = \frac{\bar{x}_1 - \bar{x}_2}{SD_{pooled}}, \text{ při korekci pro } N < 50 \; H_g = \frac{\bar{x}_1 - \bar{x}_2}{SD_{pooled}} \cdot \left(\frac{N-3}{N-2{,}25}\right) \cdot \sqrt{\frac{N-2}{N}} \tag{1.2}$$

$$\Delta = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{S_k^2}}, \tag{1.3}$$

kde v rámci těchto vzorců SD_{pooled} představuje průměr vnitroskupinového součtu čtverců, S_k^2 je rozptyl kontrolní skupiny, $\bar{x}_1$ a $\bar{x}_2$ jsou pak průměry, N je součet počtu jednotek v prvním a druhém porovnávaném souboru.

Jak uvádí Soukup (2013), hodnoty věcné významnosti dané výpočtem z těchto vzorců jsou velmi podobné Cohenovu *d*, přičemž interpretace výsledku je také stejná (viz výše). Je však nutné znát rozdíly v pravidlech používání těchto tří měr. Ačkoliv Soukup (2013) zmiňuje srovnatelnost těchto měr, žádné rozdíly v pravidlech jejich užití neuvádí (popř. počítá všechny míry pro stejná data). Ialongo (2016) uvádí, že Cohenovo *d* je vhodné použít jen v případě, že velikosti obou testovaných skupin jsou stejně nebo podobně velké a rozptyl obou výběrů se významně neliší. Naopak Hedgesovo *g* je vhodné použít v případech, kdy byl *t*-test aplikován na velmi malých souborech (viz též Lakens, 2013), nebo výzkumné soubory nebyly početně stejné.

Na tomto místě je nutné dodat, že Cohenovo *d* je taktéž možné převést na chybou nezatížené standardizované Hedgesovo *g* a jeho varianci (Borenstein et al., 2009). A konečně Glassovo Δ je vhodné použít v případech s velkým rozdílem směrodatných odchylek nebo pro analýzu výsledků experimentů s přítomností kontrolní skupiny.

Míry měřící rozptyl

Míry věcné významnosti měřící rozptyl jsou určeny pro případy, v nichž máme více než dvě srovnatelné skupiny.

Nejjednodušším ukazatelem věcné významnosti u více než dvou skupin je ukazatel čtverce eta (η^2), též označován jako Fisherova eta. Tento ukazatel je definován jako podíl meziskupinového součtu čtverců na celkovém součtu čtverců (srov. vzorec 2.1). Výsledek se pohybuje v intervalu <0; 1>, přičemž po vynásobení hodnoty stem tento ukazatel interpretujeme jako podíl (procento) vysvětleného rozptylu za pomoci rozdělení do skupin. Nevýhodou tohoto ukazatele je, že se jedná o zkreslený odhad charakteristiky v základním souboru; při vysokém počtu respondentů ve výzkumném souboru je však toto zkreslení minimální. Tuto nevýhodu zkreslenosti řeší Haysova omega (ω^2). Na rozdíl od ukazatele η^2 zohledňuje tato míra vnitroskupinový průměrný součet čtverců a počet srovnávaných skupin. Její výhodou je oproti předchozímu koeficientu ten, že není zkreslenou mírou. Při velkém výzkumném souboru je však hodnota těchto dvou ukazatelů podobná (Lenhard & Lenhard, 2016). Pro doplnění uvádíme vzorce pro obě tyto míry (2.1):

$$\eta^2 = \frac{SS_b}{SS_T}; \omega = \frac{SS_b - (v-1) \cdot MS_w}{SS_t + MS_w}, \tag{2.1}$$

kde SS_T je celkový součet čtverců, SS_b je meziskupinový součet čtverců, v je počet srovnávaných skupin, MS_w je průměrný vnitroskupinový součet čtverců.

Jedněmi z nejpoužívanějších měr věcné významnosti je koeficient korelace (r) a koeficient determinace (r^2). Korelační koeficient nám vysvětluje závislost (souvislost) dvou proměnných; jinými slovy říká, zda vyšší hodnoty jedné proměnné budou zároveň indikovat vyšší hodnoty druhé proměnné. V případě, že pracujeme s kardinálními proměnnými a předpokládáme lineární souvislost, lze využít vzorec 2.2:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{S_x \cdot S_y} \text{ (případně } \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x S_y}\text{)}, \tag{2.2}$$

kde $\bar{x}$ a $\bar{y}$ jsou výběrové průměry, s_x a s_y jsou výběrové směrodatné odchylky. Někdy se také využívá vzorec, který je uvedený v závorce.

V případě, že jsou data ordinální povahy, používá se k výpočtu Spearmanova koeficientu pořadové korelace (2.3):

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n\,(n^2-1)}, \text{ kde } d_i = x_{ri} - y_{ri}, \text{ kde} \quad (2.3)$$

kde n je rozsah výzkumného souboru, d_i^2 je druhá mocnina rozdílu pořadí každého měření. Také je nutné doplnit, že x_{ri} udává pořadí hodnoty x_i v rámci vzestupně uspořádaných hodnot x_1, ..., x_n, číslo y_{ri} jako pořadí hodnoty y_i v rámci vzestupně uspořádaných hodnot y_1, ..., y_n.

Oba korelační koeficienty lze vypočítat i v obecně dostupných počítačových programech, např. v MS Excel nebo pomocí online kalkulátorů. Pro interpretaci korelačního koeficientu opět existují tabulky hodnot, které ukazují míru významnosti jeho hodnot; na intervalech určujících významnost různých hodnot koeficientů se však autoři neshodují (viz tabulku 1).

Tabulka 1
Srovnání věcné významnosti korelačního koeficientu r *dle různých autorů*

významnost / autor	**Cohen (1988)**	**Hendl (2012)**	**Chráska (2016)**
Malá	do 0,3	do 0,3	do 0,4
Střední	0,3–0,5	0,3–0,7	0,4–0,7
Vysoká	od 0,5	od 0,7	od 0,7

Zdroj: Cohen, 1988; Hendl, 2012; Chráska, 2016

Při interpretaci věcné významnosti souvislosti dvou veličin (zejména kardinálních) se nejčastěji využívá právě Pearsonův korelační koeficient. Za nutné však považujeme neopomenout koeficient determinace, neboť ho je možné lépe a názorně interpretovat. Koeficient determinace se vypočítá jako druhá mocnina koeficientu korelace (r^2) a interpretujeme ho jako procento (podíl) vysvětleného rozptylu. Jinými slovy nám tento koeficient říká, z kolika procent je výsledek vysvětlen danou proměnnou a kolik procent vysvětlení připadá na další (nezjištěné) faktory. I pro hodnoty koeficientu determinace byly stanoveny intervaly hodnot, které udávají jejich doporučenou interpretaci; například Soukup (2013) uvádí, že lze uvažovat i o druhých mocninách z doporučovaných hodnot pro korelační koeficient dle Cohena (1988), tedy o hodnotách 0,01 (malá významnost); 0,09 (střední významnost) a 0,25 (velká významnost), je ovšem nutné být k těmto intervalům kritický, správně je interpretovat a uvést jejich limity. Sigmundová a Sigmund (2012) nicméně ještě dodávají, že pokud je výsledek koeficientu determinace větší než 0,1, lze vztah proměnných považovat za významný. I toto doporučení je ale diskutabilní, vždy je třeba posoudit hodnotu v kontextu dané problematiky a výsledků dřívějších studií.

Na tomto místě je nutné dodat, že oba zmíněné korelační koeficienty lze interpretovat z pohledu statistické i věcné významnosti. Pokud mluvíme o Pearsonově nebo Spearmanově korelačním koeficientu jako o míře udávající statistickou významnost, pak máme na mysli *p*-hodnoty, které využíváme pro statistické testování. V rámci závislosti statistické významnosti velikosti korelačního koeficientu na počtu korelovaných dat se ukazuje, že nelze slepě spoléhat pouze na *p*-hodnotu, a to i v případě volby odpovídajícího statistického prostředku (Sigmundová & Sigmund, 2012). Právě *p*-hodnota reprezentuje statistickou významnost. Jsme si tedy vědomi toho, že již samotnou hodnotu r (r_s, r_p) je možné interpretovat jako věcnou významnost, jelikož již sama tato hodnota hovoří o síle závislosti (viz výše).

Míry určené pro testové kritérium s rozdělením chí-kvadrát

Volba koeficientu pro výpočet věcné významnosti pro výsledek chí-kvadrát testu je závislá na počtu kategorií, které mají proměnné v kontingenční tabulce. Pokud se jedná o čtyřpolní tabulku (jedná se tedy o 2 × 2 proměnné, např. chlapci a dívky; správná a chybná odpověď v didaktickém testu), je pro výpočet věcné významnosti určen koeficient Cramerovo φ, který se vypočte dle vzorce 3.1:

$$\varphi = \sqrt{\frac{\chi^2}{n}} \tag{3.1}$$

kde χ^2 je vypočítaná hodnota ze statistického testu a n je rozsah souboru.

Výsledek nabývá hodnot <0; 1>. V současnosti se přestává užívat přepočet výsledku na procenta (po vynásobení výsledku stem), protože povaha dat toto neumožňuje, ale upřednostňuje se intepretace výsledku v kontextu hraničních hodnot intervalu, tedy 0 a 1. Dále jsou pro hodnoty φ znovu určeny tabulkové hodnoty pro interpretaci důležitosti výsledku: **(1)** 0,1–0,29 = malý efekt; **(2)** 0,3–0,49 = střední efekt; **(3)** 0,5 a více = velký efekt.

V případě, že je rozměr kontingenční tabulky větší než 2 × 2, je nutné použít pro výpočet věcné významnosti jiný vzorec, nejčastěji užívané je Cramerovo *V*. Jeho výpočet se provede dle vzorce 3.2:

$$V = \sqrt{\frac{\chi^2}{n\,(k-1)}} \tag{3.2}$$

kde χ^2 je výsledek chí-kvadrát testu pro kontingenční tabulku (testové kritérium), n je rozsah souboru, k je menší z počtu řádků, resp. sloupců.

Výsledek tohoto koeficientu nabývá hodnot <0; 1>, přičemž čím více se hodnota Cramerova *V* blíží 1, tím lze mluvit o vyšší míře věcné významnosti, naopak čím více se hodnota Cramerova *V* blíží 0, tím více můžeme mluvit i nízké věcné významnosti. U Cramerova *V* sice Cohen (1988) stanovil určitá doporučení, ta jsou ale v literatuře spíše kritizována, proto je neuvádíme.

Míry věcné významnosti pro neparametrické postupy

Sociální vědci často pracují s ordinálními veličinami a namísto parametrických postupů (*t*-test, analýza rozptylu, korelace dle Pearsona) volí jejich neparametrické alternativy (Mann Whytneův test, Kruskal-Wallisův test či korelace dle Spearmana, viz výše). Práce s těmito testy je obtížnější (ruční výpočty jsou téměř nemožné) a také věcná interpretace zde bývá obtížnější, čímž bývá často autory článků opomíjena. Nicméně i pro tyto situace vyvinuli statistici samostatné míry věcné významnosti, případně odvodili vzorce, které výsledek testového kritéria neparametrického testu převedou na některou z dříve prezentovaných měr věcné významnosti. Aniž bychom zacházeli do detailů, uvedeme základní možnosti pro dva neparametrické testy: Mann-Whitney pro dvě nezávislé skupiny a Kruskal-Wallis pro více než dvě nezávislé skupiny. Nejčastější bývá přepočet testového kritéria Mann Whytneova testu na Cohenovo *d*, v případě Kruskal-Wallisova testu pak na ukazatel η^2. Oba přepočty opět nabízí pomůcka na portálu německých psychometriků (Lenhard & Lenhard, 2016).[1] Výhodou takto užitých přepočtů je, že získáme nám známé míry, tj. Cohenovo *d* a η^2, která lze poměrně snadno vyhodnocovat a interpretovat. Druhou výhodou je možnost srovnání výsledků získaných v jedné studii parametrickým postupem a druhých získaných neparametricky, což by specifické míry pro neparametrické postupy znemožnilo, proto je zde ani neuvádíme.

Převody mezi jednotlivými mírami věcné významnosti

Až dosud jsme diskutovali jednotlivé míry věcné významnosti a přiřazovali je k různým výzkumným situacím: Cohenovo *d* ke srovnání průměrů dvou skupin, eta (či eta^2) ke srovnání průměrů u více skupin, korelační koeficient či koeficient determinace k vyjádření souvislosti kvantitativních proměnných. Platí ale, že problém porovnání skupin je převoditelný na problém souvislosti proměnných, obdobně platí, že dvě skupiny jsou jen speciálním případem více skupin. Proto statistici odvodili vzorce pro převody jednotlivých měr věcné významnosti mezi sebou. Těchto převodních vztahů (vzorců) jsou desítky. Zájemce o praktické převody hodnot jednotlivých měr věcné významnosti lze odkázat na online pomůcky, nejvíce možností nabízí portál německých psychometriků (Lenhard & Lenhard, 2016).[2] Konkrétně lze z hodnoty Cohenova *d*, korelačního koeficientu, η^2 vypočítat všechny míry uvedené v předchozí části věty. Stejný kalkulátor (jen v jiných záložkách) umí spočítat míry věcné významnosti z testové statistiky (typicky z hodnoty *t*, *Z* či χ^2).

---

1 www.psychometrica.de/effect_size.html. Přepočty nabízí záložka 11 označená „Effect size calculator for non-parametric tests".

2 www.psychometrica.de/effect_size.html. Přepočty nabízí 14. záložka označená jako „Transformation of the effect sizes".

Příklady aplikace měr věcné významnosti a jejich intervalů spolehlivosti

Příklady, na kterých chceme demonstrovat užití měr věcné významnosti společně s intervaly spolehlivosti, vybereme vždy takové, aby demonstrovaly výhody a možné limity interpretace konkrétních měr věcné významnosti. Výběr dat k demonstraci výpočtů bude zahrnovat velká mezinárodní šetření PISA s výběrem v řádu několika tisíc náhodně vybraných respondentů, smyšlené příklady pro vhodnou demonstraci limitů interpretace intervalů spolehlivosti i český výzkum s výběrem několika set respondentů jako příklad blížící se výzkumné praxi na českých univerzitách.

Kromě samotné míry věcné významnosti budeme počítat a interpretovat i intervaly spolehlivosti těchto měr (srov. výklad výše). Pro čtenářský komfort většinou popíšeme, jak lze výsledky získat ručním výpočtem a vždy ukážeme možnosti výpočtu pomocí softwaru. Výpočty uvedených měr lze samozřejmě aplikovat i na jiná data, musí však být dodržena pravidla pro použití konkrétních měr věcné významnosti uvedená v předchozí části článku. Obecně také platí, že pokud chceme míry zobecňovat skrze intervaly spolehlivosti, musí být užit tomu odpovídající design výzkumu, tj. buď náhodný výběr nebo randomizovaný experiment. Ostatně jen pro tyto designy je možné uplatnit i *p*-hodnoty, resp. testy statistické významnosti (pro úplnost a srovnání je budeme uvádět také).

Aplikace míry měřící rozdíl

Jak bylo uvedeno v části *Míry měřící rozdíl*, tyto míry věcné významnosti je možné použít zejména v případě, kdy máme dvě nezávislé skupiny. Pro jednoduchost je tedy použijeme na dnes již klasické srovnání chlapců a dívek a jejich výsledku z matematicky, resp. čtení, jak byl zachycen testem PISA v roce 2018. Celý výpočet a interpretaci výsledků uvádíme pro čtenářův přehled v samostatné části v rámci příkladu 1.

Příklad 1

Pro statistické otestování (ne)shody průměrů mezi chlapci a dívkami v rámci šetření PISA 2018 (Česko, $N = 7019$) lze z důvodu rozsahu souboru a přibližně normálního rozdělení obou závislých proměnných užít dvouvýběrový *t*-test (Welchova verze[3]). Předpokládáme 5% hladinu statistické významnosti, oboustranný test a nulovou hypotézu předpokládající shodnou průměrnou

[3] Pro obě skupiny ukázal Leveneho test neshodu rozptylů, proto byl pro srovnání průměrů užit Welchův test, který shodu rozptylů nepředpokládá.

úroveň gramotností u dívek a chlapců. Zde i v dalších výsledcích využíváme SPSS, které pro malé *p*-hodnoty tiskne hodnotu 0,000. Protože tento formát pokládáme za nevhodný, užíváme pro takové výsledky označení $P < 0{,}0005$, což věcně odpovídá zápisu užitému SPSS. Výsledky statistického testování jsou pak následující:

Čtení: $t = 14{,}246$; $df = 7016{,}132$; $P < 0{,}0005$
Matematika: $t = -3{,}037$; $df = 7016{,}740$; $P = 0{,}002$

Z výsledků snadno zjistíme, že v obou případech je při klasickém postupu výsledek vyhodnocen jako statisticky významný, a to nejen na 5% hladině statistické významnosti, ale i na 1% či 0,5% hladině statistické významnosti, kterou doporučují světoví statistici jako vhodnou modifikaci (srov. výše). Mnozí analytici by tak již zde mohli uzavřít, že rozdíly mezi chlapci a dívkami ve čtení i matematice, existují a dále netřeba pátrat. Nicméně jak jsme již dříve naznačili v textu, není to vhodný postup. Kromě konstatování statistické významnosti (pokud připustíme, že ji má smysl užívat) by bylo vhodné vyhodnotit též významnost věcnou. Samotná pochybnost je ještě umocněna velikostí souboru. Jak argumentují například Soukup a Rabušic (2007), pro velké soubory v řádu tisíců či více jednotek není užívání statistické významnosti vůbec potřebné. S ohledem na srovnání průměrů u dvou skupin využijeme její nejčastější ukazatel, tj. Cohenovo *d*. Pro jeho výpočet je potřebné znát průměry testů v obou skupinách, směrodatné odchylky a velikosti obou skupin. Pro ruční výpočet vybereme jen čtenářskou gramotnost, pro obě gramotnosti pak ukážeme též výpočet v softwaru. Výpočet je tedy následující:

Průměr ze čtenářského testu pro dívky: $\bar{x}_1 = 506{,}95$
Průměr ze čtenářského testu pro chlapce: $\bar{x}_2 = 474{,}59$
Počet dívek: $n_1 = 3432$
Počet chlapců: $n_2 = 3587$
Směrodatná odchylka výsledků testu pro dívky: $s_1 = 92{,}52$
Směrodatná odchylka výsledků testů pro chlapce: $s_2 = 97{,}73$

$$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2}} = \frac{506{,}95 - 474{,}59}{\sqrt{9066{,}245}} = 0{,}34,$$

kde

$$s^2 = \frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2} = \frac{3432 \cdot 92{,}52^2 + 3587 \cdot 97{,}73^2}{3432 + 3587} = 9066{,}245$$

Výsledek Cohenova *d* se rovná 0,34, což značí malý efekt. Rozdíl mezi 15letými dívkami a chlapci ve čtení je tedy malý, dívky jsou v průměru mírně lepší. Mohli bychom tedy konstatovat, že rozdíl mezi dívkami a chlapci v našem souboru (nezapomínejme, že míra věcně významnosti je fakticky popisná

statistika našich dat) je spíše malý, mírně lepší ve čtení byly dívky, rozdíl průměrů v testu činil cca 32 bodů, což při směrodatné odchylce cca 95 bodů poskytne hodnotu Cohenova *d* o velikosti 0,34, tj. rozdíl v průměrech je cca třetina směrodatné odchylky. Pokud si položíme otázku, jak velký rozdíl by mohl být v celé populaci patnáctiletých v Česku (z nichž bylo pro účely PISA 2018 vybíráno), měli bychom užít interval spolehlivosti (zde pro Cohenovo *d*). Ruční výpočet již není snadný. Zatímco výpočty intervalů spolehlivosti pro běžné charakteristiky (průměr, proporce či rozptyl) jsou založeny na běžně známých rozděleních typu normálního, *t*- a *F*-rozdělení či chí-kvadrát rozdělení, pro výpočty intervalů spolehlivost měr věcné významnosti se užívají různé transformace (viz dále u korelace) nebo necentrální obdoby *t*- či *F*-rozdělení, která běžně nenajdeme v tabulkách. Využijeme proto software, resp. se odkážeme i na online pomůcky, které výpočet usnadňují. V našem příkladu využijeme SPSS ve verzi 27, která mj. implementovala výpočet měr věcné významnosti a jejich intervalů spolehlivosti (pro *t*-test, resp. analýzu rozptylu).[4] Dle slibu vypočteme na stejných datech (PISA 2018, Česko, $N = 7019$) míru věcné významnosti (Cohenovo *d* včetně 95% intervalů spolehlivosti pro srovnání výkonu chlapců a dívek) a intervaly spolehlivosti nejen pro čtenářskou, ale i pro matematickou gramotnost:

Čtení: Cohenovo $d = 0{,}34$; dolní mez = 0,29; horní mez = 0,39
Matematika: Cohenovo $d = -0{,}07$; dolní mez = −0,12; horní mez = −0,03

Z vypočtených hodnot je patrné, že náš výpočet Cohenova *d* pro čtenářskou gramotnost (viz výše) byl v pořádku, i software nabídl stejnou hodnotu. Kromě ní nabízí v dalších dvou sloupcích intervalový odhad velikosti Cohenova *d* pro celou populaci (tj. patnáctileté žáky). Pro čtenářskou gramotnost je interval spolehlivosti od 0,29 do 0,39; máme tak jasnější představu o velikost efektu v celé populaci. Rozdíly čtenářské gramotnosti 15letých chlapců a dívek byly v roce 2018 spíše malé, ale nebyly zcela zanedbatelné. Zajímavější jsou údaje pro matematickou gramotnost. Již samotný údaj o Cohenově *d* (−0,07) nás nutí revidovat předčasný závěr, jelikož jsme měli tendenci prohlásit rozdíl za existující a analýzu ukončit. Když nahlédneme na hodnotu Cohenova *d*, případně i na příslušný interval spolehlivosti, lze konstatovat, že věcně je rozdíl mezi chlapci a dívkami v matematické gramotnosti (měřené PISA testem) zcela zanedbatelný, o opravdu velice málo jsou lepší chlapci. Pokud bychom využili dnes již částečně překonané kate-

[4] Cohenovo *d* pro dvouvýběrový test nalezneme v proceduře *Analyze-Compare Means-Two independent sample t-test*. Po zadání proměnných a spuštění procedura poskytne jako jeden z výstupů Cohenovo *d* včetně intervalu spolehlivosti (vypočte i další ukazatele pospané v teoretické části, tj. Glassovo delta a Hedegesovo *g*).

gorie statistické významnosti, mohli bychom věc uzavřít tak, že rozdíly v matematice jsou sice statisticky významné, věcně (prakticky) ovšem bez významu a nemá smysl se jimi zabývat. Ještě jednu možnost nabízí zde vypočtené Cohenovo *d*. I v případě odlišných měřících stupnicích testů (PISA testy mají stupnice srovnatelné, mezinárodně je standardizují na průměr o velikosti 500 a směrodatnou odchylku 100) můžeme provést srovnání věcné významnosti mezi matematikou a čtením. Z hodnot Cohenova *d* pro čtení a matematiku tak jasně plyne, že rozdíly ve čtení jsou téměř pětinásobné oproti rozdílům v matematice. I pro možnost srovnání při různých měřítcích, pro srovnání v čase či možnost shrnutí výsledků skrze metaanalýzu je více než vhodné míry věcné významnosti publikovat. Dodejme, že alternativou pro výpočet Cohenova *d* s intervalem spolehlivosti může být např. volně dostupný software R (například v balíčku *effsize*), dále pak Jamovi, prostředí obdobné SPSS vystavěné na algoritmech R a také mnohé online kalkulačky. Za všechny uvádíme dva odkazy na velmi povedené pomůcky.[5] Pro výpočet je vždy potřebné zadat velikost obou skupin, dále jejich průměry a směrodatné odchylky (tj. údaje které jsme využívali pro ruční výpočet). Poté již kalkulátor vypočte výsledky. Dodejme, že interpretace výsledků stále zůstává na analytikovi, zde se na pomoc softwaru spoléhat nelze.

Aplikace měr měřících rozptyl

Srovnání více skupin

V případě, že máme více než dvě skupiny, které mezi sebou porovnáváme, je nutné využít míry měřící rozptyl. Opět můžeme vyjít z dat PISA 2018 za Česko. Pro lepší srovnatelnost se nyní zaměříme jen na žáky v 1. ročníku středních škol (víceletá gymnázia, čtyřletá gymnázia, střední odborné školy s maturitou, střední odborné školy bez maturity) a srovnáme jejich výsledky ve čtení a matematice dle typu školy, kterou navštěvují. Pro úplnost dodejme, že pro malé počty žáků a horší reprezentativnost nebylo pracováno s konzervatořemi a speciálními středními školami. S jistou mírou nepřesnosti tak zjistíme, jak jsou odlišné výchozí gramotnosti žáků nastupujících do 1. ročníku, resp. srovnání ponecháme v analýze i žáky víceletých gymnázií, kteří jsou již na gymnáziu třetí či pátý rok. Pro analytické zachycení rozdílů lze kromě popisné statistiky a srovnání průměrů užít analýzy rozptylu, pro věcné zachycení rozdílů pak ukazatelů typu čtverce eta (η^2 – Fisherova eta), resp. Haysovo omega (s ohledem na velikost dat lze čekat, že obě míry poskytnou srovnatelné výsledky). Postup výpočtu a interpretaci dat uvádíme v příkladu 2.

---

[5] (1) https://www.danielsoper.com/statcalc/category.aspx?id=8; (2) https://www.psychometrica.de/effect_size.html.

Příklad 2

Nejdříve opět prezentujeme výsledky analýzy rozptylu, pro čtenářskou i matematickou gramotnost (tabulka 2).

Tabulka 2

Analýza rozptylu pro srovnání výkonu žáků z různých středních škol v matematice a čtení

Čtení					
	Součet čtverců	Stupně volnosti	Průměrné čtverce	*F*	*p*-hodnota
Meziskupinový	9785162,932	3	3261720,977	629,555	< 0,0005
Vnitroskupinový	15563719,623	3004	5180,999		
Celkem	25348882,556	3007			
Matematika					
	Součet čtverců	Stupně volnosti	Průměrné čtverce	*F*	*p*-hodnota
Meziskupinový	8332388,711	3	2777462,904	545,402	< 0,0005
Vnitroskupinový	15297891,669	3004	5092,507		
Celkem	23630280,380	3007			

Zdroj: PISA 2018 (Česko), N = 3,008

Z tabulky 2 plyne, že pro čtenářskou i pro matematickou gramotnost je *p*-hodnota velmi nízká, tedy rozdíl je zobecnitelný na celou populaci, tj. rozdíly mezi typy škol existují. Mohli bychom proto konstatovat, že výsledek je statisticky významný (nezaměřujeme se zde detailně na substantivní popis, ale v souladu s očekáváním jsou nejlepší výsledky žáků z gymnázií a nejhorší u žáků na učebních oborech, srov. dále). Otázkou však zůstává (po předchozí analýze rozdílů mezi chlapci a dívkami v příkladu 1), zda jsou rozdíly mezi typy škol též věcně významné – a pokud ano, jak výrazně.

Nyní uvedeme postup ručního výpočtu. Předpokládáme, že sledujeme veličinu Y v k skupinách, v i-té skupině je n_i pozorování; j-té pozorování v i-té skupině budeme označovat y_{ij}. Pro čtenářskou gramotnost tak výpočet probíhá následovně:

Celkový součet čtverců: $SS_T = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}\right)^2 = 25348882{,}6$

Meziskupinový součet čtverců: $SS_b = \sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 = 9785162{,}9$

V těchto vzorcích $\bar{y}$ značí celkový průměr a $\bar{y}_i$ značí průměr v i-té skupině.

$$\eta^2 = \frac{SS_b}{SS_T} = \frac{9785162{,}9}{25348882{,}6} = 0{,}39$$

Můžeme konstatovat, že téměř 40 % rozptylu výsledků žáků ve čtení lze vysvětlit skrze typ navštěvované střední školy. Nelze však věcně konstatovat,

že tyto rozdíly způsobila výuka na středních školách, neboť se jedná o žáky v prvním ročníku. Tato vysoká hodnota věcné významnosti ukazuje na výraznou selekci českých dětí do gymnázií, středních maturitních oborů a středních nematuritních oborů (neprezentujeme zde detailní popisné statistiky, ale žáci nematuritních oborů mají výsledek v testu čtenářské gramotnosti téměř o 200 bodů nižší, což by – jak víme z předchozích srovnání – znamenalo Cohenovo *d* o velikosti cca 2). Uzavřeme, že výsledek je zde nejen statisticky, ale i věcně významný.

Opět ukažme, že ruční výpočet není v dnešní době potřebný. Můžeme využít SPSS, verzi 27, která kromě hodnoty η^2 a ω vypočte i intervaly spolehlivosti těchto měr.[6] Opět provedeme srovnání pro matematickou a čtenářskou gramotnost. Výsledky nabízí tabulka 3.

Tabulka 3
Fisherovo η^2 a Haysovo ω^2 (vč. 95% intervalů spolehlivosti) pro srovnání výkonu v matematice a čtení dle typu navštěvované střední školy

Sledované proměnné	*Matematika*			*Čtení*		
	Odhad	Dolní mez	Horní mez	Odhad	Dolní mez	Horní mez
η^2	0,353	0,327	0,377	0,386	0,361	0,410
ω^2	0,352	0,326	0,376	0,385	0,360	0,409

Zdroj: PISA 2018 (Česko), N = 3,008

Z tabulky 3 zjistíme, že náš ruční výpočet Fisherovy Ety^2 byl správný, přesná hodnota pro čtení na tři desetinná místa je 0,386. Platí již uvedený komentář o výrazných rozdílech mezi typy škol. Nic na tom nezmění ani interval spolehlivosti, pro Fisherovu eta^2 v případě čtenářské gramotnosti je mezi 0,36 a 0,41.[7] Můžeme též konstatovat, že nezkreslený odhad, který na rozdíl od Fisherovy ety^2 poskytuje omega, není nijak odlišný, bodový odhad činí v případě čtenářské gramotnosti 0,385 a liší se tedy až na třetím desetinném místě; i interval spolehlivosti vykazuje jen kosmetické odlišnosti. Máme tedy empirický „důkaz", že pro velké soubory jsou hodnoty η^2 i ω^2 velmi podobné.

---

6 V SPSS najdeme příslušné výpočty skrze proceduru Analyze-Compare means-One way ANOVA. Výpočty eta a omega (vč. intervalů spolehlivosti) jsou defaultním výstupem, získáme je tedy po zadání proměnných a spuštění procedury.

7 Interval spolehlivosti je poměrně malý, což je zejména způsobeno tím, že náš výzkumný soubor je velký. Pro malé soubory by byl interval za jinak stejných okolností výrazně větší. Platí tedy, že zatímco hodnoty míry věcné významnosti velikost souboru neovlivňuje, jejich interval spolehlivosti ovlivněn velikostí souboru je (jde o obecnou vlastnost intervalů spolehlivosti).

To je výhodné zejména pro případný ruční výpočet, protože u ω^2 není zcela snadný (srov. Soukup, 2013).

Závěrem doplňme srovnání měr věcné významnosti pro matematickou a čtenářskou gramotnost, pro srovnání vyberme η^2. Z tabulky 3 je patrné, že hodnoty η^2 se liší pro obě testované oblasti jen minimálně, a lze konstatovat, že rozdíly mezi žáky různých typů středních škol v matematice a čtení jsou obdobné (velké).

Souvislost dvou a více proměnných (korelace)

Dalším ukazatelem měr věcné významnosti měřící rozptyl je koeficient korelace, resp. z něho odvozený koeficient determinace. Pro ilustraci korelačního koeficientu a jeho intervalu spolehlivosti využijeme naposledy česká data ze šetření PISA 2018. Konkrétně si položíme otázku: jaká je souvislost mezi výsledkem testu čtenářské gramotnosti a socio-ekonomicko-kulturním zázemím vyjádřeném v PISA šetření syntetickým ukazatelem označovaném jako ESCS? Výpočet a interpretaci těchto dat uvádíme v příkladu 3.

Příklad 3

S ohledem na charakter obou proměnných (kardinální proměnné se symetrickým rozdělením) zvolíme Pearsonův korelační koeficient a následný test jeho nulové hodnoty. Základní výsledek, který bychom získali například z SPSS[8], by měl podobu zobrazenou v tabulce 4.

Tabulka 4

Korelační koeficient pro souvislost čtenářské gramotnosti a socio-ekonomicko-kulturního zázemí a test jeho statistické významnosti

	Čtení	ESCS
Čtení	1	0,404
		< 0,0005
	7019	6920

Hodnota korelačního koeficientu je po zaokrouhlení na jedno desetinné místo 0,4; výsledkem statistického testu je velice nízká p-hodnota (menší než 0,0005), a proto bychom výsledek označili za průkazný, resp. zobecnitelný.

[8] Výpočet získáme skrze proceduru Analyze-Correlate-Bivariate. Po zadání minimálně dvojice proměnných vypočte procedura jejich korelace dle Pearsonova vzorce. Pokud chce analytik jiný koeficient, musí to ve vstupním dialogu nastavit.

V případě interpretace věcné významnosti pro tento příklad bychom mohli postupovat následovně. Náhledem na doporučení podaná dříve zjistíme, že přiléhavým slovním označením používaným pro takto vysokou korelaci by byla střední souvislost. Opět dodejme, že tato generická doporučení nejsou příliš vhodná k užívání. Mnohem lepší by bylo provést srovnání získané korelace s dalšími výsledky (například za jiné země účastnící se šetření PISA 2018). Zjistili bychom tak, že sociální podmíněnost školních výsledků v Česku je jedna z největších. Důvody, které za tím stojí, zde detailně rozebírat nebudeme a odkážeme na bohatou literaturu z oblasti sociologie vzdělání (např. Matějů et al., 2007).

Místo toho nabídneme jinou možnou interpretaci korelačního koeficientu. Pokud umocníme korelační koeficient na druhou, získáme koeficient determinace (r^2). Ten udává procentuální podíl variability, a je tak obdobou η^2 či ω^2. V našem příkladu bychom umocněním 0,404 získali hodnotu 0,165, tj. 17 % rozptylu čtenářské gramotnosti lze vysvětlit skrze socio-ekonomicko-kulturní zázemí žáka. Není to tedy extrémní vliv, ale nejde též o zanedbatelnou hodnotu.

Analogicky jako u Cohenova *d* a předchozích měr zaměřených na rozptyl by bylo i pro korelaci vhodné spočítat interval spolehlivosti. I zde platí, že nejde úplně o jednoduchý postup, nejužívanější možností je tzv. Fisherova transformace (pro detailní popis viz Soukup, 2013). Pro výpočet lze použít buď speciální software (interval spolehlivosti umí počítat R, z dalších volně dostupných softwarů pak Jamovi či JASP) nebo existují i mnohé online pomůcky.[9] V často užívaném SPSS se nabízí dvě možnosti. První je využít odhad skrze Bootstrap (pro detailní ukázku viz Rabušic et al., 2019, s. 307–310). Mírně uživatelsky náročnější je využít malé programy pro SPSS, které připravili různí autoři. Pro ilustraci využijeme pomůcku, která je dostupná přímo na stránkách IBM (https://www.ibm.com/support/pages/confidence-intervals-correlations). Modifikaci (počeštění) této pomůcky pro náš výpočet nabízíme v online příloze článku. Zadáme-li do pomůcky naše údaje (z tabulky 4), tj. přesnou velikost korelace 0,404, počet žáků pro korelaci (N = 6920) a požadovanou míru spolehlivosti (volíme zde 0,95), získáme výstup (SPSS zobrazí výstup jednak ve výstupním okně, jednak i v datovém okně), z nějž bude patrné, že *lo_r* = 0,384 (představuje dolní mez intervalu spolehlivosti) a *hi_r* = 0,424 (představuje horní mez intervalu spolehlivosti).

Doplňme, že by samozřejmě bylo následně možné dolní a horní mez umocnit na druhou a získat tak interval spolehlivosti pro koeficient determinace. O tom se více zmíníme v další části článku.

[9] Například danielsoper.com/statcalc/calculator.aspx?id=28.

Souvislost dvou a více proměnných (regrese)

Obdobně jako u korelačních úloh lze i pro regresní analýzu počítat koeficient determinace (r^2). Interpretace tohoto ukazatele je analogická, ideálně po vynásobení stem interpretujeme v procentech jako míru vysvětlení závisle proměnné skrze nezávisle proměnné. Je častým nešvarem, že autoři kvantitativních studií r^2 vůbec nepublikují, případně jej neinterpretují, častěji v případech, kdy jsou hodnoty tohoto ukazatele malé. Zcela zřídka se pak objevují intervaly spolehlivosti pro koeficient determinace a jejich interpretace. Přitom právě tyto výsledky by mohly vést k nepřeceňování výsledků regresní analýzy. Zde nabídneme jen jednoduchou ilustraci na smyšleném příkladu (příklad 4). Důvodem výběru těchto ilustrativních dat je ukázat limity interpretace výsledku na relativně nižším rozsahu výzkumného souboru.

Příklad 4

Mějme regresní model, kde R^2 dosáhlo hodnoty 0,2. Tento model byl vypočten na výběrovém souboru obsahujícím 100 jednotek a v modelu byly využity dvě vysvětlující proměnné. Formálně zapsáno: $R^2 = 0{,}2$; $n = 100$ a $k = 2$.

Nyní je vhodné opět využít některou z online pomůcek, pro zadání využijeme právě popsané hodnoty. Využijeme opět pomůcku připravenou Danielem Sopperem.[10]

Výsledek výpočtu pro 95% interval spolehlivosti poskytne dolní mez 0,06 a horní mez 0,34. Zatímco hodnota R^2 byla poměrně optimistická (20% míra vysvětlení), dolní mez intervalu spolehlivosti nám sděluje, že míra vysvětlení může být v našem modelu v populaci velice malá (6% míra vysvětlení).

Aplikace měr určených pro testové kritérium chí-kvadrát

V případě, že posuzujeme souvislost dvou nebo více nominálních, resp. dichotomických proměnných, využijeme ke statistické analýze klasicky chí-kvadrát test pro čtyřpolní nebo kontingenční tabulku. Kromě výsledku testu bývá zvykem prezentovat i některé kontingenční koeficienty (srov. výklad výše). I pro tyto koeficienty je možné počítat intervaly spolehlivosti a tím splnit doporučení, že se má publikovat míra věcné významnosti včetně intervalu spolehlivosti. Opět zde narážíme na to, že běžně užívané SPSS (ale i jiné platformy) intervaly spolehlivosti pro kontingenční koeficienty nezobrazují, a proto je uživatelé těchto platforem nepublikují. Opět se dá využít speciálního software, nejvíce možností nabízí *R* (např. funkce confintCramersV v balíčku

[10] Pomůcka online dostupná z: https://www.danielsoper.com/statcalc/calculator.aspx?id=28. V případě indexu determinace tato pomůcka počítá 90%, 95% a 99% interval spolehlivosti.

MBESS), další volbou jsou online pomůcky (jež bývají připraveny jen pro tabulky 2×2[11]). V rámci SPSS lze pak pro odhad využít Bootstrap. Opět si ukážeme jednoduchý příklad (příklad 5), kde využijeme právě Bootstrapu v SPSS pro Cramerovo *V*.

Příklad 5

Převzali jsme příklad analyzovaný v článku Paprštеinová et al. (2011, s. 171), který dával do vztahu typ školy, kde učitel působí, a subjektivní hodnocení mimopracovní pohybové aktivity (viz tabulku 5 níže). V původním článku byla publikována jen *p*-hodnota pro chí-kvadrát test, nic dalšího uvedeno nebylo. Dodejme navíc, že použití testu bylo v tomto případě diskutabilní zejména kvůli tomu, že počet buněk tabulky s nízkými očekávanými četnostmi byl poměrně velký (cca jedna čtvrtina). Uveďme, že následující příklad, ačkoliv byl aplikován na výsledky dotazníkového šetření, lze vhodně aplikovat i na data vycházející z didaktického testování. Tabulku jsme načetli do SPSS (příkazy k tomu potřebné včetně analytických postupů jsou v online příloze článku) a získali jsme tyto výsledky:

Testové kritérium chí-kvadrát: $\chi^2 = 30{,}7$
p-hodnota = 0,002
Počet respondentů: $n = 484$
Počet stupňů volnosti: $df = 12$
Cramerovo $V = 0{,}145$

Z pohledu statistické významnosti je výsledek průkazný, *p*-hodnota je malá. Souvislost mezi mírou pohybové aktivity a typem školy, kde učitel působí, tedy nejspíše existuje. Při pohledu na hodnotu Cramerova *V* ale můžeme konstatovat, že tato souvislost je poměrně slabá. Statisticky jde tedy o vysoce významný výsledek, věcný význam je spíše malý. Namístě je tedy vypočítat interval spolehlivosti Cramerova *V*, který by mohl získaný výsledek ještě více zpochybnit. Zde využijeme pro odhad Bootstrap v SPSS a necháme udělat 1000 bootstrapových výběrů.[12] Výsledný interval spolehlivosti pro Cramerovo *V* bude následující:

Dolní mez: 0,12
Horní mez: 0,22

---

[11] Například: http://www.vassarstats.net/odds2x2.html; nebo velmi detailní: https://statpages.info/ctab2x2.html.

[12] V rámci procedury na kontingenční tabulky *Analyze-Descriptive statistics-Crosstabs* musíme v dílčí volbě *Statistics* zaškrtnout *Phi and Cramer's V* a poté navštívit dílčí volbu *Bootstrap*. Zde musíme zaškrntout volbu *Perform bootstrapping*. Po spuštění pak získáme bootstrapový interval spolehlivosti pro Cramerovo *V* (defaultně 95%).

Interval spolehlivosti nám tedy říká, že Cramerovo V může být i jen 0,12, tedy ještě menší než naznačovala jeho prostá hodnota ve výběru.

Tabulka 5
Kontingenční tabulka pro výpočty v příkladu 5

Fyzická aktivita / Typ školy	ZŠ	Gymnázia	Víceletá gymnázia	SOŠ	VŠ	Celkem
Velmi nízká	15	2	4	19	14	54
Nízká	68	12	7	59	19	165
Průměrná	112	21	16	70	16	235
Vysoká	6	4	0	13	6	29
Celkem	201	39	27	161	55	483

Zdroj: Papršteinová et al. (2011, s. 171)

Kritická diskuze aplikace měr věcné významnosti do výzkumu a jejich srovnání s aplikací měr statistické významnosti

Aplikace výpočtů věcné významnosti představené v tomto textu má samozřejmě své výhody a nevýhody. Tabulka 6 shrnuje pozitiva a negativa aplikace věcné významnosti do výzkumných studií a srovnává je s výpočty statistické významnosti. Autoři studie, v níž jsou aplikovány míry věcné významnosti, by měli v diskuzi výsledků upozornit na limity těchto výpočtů, aby si i méně znalý čtenář mohl vytvořit kritický pohled na výsledky předkládané studie a jejich využitelnost v praxi.

Tabulka 6
Výhody a nevýhody aplikace výpočtů statistické a věcné významnosti

	Statistická významnost	Věcná významnost
Pozitiva aplikace	• možnost zobecnění výsledků z výběru na populaci • dostupná literatura k problematice • dnes běžná publikační praxe a součást výuky	• hodnota míry není závislá na velikosti výběrového souboru • zajišťuje srovnatelnost mezi jednotlivými výzkumy (metaanalýza) • vypovídá o velikosti rozdílu nebo o míře vysvětlení souvislosti (často v procentech)
Negativa aplikace	• závisí na velikosti výběru a v případě velkých souborů je užíván zbytečně, vede k přeceňování výsledků • práce s kritickou hodnotou 5 % („dogma“) • někteří autoři neprezentují statisticky nevýznamné výsledky a považují je za nedůležité • aplikace na nevhodné soubory (a především absence zdůvodněných limitů studie v její diskuzi) • časté zaměňování s věcnou významností	• bez intervalu spolehlivosti se jedná o popisnou charakteristiku (nelze zobecnit na základní soubor) • doporučené tabulkové hodnoty a jejich slepé následování • většina publikací je nepoužívá, absence výuky na vysokých školách, nedostatek odborné literatury • často je třeba využít speciální software či online pomůcky (situace se však zlepšuje) • v praxi výpočty nejsou vyžadovány (např. redakční rady odborných časopisů)

Pro efektivní aplikaci měr věcné významnosti se nabízí i otázka, v jakém sledu použít výpočet věcné a statistické významnosti. V této studii vycházíme z určité tuzemské tradice již proběhlých výzkumů především v kinantropologii (např. Blahuš, 2000; Sigmundová & Sigmund, 2012) a používáme výpočet věcné významnost až po statisticky významném testování nulové hypotézy (viz příklady 1–5). V současnosti je doporučován postup upřednostňující **používání obou významností současně**, samozřejmě se správnou interpretací. Například APA i AERA (např. APA, 2001, 2010; AERA, 2006) doporučuje nejprve (!) posoudit věcnou významnost a následně posoudit statistickou významnost, tedy možnost zobecnit výsledky. Jako vhodné se však jeví, i na základě informací z předcházejících částí článku, doplnit výpočet věcné významnosti o intervaly spolehlivosti, jež mají ambici tento v základě popisný ukazatel zobecnit na základní soubor.

Jak vyplývá z tabulky 6, věcná významnost má v užívání několik limitů. Autoři by neměli zapomínat se v diskuzi kriticky vyjádřit k jejich limitům stejně jako k limitům konkrétní studie a k možnostem potenciálního zkreslení výsledků různými faktory. Jedině tak lze nakládat s výsledky empirických kvantitativních studií ve správných souvislostech a nepodnikat chybné či zkreslující závěry.

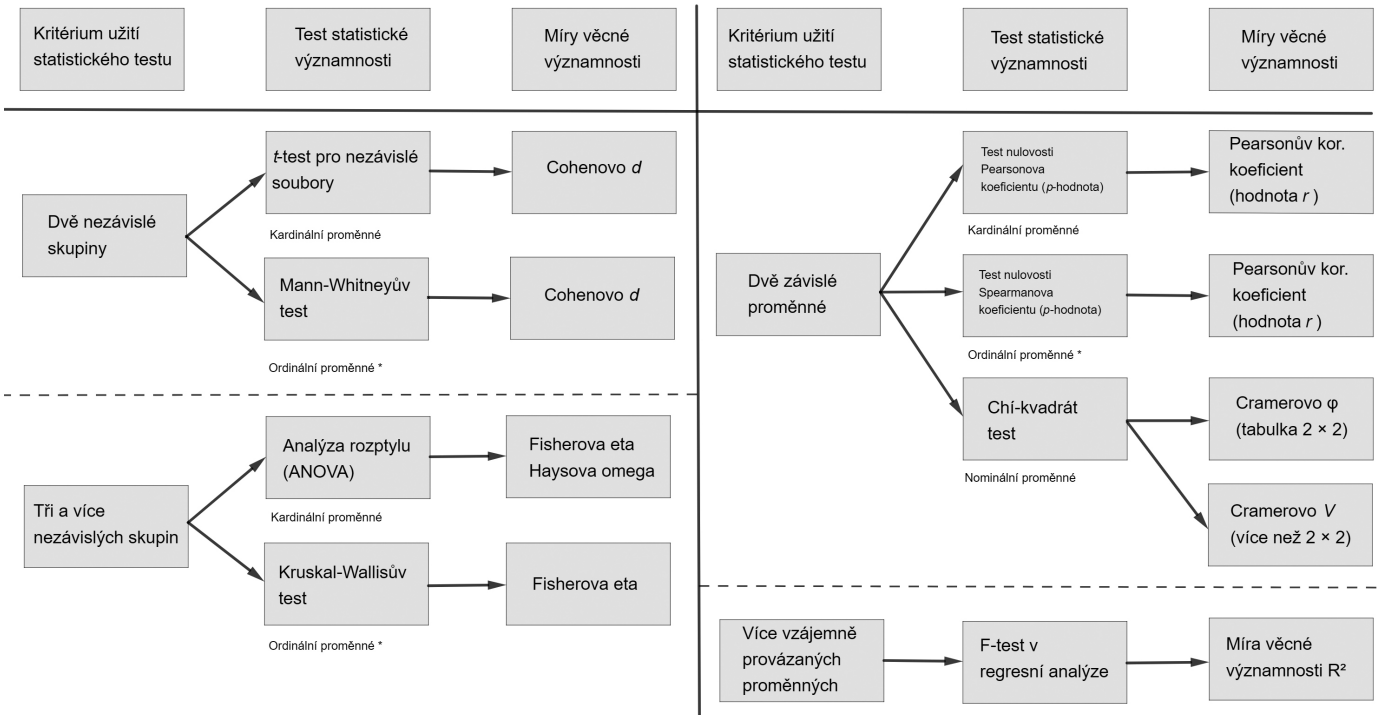
Závěrečné shrnutí

Koncept věcné významnosti má praktické užití v oblasti pedagogického výzkumu. Oproti statistické významnosti nabízí řadu výhod, které mají potenciál obohatit interpretaci výsledků o nové skutečnosti. Předkládaný text dává praktický návod na použití nejdůležitějších měr věcné významnosti v analýzách tradičních výzkumných metod pedagogiky a jejich dílčích subdisciplín (např. oborová didaktika) včetně jejich výpočtu intervalů spolehlivosti (často pomocí online kalkulátorů nebo často používaného programu).

Jak bylo řečeno v úvodní kapitole, použití různých statistických procedur se řídí jasnými pravidly, která je nutné respektovat. Obrázek 1 dává do vztahu základní míry věcné významnosti s příslušnými testy statistické významnosti; toto schéma je vždy rozděleno podle toho, s kolika skupinami pracujeme, s jakými typy proměnných pracujeme nebo zda je řešena otázka rozdílu či závislosti. Jedná se o orientační návod odpovídající výše zmíněnému textu, nejsou zde zahrnuty všechny možnosti volby statistického testu pro porovnání skupin a hodnocení jedné skupiny tak, jak uvádí Hendl (2012).

Obrázek 1

Příklady použití konkrétních měr věcné významnosti v závislosti na kritérii výběru statistického testu a testu statistické významnosti



* Příslušné testy (Mann-Whitney, Kruskal-Wallis nebo Spearman) je možné realizovat taktéž pro kardinální proměnné. Jde o neparametrické metody, u kterých se vychází z pořadí hodnot a u kardinálních proměnných není vyžadována jejich normalita. Testy uvedené u kardinálních proměnných jsou pro tento typ proměnné typické, využíváme je zejména pro větší datové soubory (kvůli centrální limitní větě), u menších souborů pak při dodržení normality dat.

Literatura

AERA. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33–40. https://doi.org/10.3102/0013189X035006033

American Psychological Association [APA]. (2001). *Publication manual of the American Psychological Association*.

American Psychological Association [APA]. (2010). *Publication manual of the American Psychological Association* (6th ed.).

American Psychological Association [APA]. (2020). *Publication manual of the American Psychological Association* (7th ed.). APA. https://doi.org/10.1037/0000165-000

Benjamin, D. J., Berger, J. O., Johannesson, M., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behavior, 2*(1), 6–10. https://doi.org/10.1038/s41562-017-0189-z.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention, 2*(3), 129–141. https://doi.org/10.1080/17489530802446302.

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

Blume, J. D., Greevy, R.A., Welty, F.W., Smith, J. R., & Dupont, W. D. (2020). An introduction to second-generation p-values. *The American Statistician, 73*(1), 157–167. https://doi.org/10.1080/00031305.2018.1537893.

Blume, J. D., McGowan, L. D., Greevy, R. A., & Dupont, W. D. (2018). Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses. *PLoS One, 13*(3), 1–17. https://doi.org/10.1371/journal.pone.0188299.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to metaanalysis*. John Wiley & Sons.

Cohen, J. (1988). *Statistical power analysis for the behavioral science*. Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037//0033-2909.112.1.155.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education*. Routledge.

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. https://doi.org/10.1080/00220670109598763.

Ferjenčík, J. (2010). *Úvod do metodologie psychologického výzkumu*. Portál.

Fico, M. (2020). Nástroje na zlepšenie štatistickej inferencie – analýza p-kriviek a ekvivalenčné testovanie. *Sociológia, 52*(4), 323–353. https://doi.org/10.31577/sociologia.2020.52.4.14.

Harris, R. J. (1997). *Reforming significance testing via three-valued logic. What if there were no significance tests?* Erlbaum.

Hedges L. V. (2008). What are effect sizes and why do we need them? *Developmental Psychology Perspectives, 2*(3), 167–171. https://doi.org/10.1111/j.1750-8606.2008.00060.x.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for metaanalysis*. Academic Press.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*(3), 224–239. https://doi.org/10.1002/jrsm.1052.

Hendl, J. (2012). *Přehled statistických metod: analýza a metaanalýza dat*. Portál.

Hendl, J. (2014). *Statistika v aplikacích*. Portál.

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the neyman-pearson decision-theoretic framework and rise of the neofisherian. *Annales Zoologici Fennici, 46*(5), 311–349. https://doi.org/10.5735/086.046.0501.

Chráska, M. (2016). *Metody pedagogického výzkumu: základy kvantitativního výzkumu*. Grada.

Ialongo, C. (2016). Understanding the effect size and its measures. *Biochem Med, 26(*2), 150–163. https://doi.org/10.11613/BM.2016.015

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review, 67*(3), 160–167. https://doi.org/10.1037/h0047595.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences*. APA.

Kočvarová, I., & Soukup, P. (2018). Výuka kvantitativní analýzy dat jako součást metodologie výzkumu v pedagogických studijních programech veřejných vysokých škol v ČR. *Orbis Scholae*, *12*(3), 127–145. https://doi.org/10.14712/23363177.2019.2.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*(863), 1–38. https://doi.org/10.3389/fpsyg.2013.00863.

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

Lenhard, W., & Lenhard, A. (2016). *Calculation of effect sizes*. Psychometrica. https://www.psychometrica.de/effect_size.html.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject design: Percentage of data points exceeding the median. *Behavior Modification, 30*(5), 598–617. https://doi.org/10.1177/0145445504272974.

Matějů, P., Soukup, P., & Basl, J. (2007). *Educational aspirations in a comparative perspective. The role of individual, contextual and stuctual factors in the formation of educational aspirations in OECD countries*. Sociologický ústav AV ČR.

McShane, B. B., Gal, D., Gelman, A., Robert, Ch., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician, 73*(1), 235–245. https://doi.org/10.1080/00031305.2018.1527253.

Paprštеinová, M., Smejkalová, J., Hodačová, L., & Čermáková, E. (2011). Zdravotní stav a životní styl učitelů různých stupňů škol. *Pedagogika, LXI* (2), 164–174.

PISA. (2018). *OECD's Programme for International Student Assessment*. https://www.oecd.org/pisa/publications/pisa-2018-results.htm

Rabušic, L., Soukup, P., & Mareš, P. (2019). *Statistická analýza sociálněvědních dat (prostřednictvím SPSS)*. Masarykova univerzita.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615–620. https://doi.org/10.1007/s10979-005-6832-7.

Shourki, M. M., & Edge, V. L. (1996). *Statistical methods for health sciences*. CRC Press.

Sigmundová, D., & Sigmund, E. (2012). Statistická a věcná významnost a použití dat o pohybové aktivitě. *Tělesná kultura, 35*(1), 55–72. https://doi.org/10.5507/tk.2012.004.

Soukup, P. (2013). Věcná významnost výsledků a její možnosti měření. *Data a výzkum, 7*(2), 125–148. http://dx.doi.org/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. https://doi.org/10.5817/PedOr2016-2-182.

Soukup, P., & Kočvarová, I. (2016). Velikost a reprezentativita výběrového souboru v kvantitativně orientovaném pedagogickém výzkumu. *Pedagogická orientace, 26*(3), 512–536. https://doi.org/10.5507/fzv.

Soukup, P., & Rabušic, L. (2007). Několik poznámek k jedné obsesi českých sociálních věd – statistické významnosti. *Sociologický časopis/Czech Sociological Review*, *43*(2), 379–395. https://doi.org/10.13060/00380288.2007.43.2.06.

Sullivan, G. M., & Feinn, R. (2012). Using effect size – or why the p value is not enough. *Journal of Graduate Medical Education, 4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1.

Steiger, J. H., & Fouladi, R. T. (1997). *Noncentrality interval estimation and the evaluation of statistical models. What if there were no significance tests*? Erlbaum.

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *TRENDS in Sport Sciences, 21*(1), 19–25.

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

Valentine, J. C., & Cooper, H. (2003). *Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes*. Western Michigan University. https://wmich.edu/sites/default/files/attachments/u58/2015/Effect_Size_Substantive_Interpretation_Guidelines.pdf.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "$p < 0.05$". *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*(8), 594–604. https://doi.org/10.1037/0003-066X.54.8.594.

Kontakt na autory
Petr Soukup
Institut sociologických studií, Fakulta sociálních věd, Univerzita Karlova
E-mail: soukup@fsv.cuni.cz

Petr Trahorsch
Katedra geografie, Přírodovědecká fakulta, Univerzita J. E. Purkyně v Ústí nad Labem
E-mail: petr.trahorsch@ujep.cz

Vlastimil Chytrý
Katedra preprimárního a primárního vzdělávání, Pedagogická fakulta, Univerzita J. E. Purkyně v Ústí nad Labem
E-mail: vlchytry@gmail.com

Corresponding authors
Petr Soukup
Institute of Sociogical Studies, Faculty of Social Sciences, Charles University
E-mail: soukup@fsv.cuni.cz

Petr Trahorsch
Department of Geography, Faculty of Science, J. E. Purkyně University in Ústí nad Labem
E-mail: petr.trahorsch@ujep.cz

Vlastimil Chytrý
Department of Pre-primary and Primary Education, Faculty of Education, J. E. Purkyně University in Ústí nad Labem
E-mail: vlchytry@gmail.com

Příloha: přehled klíčových měr věcné významnosti, jejich výpočet a pravidla užití

Název míry	Výpočet (vzorec)	Interpretace, zhodnocení výsledku	Pravidla použití, nevýhody
Cohenovo *d*	$d = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2}}$ $\bar{x}_1$ a $\bar{x}_2$ jsou průměry první a druhé skupiny s^2 je rozptyl společný oběma skupinám, který se vypočte dle vzorce: $s^2 = \frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2}$ n_1 a n_2 je velikost prvního a druhého souboru; s_1^2 a s_2^2 jsou rozptyly v první a druhé skupině.	0,2–0,5: malý efekt 0,5–0,8: střední efekt 0,8 a více: velký efekt	– Míra měřící rozdíl výsledků u dvou skupin – Velikosti obou testovaných skupin jsou stejně nebo podobně velké – Rozptyl obou výběrů se významně neliší
Hedgesovo *g*	$g = (\bar{x}_1 - \bar{x}_2)/\sqrt{MS_w}$ $\bar{x}_1$ a $\bar{x}_2$ jsou průměry první a druhé skupiny MS_w je průměr vnitroskupinového součtu	0,2–0,5: malý efekt 0,5–0,8: střední efekt 0,8 a více: velký efekt	– Míra měřící rozdíl výsledků u dvou skupin – Je vhodné použít v případech, kdy byl *t*-test aplikován na velmi malých souborech – Pro výzkumné soubory, které nebyly početně stejné
Glassovo delta	$\Delta = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_k^2}}$ $\bar{x}_1$ a $\bar{x}_2$ jsou průměry první a druhé skupiny $\sqrt{s_k^2}$ je rozptyl srovnávací (kontrolní) skupiny	0,2–0,5: malý efekt 0,5–0,8: střední efekt 0,8 a více: velký efekt	– Míra měřící rozdíl výsledků u dvou skupin – Je vhodné použít v případech s velkých rozdílem směrodatných odchylek nebo pro kontrolní skupinu – Je výhradně určen pro analýzu výsledku experimentů (přítomnost kontrolní skupiny)
Fisherova eta (η^2)	$\eta^2 = SS_b/SS_T$ SS_T je celkový součet čtverců SS_b je meziskupinový součet čtverců	po vynásobení 100 jako % vysvětlená daným faktorem	– Měří rozptyl mezi více než dvěma skupinami (užití např. při analýze rozptylu) – Zkresluje odhad charakteristiky v základním souboru – Vhodné pro experimentální designy výzkumu

Haysova omega	$= \frac{SS_b-(v-1)\cdot MS_w}{SS_t+MS_w}$, kde SS_t je celkový součet čtverců, SS_b je meziskupinový součet čtverců, v je počet srovnávaných skupin, MS_w je průměrný vnitroskupinový součet čtverců.	po vynásobení 100 jako % vysvětlená daným faktorem	– Měří rozptyl mezi více než dvěma skupinami (užití např. při analýze rozptylu) – Nezkresluje odhad charakteristiky v základním souboru – Vhodné pro experimentální designy výzkumu
korelační koeficient *r*	$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{S_x \cdot S_y}$ (případně $\frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{(n-1)S_x S_y}$) $\bar{x}$ a $\bar{y}$ jsou výběrové průměry S_x a S_y jsou výběrové směrodatné odchylky	dle Hendla (2012) do 0,3: malá významnost 0,3–0,7: střední významnost od 0,7: vysoká významnost	– Určeno pro testování souvislosti mezi dvěma proměnnými – Data by měla být metrické povahy, normálního rozdělení a z náhodného výběru – Lze přepočítat i testová kritéria *t*-testu a Cohenova *d* na koeficient *r*
koeficient determinace	r^2 *r* je korelační koeficient	0,01: malá významnost 0,09: střední významnost 0,25: velká významnost po vynásobení 100 jako % vysvětlená daným faktorem	– Výpočet z korelačního koeficientu – Hodnotu podílu vysvětleného rozptylu ze základního souboru nadhodnocuje > užití upraveného koeficientu determinace
upravený koeficient determinace	$R^2_{adj} = \frac{1-(1-r^2)\times(n-1)}{n-k-1}$, *n* je velikost výběru *r* je korelační koeficient *k* je počet nezávisle proměnných.	po vynásobení 100 jako % vysvětlená daným faktorem	– Lze použít v případech malých výběrů, neboť koeficient determinace v těchto případech výsledek nadhodnocuje
Cramerovo	$\varphi = \sqrt{\frac{\chi^2}{n}}$, χ^2 je vypočítaná hodnota ze statistického testu *n* je rozsah souboru.	0,1–0,29: malý efekt 0,3–0,49: střední efekt 0,5 a více: velký efekt po vynásobení 100 jako % vysvětlená daným faktorem	– Určeno pro výsledek statistického testu z čtyřpolní tabulky
Cramerovo	$V = \sqrt{\frac{\chi^2}{n(k-1)}}$, kde χ^2 je výsledek chí-kvadrát testu pro kontingenční tabulku (testové kritérium), *n* je rozsah souboru a *k* je menší hodnota z počtu řádků, resp. počtu sloupců.	hodnota z intervalu <0; 1> hodnota blížící se 1 značí vysokou míru věcné významnosti, hodnota blížící se 0 naopak minimální hodnotu věcné významnosti	– Určeno pro výsledek statistického testu z kontingenční tabulky (více než 2 × 2)

Online příloha článku

I. Pomůcka pro stanovení intervalu spolehlivosti pro korelaci v SPSS
Návod: Pomůcku stačí zkopírovat do syntax okna SPSS, na počátku je nutno zadat vaši hodnotu pro korelaci, velikost soboru a míru spolehlivosti a poté stisknout Ctrl + A a Ctrl + R, SPSS provede výpočet a zobrazí ho jak do výstupu, tak i do datové matice.
* V následujícím zadání je nutné napsat korelaci, velikost souboru a hladinu spolehlivosti (zde je zadána korelace 0,404, velikost 6920 a spolehlivost 0,95).

```
data list free / r n conflev .
begin data .
,404 6920 ,95
end data.

* Dále se provádí výpočet.

* Fisherova transformace
compute fz = .5*ln((1+r)/(1-r)).
compute sez = 1/sqrt(n-3).
compute critz = abs(idf.normal((1 - conflev)/2,0,1)).
* Dolní a horní mez pro transformovanou veličinu
compute lo_fz = fz - critz*sez .
compute hi_fz = fz + critz*sez .
* Zpětná transformace
compute lo_r = (exp(2*lo_fz) - 1)/(exp(2*lo_fz) + 1).
compute hi_r = (exp(2*hi_fz) - 1)/(exp(2*hi_fz) + 1).
formats r conflev to hi_r (f10.4) / n (f8).
list r n conflev lo_r hi_r .
```

II. Způsob načtení kontingenční tabulky do SPSS příkazem a následné analýzy

```
data list free / aktiv skolatyp w .
begin data .
1  1  15
2  1  68
3  1  112
4  1  6
1  2  2
2  2  12
3  2  21
4  2  4
1  3  4
2  3  7
```

```
3   3   16
4   3   0
1   4   19
2   4   59
3   4   70
4   4   13
1   5   14
2   5   19
3   5   16
4   5   6
end data.

wei by w.
val lab aktiv  1‘velmi nízká‘ 2‘nízká‘ 3‘průměrná‘ 4‘vysoká‘.
val lab skolatyp 1‘zš‘ 2‘gym‘ 3 ‚vícel. gym.‘ 4 ‚soš‘ 5 ‚vš‘.

BOOTSTRAP
/SAMPLING METHOD=SIMPLE
/VARIABLES INPUT=aktiv skolatyp
/CRITERIA CILEVEL=95 CITYPE=PERCENTILE  NSAMPLES=1000
/MISSING USERMISSING=EXCLUDE.
CROSSTABS
/TABLES=aktiv BY skolatyp
/FORMAT=AVALUE TABLES
/STATISTICS=CHISQ PHI
/CELLS=COUNT
/COUNT ROUND CELL.
```