Statistická významnost proti vědecké průkaznosti výsledků výzkumu

Petr Blahuš

Katedra kinantropologie, Univerzita Karlova v Praze, Fakulta tělesné výchovy a sportu

Česká kinantropologie, 2000, Vol. 4, č. 2, s. 53-72

Citace:

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

SOUHRN

Statistická významnost výsledků výzkumu je v posledních letech podrobována kritice ze strany metodologů, statistiků i samotných výzkumníků, přestože mnozí u jejího mnohdy bezmyšlenkovitého používání zůstávají. Její smysluplné použití je však omezeno jen na reprezentativní výběry pořízené metodami náhodného vybírání a na randomizované řízen experimenty. Proto se prosazuje upřednostněně používat nestatistické hodnocení velikosti rozdílu či vztahu ve výzkumných výsledcích, tzv. „size of effect", zvláště pomocí tzv. koeficientu ω^2 *jakožto podílu, resp. procenta vysvětleného rozptylu. Jednou z jeho hlavních výhod je, že nezávisí na rozsahu výběru N. Naopak jednou z hlavních nevýhod testování nulové hypotézy podle statistické teorie rozhodování je závislost statistické významnost na N, takže u velmi velkých výběrů jsou i nepatrný rozdíl nebo korelace statisticky významné a u malých výběrů i velký rozdíl či vysoká korelace statisticky nevýznamné. Prosazování používání tzv. „věcné významnosti" má už delší tradici v kinantropologických výzkumech, avšak dosud stále značný počet výzkumů končí mnohdy nesprávnou interpretací výsledků, díky slepě používané statistické významnosti, jak ji nabízí většina statistického software. V článku se kromě objasnění smyslu a správné interpretace statistické významnosti* α *navrhuje i správný postup ve výzkumu:*

1. nejprve zhodnotit věcnou významnost jak absolutně (v jednotkách měření), tak i relativně k podílu vlivu ostatních faktorů (pomocí ω^2*), a jen jde-li o randomizovaný výzkum pak*

2. použít statistickou významnost α *jakožto riziko zobecnění. Uvedení řady konkrétních příkladů z výzkumné praxe ilustruje problémy slepého používání statistické významnosti a její správné místo v metodologii konkrétních výzkumů.*

***Klíčová slova**: statistická významnost, nulová hypotéza, statistické testy, věcná-praktická významnost, interpretace výsledků, metodologie výzkumu*

ÚVOD

V posledních zhruba dvaceti letech zesílila kritika nesprávného používání statistických testů významnosti ve výsledcích výzkumu - především jako obrana proti kritice tzv. kvantitativního výzkumu ve vědách o chování, kam kinantropologie jistě patří. Tento trend se již v minulosti objevoval v několika vlnách, zhruba od 50. let, zdůrazňováním významu interpretace tzv. „praktické" významnosti oproti slepému a automatickému testování nulové hypotézy podle statistické rozhodovací teorie Neymana-Pearsona. Hlavní, protože teoreticky a matematicky podložená změna se datuje od

práce psychologa Hayse 1963, který jako protějšek k používání statistických testů na hladině rizika α = 0,05 navrhl koeficient „praktické“ významnosti ω^2. V tělovýchovném výzkumu u nás se tento trend prosazoval pod názvem „věcná“ významnost (srov. např. Čelikovský - Teplý 1966, Čelikovský etal. 1974, Kovář - Blahuš 1971, 1989 a další).

Nakonec v mezinárodní vědecké komunitě situace vykrystalizovala asi v posledních pěti letech: řada významných vědeckých časopisů a vědeckých společností prosadila nebo důrazně doporučuje nové standardy pro interpretaci výsledků kvantitativních výzkumů a pro požadavky na jejich přijetí k publikování.

Tři známé postuláty a tři otázky po významnosti výzkumných výsledků

Následující tři předpoklady o použití statistických metod ve výzkumu jsou dobře známy:

(i) Vědecká průkaznost výsledků výzkumu spočívá v logicky správném a metodologicky čistém designu výzkumu, např. v plánu experimentu.

(ii) Design výzkumu nemohou nahradit žádné dodatečné statistické, byť téměř akrobatické cviky s daty.

(iii) Vědecká průkaznost výsledků výzkumu nespočívá v jejich „statistické významnosti“.

Méně známé jsou však už odpovědi na tři otázky:

1. Jaký je vztah vědecké průkaznosti a statistické významnosti?

2. Co přesně znamená výrok „statistická významnost na hladině α = 0,05“?

3. Co znamená „size of effect“ a koeficient ω^2, které jsou požadovány předními vědeckými časopisy a společnostmi v oblasti behaviorálního, společenskovědního, pedagogického, psychologického apod. typu výzkumu?

Vědecká průkaznost vs. statistická významnost

Vědecká průkaznost výsledků výzkumu závisí na „statistické významnosti“ jedině tehdy, jestliže vědecká otázka si žádá použít nikoli deskriptivní, ale induktivní statistické usuzování, tj. tedy jen ve dvou případech:

a) Randomizační procedurou byl proveden náhodný reprezentativní výběr z tzv. opory výběru, tj. z evidence základního souboru, a cílem je provést zobecnění z výběru na základní soubor s přijatelnou chybou a přijatelným rizikem.

Příklad: Chceme vědět, zda nadpoloviční většina (>50 %) studentů FTVS je spokojena s organizací studia. Vylosujeme výběr 100 studentů FTVS z kompletního seznamu studijního oddělení, v osobním interview zjistíme, že procento spokojených studentů ve výběru je 60 %, tedy nadpoloviční. Ptáme se, jaké by bylo toto procento, kdybychom se bývali zeptali celého základního souboru všech 1.500 studentů, zda bychom pak také opravdu konstatovali, že procento spokojených je větší než 50 %.

b) Randomizační procedurou byl připraven design experimentu a cílem je zhodnotit, zda výsledky překračují náhodnost experimentu, kterou jsme tak do něj vložili.

Příklad: Při rozlosování osob na skupinu pokusnou a skupinu kontrolní je cílem zjistit, zda efekt faktoru působícího na pokusnou skupinu převyšuje nad náhodností rozlosování osob. (Zde randomizace osob do zmíněných skupin vlastně také představuje vytvoření dvou náhodných-reprezentativních výběrů.)

Co je a co není statistická významnost α

V duchu uvedených příkladů pak tvrzení, že výsledky jsou „statisticky významné“ na hladině $\alpha = 0{,}05$ má přesně následující význam.

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

Jde o riziko tzv. chyby I. druhu, že nesprávně zamítneme statistickou nulovou hypotézu H_0. Tj. zde hypotézu, že rozdíl mezi skutečným procentem spokojených v základním souboru a zadaným procentem 50% je nulový. Jinak též, že chybně zamítneme hypotézu, že rozdíl mezi hodnotou u výběru (60%) a pesimisticky předpokládanou možnou hodnotou v základním souboru (50%) je jen náhodný. Tedy chybně učiníme závěr, že z výběru lze provést zobecnění (zde, zobecnění, že v souboru studentů je počet spokojených větší než 50%).

b) U randomizovaného experimentu pak znamená riziko nejvýše 0,05, že zjištěný výsledný efekt pokusu je jen v rámci náhodnosti způsobené námi provedenou randomizací a nebyl způsoben faktorem, kterým jsme podle naší hypotézy záměrně působili. Tedy např. riziko, že výsledný rozdíl mezi výsledky pokusné a kontrolní skupiny je jen v rámci náhodnosti rozlosování osob.

Tedy oznamovací výrok, že výsledek výzkumu je „statisticky významný“ ***ne***znamená

- ani „významný“ ve smyslu vědeckého důkazu

- ani „důležitý“ nebo „podstatný“ ve smyslu laického jazyka či selského rozumu,

ke kterýmžto záměnám bohužel velmi často dochází. Důvody pro takovou záměnu jsou dvojí - jednak neznalost, jednak psychická pohodlnost, jak se o tom zmíníme níže.

Co tedy tento výrok znamená? Jeho význam je jen a jedině:

„statisticky zobecnitelný“ z reprezentativního-randomizovaného výběru na základní soubor

a to se zvoleným rizikem a tomu odpovídajícím konfidenčním intervalem v důsledku náhodné výběrové chyby. A to opravdu jen za předpokladu, že výběr byl proveden randomizační metodou vybírání z opory výběru a že je jasné na jaký přesně definovaný základní soubor chceme zobecnit.

Příklad:

Ve výše uváděném případě zjišťování spokojenosti studentů je *konfidenční interval* výběrové náhodné chyby při riziku 0,05 roven $\pm$ 9,6 %, tj. skutečné procento spokojených ze všech 1 500 studentů základního souboru leží v konfidenčním intervalu náhodné výběrové chyby **od 50,4 %** do 69,6 %.

Tedy zamítáme nulovou hypotézu H_0, že procento spokojených studentů je jen 50 %, ovšem zamítáme ji s rizikem 0,05 (!), že možná tomu tak není.

Věcná / praktická významnost a „velikost účinku“

Pojmy „size of effect“ a koeficient ω^2 vycházejí původně z přístupu, podle něhož se v některých zahraničních vědeckých časopisech v poslední době hodnotí „nestatistická“ velikost významnosti, u nás často nazývané věcná významnost.

Název „size of effect“ nebo též „effect size“ pochází z aplikací v kontrolovaných experimentech, kde se ve výzkumu nejvíce projevoval rozpor mezi statistickou významností, jakožto zobecnitelností, a skutečnou „podstatnou velikostí experimentálního účinku“.

Význam termínu „size of effect“ se pak analogicky přenáší i na neexperimentální, ale jen popisné induktivní studie zobecňující z reprezentativních výběrů na základní soubory.

Jinde se v literatuře též setkáváme s ekvivalentními termíny, např. významnost

praktická - practical significance

logická - logical

„podstatnost“ - substantive significance

„výsledková důležitost“ - result importance

„výsledková smysluplnost“ - result meaningfulness

Doposud výzkumníci hodnotili věcnou významnost, tj. „size of effect“, výhradně v naměřených jednotkách, např. rozdíl průměrů v cm, sekundách, bodech škál atp., což je ovšem i nadále zásadně nutné. Nyní se však užívá i několik dalších statistických indexů, z nichž většina je založena na obdobném principu jako koeficient ω^2, který byl již v r. 1963 navržen psychologem Haysem. Stručně a zjednodušeně řečeno: koeficient ω^2 = podíl „vysvětleného“ rozptylu.

Jde o jistou analogii druhé mocniny korelačního koeficientu závisle proměnné Y na nezávislé proměnné X, dobře známý tzv. koeficient determinace r^2, obvykle převedený na r^2.100 %, tj. procento rozptylu Y „vysvětlené“ (či lépe řečeno „odhadnutelné“) korelační-regresní závislostí na X.

Tedy ω^2 je číslo mezi 0 a 1. Např. je-li v experimentu ω^2 = 0,10, znamená to, že experimentální efekt lze z 10 % procent připsat vlivu záměrného působení experimentálním faktorem a 90% jiným, zpravidla neznámým vlivům ap.

Dvojí zhodnocení věcné významnosti

Hodnocení věcné významnosti je nutno provádět dvojím způsobem současně:

- určit předem minimální hodnotu velikosti v jednotkách měření (cm, sekundy ap.), kterou budeme považovat za podporu naší hypotézy, tj. posouzení věcné významnosti přímo a v absolutní reálné velikosti,

- určit předem minimální vysvětlené procento rozptylu, resp. jeho podíl, tj. koeficient ω^2, jež budeme považovat za obsahově podstatné v relaci k ostatním nesledovaným vlivům, přičemž je i při intuitivním pohledu zřejmé, že tento podíl (alespoň v rámci deskriptivní statistiky daného výběru) nezávisí na velikosti rozsahu výběru N, tj. jde o posouzení věcné významnosti relativně k významnosti vlivu ostatních faktorů. Je záležitostí vědecké etiky tyto dvě hodnoty formulovat předem, na základě podrobné znalosti teorie o problému a jako součást důkladného zdůvodnění hypotézy. Tak se lze vyhnout případným námitkám, že dodatečně přijaté zdůvodnění bylo spekulativní, nebo že přání výzkumníka vedlo k oportunisticky přizpůsobené volbě věcné významnosti tak, aby se spíše podpořilo jeho zbožné přání.

Reprezentativnost a rozsah výběru „N“ - vliv na statistickou významnost

Definice ze statistické teorie výběru:

Výběr je reprezentativní, jestliže výběrová procedura zajistila, že každý prvek základního souboru měl stejnou pravděpodobnost, aby byl vybrán. To zajišťují jedině randomizační - znáhodňovací metody: dobře promíchané losování, tabulky náhodných čísel, randomizace na počítači, vícestupňový znáhodněný výběr atp. Např. tzv. záměrný kvótní výběr není reprezentativní, jeho chybu a riziko nelze statisticky vypočítat, musí je odhadnout expert, který kvóty *expertně* záměrně předepsal. Reprezentativní soubor tedy není takový, o kterém si někdo myslí, že by mohl *něco* dost dobře reprezentovat (a navíc dokonce ani nedokáže přesně definovat *co?* - jaký konkrétní základní soubor?).

Známé, školácky triviální příklady říkají, že náhodný výběr není ten, který vznikl tak, že se do něj pokusné osoby „náhodou" dobrovolně přihlásily. Nebo že znáhodnění mezi kontrolní a pokusnou skupinou není, když pokusnou skupinou je třída žáků, jejíž třídní učitelka „náhodou" má zájem v experimentu „vědecky dokázat", že její didaktická metoda je lepší ap.

Někdy se také zapomíná na základní vlastnosti statisticky reprezentativního výběru:

- není-li výběr reprezentativní, tj. není-li pořízen znáhodňovací procedurou, pak ovšem příslušné matematicky odvozené *vzorečky neplatí*, byly totiž odvozeny z postulátů náhodnosti, a vypočtené riziko α je *falešné*, stejně tak jako interval konfidence

- „reprezentativnost" není odstupňovaná vlastnost, výběr nemůže být „více reprezentativnější" nebo „méně reprezentativnější" - především *větší* výběr není „reprezentativnější"

- reprezentativnost výběru totiž nezávisí na rozsahu výběru N :

- i výběr při N = 3 může být reprezentativní, má ovšem velkou výběrovou chybu, tj. konfidenční interval a velké riziko α, oboje však umíme vypočítat i výběr při N = 1 000 000 může být nereprezentativní, ale jeho chybu a riziko α nemůžeme vypočítat, protože pro něj příslušné vzorce neplatí a pokud je nesprávně použijeme, pak vypočítáme nesmyslné hodnoty

Statistická zobecnitelnost a velikost N

Je-li výběr reprezentativní, pak statistická významnost, tj. zobecnitelnost velmi silně závisí na rozsahu výběru N. Tak např. - kdyby vylosovaných studentů bylo jen o 8 méně, tj. při N = 92, pak by při riziku α=0,05 byl konfidenční interval od 49,009 % do 70,001 %, a nulovou hypotézu, že 50 % do něj spadá, nemůžeme zamítnout. Naopak je-li rozah výběru velmi velký, pak i sebemenší rozdíl je statisticky zobecnitelný.

Jak uvádějí mnozí autoři, ať z hlediska filozofie vědy či z hlediska běžné praxe je obtížné předpokládat, že mezi vlastnostmi, které výzkumem sledujeme, by nebyl absolutně žádný vztah, tedy, že by nulová hypotéza v populaci platila přesně. Spíše je tomu tak, že vlastnosti a jevy jsou zprostředkovaně přes síť jiných vztahů alespoň v nějaké nepatrné souvislosti, a při dostatečně velkém až obrovském počtu pozorování N pak tuto velmi slabou a prakticky bezvýznamnou závislost musíme nutně objevit.

Oba extrémy - obrovské N i velmi malé N - pak vedou k problematickým a často i protismyslným důsledkům. Uvedeme některé prototypy možných odstrašujících příkladů.

A) Příliš velké N

V tomto případě i věcně zcela bezvýznamné rozdíly či vztahy jsou zobecnitelné:

A 1) Změřeno N = 80 000 branců v ČSSR v r. 1971 a další vlna v r. 1972, rozdíl v běhu na 100 m vykazoval „zhoršení“ průměrů o 0,000 3, tj. tři desetitisíciny sekundy. Rozdíl byl pomocí Studentova t-testu shledán jako vysoce statisticky „významný“ s rizikem menším než α=0,001 a interpretován jako „vysoce významné zhoršování zdatnosti branců“.

Otázka na okraj: Na jaký základní soubor se mělo z tohoto výzkumu zobecnit s rizikem α=0,001, když byli změřeni všichni branci? Jednalo se tedy nikoli o „výběr“ ale o tzv. vyčerpávající statistické šetření, tj. byl změřen základní soubor sám. Výsledky u něj zjištěné pro něj nutně tedy platí, není na co zobecňovat a ptát se po riziku α statistického zobecnění je holý nesmysl.

A 2) Na universitách USA provedeno šetření, kolik času týdně věnují domácímu studiu studenti bílí oproti černým, rozsah výběru N = 60 000. Celková průměrná doba studia byla přes 5 hodin týdně. Přitom rozdíl 12 sekund týdně mezi černými a bílými studenty byl statisticky významný a proto interpretován jako rasová diskriminace.

A 3) Korelace - korelační koeficient **r** je při velkém N statisticky významný, i když je u výběru téměř nulový.

Např. na hladině rizika zobecnění α=0,05 jsou „významné“ hodnoty **r** při rostoucím N:

N	**r**
3	0,99
10	0,67
100	0,20
1 000	0,06
10 000	0,02
20 000	0,01

V základním souboru může být sice korelace různá od nuly, ale přitom rovna jen r = 0,01 - jakou vědeckou nebo praktickou významnost má takto slabý vzájemný vztah?

A 4) Přitom je třeba si uvědomit, že „statisticky významná korelace“ znamená: taková korelace u výběru, že v základním souboru není nulová, tj. není přesně (!) rovna nule, tj. neplatí r = 0, tedy laicky zdůrazněno, že není nulová na nekonečný počet desetinných míst. Nepochopení tohoto faktu pak může vést k podivuhodným desinterpretacím.

A 5) Např. při ověřování spolehlivosti neboli reliability diagnostické metody - diagnostického testu - byla provedena korelace prvého a opakovaného měření, tzv. test-retest stabilita, a zjištěn tzv. korelační koeficient stability. Ten byl při daném N = 100 „vysoce statisticky významný“, a to na hladině α = 0,01. To pak bylo interpretováno tak, že uvedená diagnostická metoda je „vysoce významně spolehlivá“. Přitom faktická hodnota korelačního koeficientu u daného výběru osob byla r = 0,28. Správná interpretace by měla být, že (s rizikem 1%) v základním souboru není spolehlivost dané diagnostické metody přesně nulová - a jistě, že ne každou, nepatrně se od nuly lišící spolehlivost bychom považovali za výraz kvality nějaké diagnostické metody.

B) Příliš malé N

Je-li rozsah reprezentativního výběru malý, pak ani sebevýraznější *věcně* významné výsledky nejsou statisticky zobecnitelné

B 1) Hypotetický příklad:

V posledním stadiu AIDS se novým lékem podařilo zcela uzdravit 50% pacientů, kteří již umírali. Pro malý počet N výsledek nebyl statisticky významný na tradiční hladině α=0,05, lze to interpretovat tak, že nový lék působí jen „nevýznamně“?

B 2) Reálný příklad:

U československého reprezentačního družstva běžců N = 5 byla zjištěna pořadová (Spearmanova) korelace mezi nejlepším osobním výkonem v závodním období a laboratorním testem trénovanosti r = 0,89. Tuto korelaci označil počítač správně jako statisticky nevýznamnou na hladině rizika α=0,05. To vyvolalo pochyby trenéra i metodika, zvláště vzhledem k tomu, že pouhým okem bylo jasně vidět, že pořadí výkonů a pořadí v testu souhlasí až na jedinou výjimku:

Závodník:	A	B	C	D	E	
X pořadí podle testu:	1.	2.	3.	4.	5.	korelace obou pořadí
Y pořadí podle výkonu:	1.	2.	3.	5.	4.	r = 0,89

K tomu snad jen dvě otázky pro čtenáře.

Otázka 1: Znamená to, že laboratorní test je pro kontrolu trénovanosti „nevýznamný“, tedy nevhodný?

Otázka 2: Z jakého základního souboru představuje družstvo nejlepších běžců republiky „reprezentativní“ a „náhodný“ výběr, tj. na jakou populaci se má zobecnit? (Porovnejme se situací vyčerpávajícího souboru branců v příkladu A1).

B 3) U reprezentanta ve skoku dalekém, konkrétně Gombaly, byla při jeho N=14 různých skocích zjištěna korelace r = 0,49 mezi rychlostí běhu těsně před odrazem a mezi odpovídající délkou skoku. Vzhledem k malému N byla tato korelace statisticky nevýznamná i na nepříliš přísné hladině rizika α=0,05.(Povšimněte si: Hodnota, která by už byla významná, je 0,497***!***). Byl učiněn závěr, že specifickým rysem Gombalovy techniky je, že u něj „není významný vztah“ mezi rychlostí rozběhu a délkou skoku.

Otázka: Znamená „statisticky nevýznamný vztah“ mezi rychlostí před odrazem a délkou skoku, že Gombala by mohl skákat z místa?

Poznámka: Po výpočtu regresní chyby odhadu z této „nevýznamné“ korelace 0,49 se ukázalo, že jednoduchá lineární regresní rovnice umožňuje u Gombaly předpovídat délku skoku s chybou odhadu $\pm$15 cm ze změřené rychlosti jeho běhu před odrazem.

Nezávislost koeficientu ω^2 **na velikosti N**

Jde o jednu z nejdůležitějších vlastností tohoto ukazatele věcné významnosti. Tato nezávislost je zřejmá jak z povahy koeficientu - podíl vysvětleného rozptylu - tak i z matematických výrazů pro výpočet koeficientu ω^2 na základě jiných statistických charakteristik, které původně principiálně na rozsahu N právě závisejí. Tak např. hodnotu Studentova **t**-testu dvou nezávislých výběrových průměrů je možno přepočítat na koeficient ω^2 velmi jednoduše:

$$\omega^2 = (t^2 -1) / (t^2 - N-1)$$

(kde N = $n_1 + n_2$ je součet počtů obou skupin). Tento vzoreček pro relativní věcnou významnost je dokonce uváděn i v příručce pro tělovýchovný výzkum - Thomas-Nelson „Research methods in physical activity“, 1990, s. 133.

Koeficient ω^2, jako ukazatel relativní věcné významnosti nezávislé na různých rozsazích výběrů, N, má také zásadní význam pro srovnávání výsledků z různých výzkumných studií. Tedy také tam, kde jde o syntézu řady výsledků z mnoha výzkumů, článků atp., v tzv. meta-analytických studiích. Tam není možné jen konstatovat statistickou významnost anebo nevýznamnost, nebo porovnávat hodnoty t-testů aj., právě pro neporovnatelnost různých výzkumů s různými rozsahy výběrů N.

ZDŮVODNĚNÍ A VOLBA JINÉ HLADINY RIZIKA ZOBECNĚNÍ NEŽ $\alpha = 0,05$

Je-li cílem výzkumné studie statistické zobecnění z randomizovaného výběru na základní soubor, pak je problematické a mnohdy až nesmyslné slepé používání stejného rizika, bez ohledu na povahu problému, tedy všude $\alpha =$ 0,05 nebo α=0,01.

Jistě, že při testování kvality padáků od výrobce bude i hladina rizika 0,001, tj. jedno promile, nepřijatelně vysoká. Naproti tomu většina výzkumů, které ověřují pozitivní účinek nějaké nové pedagogické metody, může být zcela jistě ověřována a zobecňována i s rizikem větším než α=0,10 nebo i α=0,20.

Příklad:

Na reprezentativním výběru žáků základních škol ČR byla ověřena nová didaktická metoda výuky plavání, která dosavadní obvyklý, školními osnovami předpokládaný počet plaveckých lekcí zkracuje velmi podstatně z hlediska věcné významnosti. Toto zkrácení není „bohužel“ statisticky významné na hladině rizika zobecnění $\alpha =$ 0,05, je ale statisticky významné neboli zobecnitelné s rizikem α = 0,15. Pokud tato didaktická metoda neskrývá nějaké problémy v bezpečnosti nebo nemá neúměrné ekonomické nároky, pak ji jistě můžeme vřele doporučit všem tělovýchovným pedagogům k širokému použití - i s rizikem 15%, že tato metodika nebude časově o mnoho efektivnější.

Tedy, jak ostatně praví většina učebnic aplikované statistiky, je třeba se zásadně vyhýbat automatickému používání hladiny rizika $\alpha =$ 0,05, avšak toto riziko je třeba zdůvodněně zvolit předem, tedy před vlastním výzkumem.

Pečlivá a teoreticky i prakticky zdůvodněná volba správné hladiny rizika se tedy má provádět před vlastním výzkumem jako součást zdůvodnění a formulace jednotlivé dílčí pracovní hypotézy, aby se předešlo dodatečným oportunistickým ústupkům na přísnost posouzení hypotézy.

Vědecké společnosti a časopisy vyžadující nebo doporučující uvádění věcné významnosti - „size of effect“ ω^2

Hodnocení věcné významnosti pomocí ω^2 se v posledních letech běžně uvádí v článcích publikovaném v časopise, který je vlajkovou lodí vědeckých periodik v našem oboru, *Research Quarterly for Exercise and Sport*. Řada dalších vědeckých časopisů a vědeckých společností pak je buď přímo vyžaduje jako podmínku pro přijetí článku k publikaci anebo je důrazně doporučuje.

APA -American Psychological Association ve svém manuálu pro publikování ve všech svých časopisech doporučuje autorům uvést zhodnocení „size of effect“: „Žádný z těchto dvou typů pravděpodobnostních údajů [tj. rizik α a β] nevyjadřuje důležitost nebo velikost účinku, protože oba závisejí na rozsahu výběru... Proto se doporučuje poskytnout informaci o velikosti ´effect-size´.“ (APA 1994, s. 18).

Některé z jejích časopisů však to uvádějí jako nutný požadavek pro přijetí rukopisu.

AERA - American Educational Research Association v současné době projednává, aby doporučení bylo přeměněno na požadavek, který také zatím platí jen u některých jejích časopisů.

AAC - Association for Assessment in Counseling. to již pro své časopisy a konference formulovala jako striktní požadavek.

Jako nutný požadavek se uvedení zhodnocení věcné významnosti - „size of effect“ pro přijetí článku k publikaci vyskytuje např. u následujících časopisů:

Journal of Experimental Education

Educational Researcher

Educational and Psychological Measurement

Measurement and Evaluation in Counseling and Development.

A jistě jsou i další. Stojí za úvahu, že alespoň v podobě doporučení by tato strategie při interpretaci výzkumných výsledků měla být prosazována i do našich časopisů *Acta Universitatis Carolinae Kinanthropologica* a také *Česká kinantropologie*.

Důvody, které vedou k slepému používání statistické „významnosti“

Jak již bylo řečeno, důvody jsou prosté: neznalost a pohodlnost. Řada metodologů a statistiků má pro tento tristní stav věci různé a také zábavné metafory. Podle Federera 1973 je to pštrosí strategie „strkání hlavy do písku“. Podle Cohena 1994 je to prostě tak rozšířený úzus a zlozvyk, že pro některé výzkumníky by i hypotézu o tom zda je Země kulatá bylo třeba otestovat alespoň na hladině 0,05.

Na výroční konferenci APA 1996 bylo uspořádáno samostatné sympózium na téma „zakázat používání statistické významnosti ve všech časopisech“ této vědecké společnosti. Někteří autoři vedou se statistickou významností metaforický soudní proces s obviňováním nešťastných následků jejího nadužívání, jiní hovoří o „vymýtání ďábla, nulové hypotézy“.

Thompson 1996 říká, že výzkumník používá test významnosti pro kontrolu své únavy se sbíráním dat, tj. aby zjistil, zda pro podporu hypotézy - coby svého zbožného přání - již změřil dost osob N, kdy mu už statistický test vyjde konečně statisticky významný a on už nebude muset pokračovat v měření dalších osob.

Snad nejlepším, i když tragikomickým vysvětlením je Thompsonovo 1996 (s.28):

„Řečeno jednoduše, u mnoha výzkumníků je jejich touha využít matematický výpočet pravděpodobnosti [α] pouze atavistickým únikem od existenciální lidské zodpovědnosti v duchu existenciálního filosofa Fromma a jeho *´Úniku před svobodou´*. Ale, bohužel,..., zda velikost rozdílu mezi průměrem skupiny A a průměrem skupiny B má nějakou praktickou důležitost je otázka, která nemůže být zodpovězena pomocí statistického testu. Je to otázka, kterou může výzkumník zodpovědět po zvážení informací nestatistické povahy.“

(Závorka a podtržení od Blahuše.)

Závěry a doporučení

1. Používat statistickou významnost neboli zobecnitelnost jen tam, kde je cílem zobecnit z přísně randomizovaného-reprezentativního výběru na jasně definovaný základní soubor, nebo u randomizovaných experimentů.

2. Před vlastním výzkumem na základě teorie a zdůvodnění hypotéz stanovit předem věcnou významnost podporující hypotézu, a to jak

a) absolutně - v jednotkách měření (cm, kg, body na škále ap.), tak i

b) relativně - v procentech vysvětleného rozptylu, např. koeficientem ω^2

3. Je- li cílem zobecnění splňující podmínky sub 1., pak také předem stanovit i hladinu rizika α statistického zobecnění na konkrétně specifikovaný základní soubor. Přitom se zásadně vyhýbat tradičním klišé α=0,05, α=0,01, která obvykla shodně nabízejí jak obsolentní tabulky, tak i moderní software. Zdůvodnit předem, proč zobecnění je vyhovující i např. při riziku α=0,15, nebo proč je nutné použít extrémně nízkou hladinu rizika zobecnění, např. α=0,001.

4. Vždy nejprve posoudit velikost věcné významnosti - „size of effect“. Teprve pak, má-li tato smysl, se ptát, zda ji lze zobecnit - tj. zda je významná i statisticky, tedy zobecnitelná. (Není jasné, jakou logickou úvahou by bylo možné zdůvodnit opačné pořadí.)

5. Dodržovat terminologickou přesnost a odlišení pojmu „statistická významnost“ a „velikost věcné - praktické významnosti“. Vyhýbat se neurčitým a obojetným výrazům typu „rozdíl průměrů byl významný“ bez rozlišení, který druh významnosti se má na mysli.

6. V souhlasu s doporučením AERA podporovat vědeckou průkaznost spíše opakováním výzkumné studie a srovnávat její výsledky s jinými prostřednictvím meta-analýz, než provádět jednorázová rozsáhlá šetření u nereprezentativních výběrů.

7. V duchu osnov vyučovacího předmětu „Metodologie diplomové práce“, zavedeného na FTVS od r. 1998/99 (Blahuš-Hendl 1998), vyžadovat výše uvedený postup od diplomantů i doktorandů, doporučit jej Radou pro doktorské studium kinantropologie školitelům.

8. Po vzoru výše uvedených časopisů zařadit alespoň doporučení do pokynů pro autory v časopisech *Acta Universitatis Carolinae Kinanthropologica* a také *Česká kinantropologie*, aby uváděli interpretaci věcné významnosti svých výsledků.

LITERATURA

- *AMERICAN PSYCHOLOGICAL ASSOCIATION - APA (1994). Publication manual of the American Psychological Association, 4th edition. Author, Washington DC.*
- *ASSOCIATION FOR ASSESSMENT IN COUNSELING - AAC. (1994). Guidelines for authors. Measurement and Evaluation in Counceling and Development, 27, 1: 341.*
- *BLAHUŠ, P. (1997). K úloze tzžv. kvantitativních metod v kinantropologii. Česká kinantropologie 1, 1: 7-18.*
- *BLAHUŠ, P. - HENDL, J. (1998). Nový vyučovací předmět “Metodologie diplomové práce“. Sdělení na semináři FTVS Praha - Společensko-vědní problémy kinantropologie, 25. 11. 1998. (V tisku v připravovaném sborníku*
- *CARVER, R. (1978). The case against statistical significance testing. Harvard Educational Review, 48: 378-399.*
- *CARVER, R. (1994). The case against statistical significance testing, revisited. Harvard Educational Review 61: 287-292.*

- *COHEN, J. (1994). The earth is round, p $<$ .05 . American Psychologist, 49:997-1003.*
- *ČELIKOVSKÝ, S. - TEPLÝ, Z. (1967). Empirické metody výzkumu v tělesné výchově. (Učební texty ČSTV.) Sportovní a turistické nakladatelství. Praha.*
- *ČELIKOVSKÝ, S. etal. (1974). Seminář o použití a interpretaci statistických metod v tělovýchovném výzkumu. Metodický dopis 2/1974 ČÚV ČSTV, Sportpropag, Praha.*
- *DANIEL, W.W. (1977). Statistical significance versus practical significance. Science Education, 61: 423-427.*
- *FEDERER, W.T. (1973). Statistics and society - data collection and interpretation. Dekker, N. York.*
- *GAY, L.R. (1976a). Educational research - competencies for analysis and application. Merrill Publishers, Columbus (OH).*
- *GAY, L.R. (1976b). Professional supplement - educational research competencies for analysis and application. Merrill Publishers, Columbus (OH).*
- *HELDREF FOUNDATION. (1997). Guidelines for contributors to the Journal of Experimental Education. Journal of Experimental Education. Vol. 65.*
- *KIRK, R.E. (1996). Practical significance: a concept whose time has come. Educational and Psychological Measurement, 56, 5: 746-759.*
- *KOVÁŘ, R.- BLAHUŠ, P. (1971). Vybrané statistické metody v antropomotorice. Univerzita Karlova, Praha. (Skripta)*
- *KOVÁŘ, R.- BLAHUŠ, P. (1989). Aplikace vybraných statistických metod v antropomotorice. Univerzita Karlova, Praha. (Skripta)*
- *LEVIN, J.R. (1996). Discussant of significance tests: should they be banned from APA journals ? Symposium at APA annual meeting, American Psychological Association, Toronto.*
- *ROBINSON, D.H. - LEVIN, J.R. (1997). Reflections on statistical and substantive significance with a slice of replication. Educational Researcher, 26, 5: 21-27.*
- *ROZEBOOM,W.W. (1960). The fallacy of the null hypothesis significance test. Psychological Bulletin, 57: 416-428.*
- *SHAVER, J. (1993). What statistical significance is, and what it is not. Journal of Experimental Education, 61,4: 293-316.*
- *SCHMIDT, F. (1996). Statistical significance testing: implications for the training of researchers. Psychological Methods, 1,2: 115-129.*
- *THOMAS, J.R.- NELSON, J.K. (1990). Research methods in physical activity. Human Kinetics, NJ, 2.vyd. (3. vyd. 1996)*
- *THOMPSON, B.(1993). Statistical significance, result importance, result generalizability. Measurement and Evaluation in Counceling and Development, 22: 2-6.*
- *THOMPSON, B. (1994a). The pivotal role of replication in psychological research: empirically evaluating the replicability of sample results. Journal of Personality, 62, 2: 157-176.*
- *THOMPSON, B. (1994b). Guidelines for authors. Educational and Psychological Measurement, 54, 4: 837-847.*
- *THOMPSON, B. (1996). AERA editorial policies regarding statistical significance testing: three suggested reforms. Educational Researcher, 25, 2: 26-30.*
- *THOMPSON, B. (1997). Editorial policies regarding statistical significance tests: further comments. Educational Researcher, 26, 5: 29-32.*

- *THOMPSON, B. - SNYDER, P.A. (1997). Statistical significance testing practices in the Journal of Experimental Education. AERA - American Educational Research Association Annual Conference, AERA - ERIC No. ED, Chicago.*
- *WANG, Ch. (1993). Sense and nonse of statistical inference. Dekker, N. York.*