Nesprávná užívání statistické významnosti a jejich možná řešení*

Petr Soukup**
Institut sociologických studií
Fakulta sociálních věd, Univerzita Karlova v Praze

Improper Use of Statistical Significance and Possible Solutions

Abstract: *The use of significance tests in social sciences is widespread mainly due to simple computation via statistical packages. Unfortunately the more social scientists use statistical significance estimates for making causal inferences the less they appear to understand about this influential concept. Statistical modelling results are usually presented in terms of their statistical significance and little other information is provided. The goal of this article is to show the limits of using statistical significance as a sole means of making inferences; and to present alternative statistical fit indicators readily available within frequentist approach to statistics: confidence intervals, minimum sample size and power analysis. Multiple working hypotheses are also explored together with two well known information criteria – AIC and BIC. This article provides practical information on how to undertake valid and reliable statistical analyses of social science data.*

Data a výzkum - SDA Info 2010, Vol. 4, No. 2: 77-104.

* Tento článek byl podpořen ze Specifického vysokoškolského výzkumu UK FSV pro rok 2010 pod č. 261 510 - Analýza současných trendů sociálního vývoje v České republice a jejich reflexe v odborném diskurzu. Článek věnuji památce na nedávno zesnulého profesora Petra Blahuše († 2010), který dlouhodobě kultivoval úroveň používání statistiky v ČR a jako první výrazněji upozornil na problémy popisované v tomto článku.
** Veškerou korespondenci posílejte na adresu: Petr Soukup, Institut sociologických studií, Fakulta sociálních věd UK, U Kříže 8, 158 00 Praha 5 – Jinonice; e-mail: soukup@fsv.cuni.cz.

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

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

Úvod

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

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

I. Statistická významnost a její slabiny

Stručná historie

Statistická významnost resp. testování nulové hypotézy (v angličtině null hypothesis statistical testing nebo jen NHST) je velmi staré. Původním autorem myšlenky je zřejmě John Arbuthnott [1710], který svým testem mínil prokázat existenci boží prozřetelnosti. Dnešní vědci si kladou cíle méně smělé[1] a využívají koncepce, kterou rozpracovali statistici ve 20. a 30. letech minulého století, zejména Ronald Fisher, Jerzy Neyman a Egon Pearson. Jejich od-

1 Arbuthnott se snažil prokázat boží prozřetelnost skrze odhalování zákona vyrovnávajícího počet narozených mužů a žen a prokazoval také nepřirozenost polygamie z tohoto zákona plynoucí. Statistická významnost je tedy de facto původně sociálněvědním konceptem, ač by to zřejmě dnešní matematici či statistici nepřiznali.

kaz dodnes využívají statistici a vědci ve všech empirických oborech. Mnozí pokládají tento koncept za vynikající, naopak lze nalézt i odpůrce, kteří navrhují vymícení statistické významnosti z vědy. Detailnější pozornost koncepci Ronalda Fischera resp. Neymanovo-Pearsonovskému pohledu je věnována v dalším textu.

Definice statistické významnosti

Nejdříve si připomeňme postup testování hypotéz, který ukazuje na definici statistické významnosti. Začněme definicí pojmu statistické hypotézy: **hypotézou rozumíme tvrzení o rozdělení pozorované náhodné veličiny**[2] (např. o rozdělení nějaké statistiky[3] náhodného výběru). Důležitý zvláštní případ nastává, je-li rozdělení výběrové statistiky známé: v takovém případě lze **hypotézu formulovat přímo jako tvrzení o hodnotě parametru příslušného rozdělení**[4],[5]. Je nutné zdůraznit, že **hypotéza se týká celého základního souboru**, z něhož jsme vybírali nebo který experimentálně zkoumáme (např. všech dospělých osob v Česku), **zatímco její testování se odehrává pouze na vybraných jedincích**, které jsme skutečně zkoumali. **Smyslem testování je správně zobecnit z vybrané podmnožiny (výběru) na celek**.

Klasicky formulujeme vždy dvě hypotézy o situaci v základním souboru: takzvanou **nulovou** (H_0) a k ní opačnou **alternativní** (H_1) hypotézu. Nulová hypotéza se tak nazývá proto, že obvykle tvrdí, že proměnné v populaci na sobě nezávisí nebo že mezi skupinami v populaci nejsou rozdíly (v průměrech, mediánech apod.). Předpokládá tedy nulovost efektu vyvolaného nějakým zásahem (např. žádný rozdíl mezi skupinami experimentálních objektů, s nimiž se různě zachází). Takto se totiž původně objevila v dílech sira Ronalda Fishera o vyhodnocování biologických experimentů (viz dále). Alternativní hypotéza oproti nulové hypotéze naopak tvrdí, že existuje nějaký (zobecnitelný) rozdíl mezi sledovanými skupinami, případně že dva fenomény spolu souvisí. Testování statistických hypotéz provádíme tak, že z výběrových dat je vypočtena testová statistika a na základě porovnání s kvantily rozdělení této statistiky (za předpokladu platnosti nulové hypotézy) se zjistí, zda je na dané hladině spolehlivosti možno nulovou hypotézu zamítnout.

Koncept testování byl původně vyvinut pro pravé experimentální uspořádání, dnes však je hojně využíván i pro data pocházející z kvaziexperimentů

2 Tato definice vychází z definic běžně uváděných v učebnici statistiky.

3 Statistikou může být např. výběrový průměr, výběrový rozptyl, proporce zkoumaného jevu ve výběru apod.

4 Například lze předpokládat to (tj. utvořit hypotézu o tom), že průměrný příjem je 23 tisíc Kč, politická strana má podporu 25 % voličů.

5 Ve statistické teorii se odlišují hypotézy jednoduché a složené (platí jak pro nulovou, tak alternativní), alternativní hypotézy pak mohou být jednostranné i dvoustranné, v detailech odkazujeme na Anděl [2003, 2005].

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

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

Poznámka: V tabulce užíváme nejběžnějšího značení, symbolu H0 pro nulovou hypotézu a H1 pro hypotézu alternativní.

(není zaručeno náhodné zařazení jedinců do skupin podrobených různým experimentálním „ošetřením“ – například když se účastníci mohou sami rozhodnout, zda se experimentu podrobí) a pseudoexperimentů (neprovádí se „ošetření“, nelze tedy použít kontrolní skupiny – například standardní jednorázové sociologické výzkumy). Nejenže pak není úplně vhodná původní terminologie, ale i při interpretaci výsledků je potřeba uvážit odlišné okolnosti vzniku dat.

Interpretace statistické významnosti

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

Jak je vidět z tabulky, rozhodnutí testu o hypotéze nemusí být vždy v pořádku. K chybě prvního druhu dochází, když je nulová hypotéza zamítnuta, přestože H_0 platí. Obdobně chyba druhého druhu nastává, když nulová hypotéza zamítnuta není, přestože neplatí. Kvalita testu je dána pravděpodobnostmi, s jakými tyto chyby mohou nastat (α a β v tabulce 1). Platí, že pro daný výběrový soubor obvykle nelze současně minimalizovat pravděpodobnosti obou druhů chyb. Z tohoto důvodu se statistici rozhodli omezit riziko chyby prvního druhu na rozumnou velikost, nejčastěji na 5 % (α = 0,05); tuto hodnotu zavedl do statistického diskurzu Ronald Fisher [1925, 1926] (více v další části „Nedostatky“ statistické významnosti). Zamítání nulové hypotézy se tedy děje nejčastěji s 5% rizikem, tj. stanovujeme pravděpodobnost zamítání nulové hypotézy při její platnosti v základním souboru na maximální hodnotu 0,05. Protože chybu druhého druhu nemáme jasně pod kontrolou, volíme v případě, že nedokážeme na základě hodnoty testové statistiky zamítnout nulovou hypotézu, opatrný závěr: „nezamítáme H_0“ namísto závěru „zamítáme H_1 a přijímáme H_0“. Dnešní počítače zjistí pro příslušný test a náš datový soubor pravděpodobnost chyby prvního druhu (tzv. Sig., P-value apod.) a tu srovnáváme s klasickou hodnotou 0,05 a pro zamítnutí H_0 (a přijetí H_1) požadujeme, aby vypočtená hodnota byla nižší. Takto vypočtená pravděpodobnost chyby prvního druhu je právě statistická významnost. Jinak řečeno **statistická významnost je pravděpodobnost, s jakou bychom – za předpokladu pravdivosti nulové hypotézy – mohli obdržet data odporující nulové hypotéze stejně či ještě více než po-**

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

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

Zdroj: ISSP 2007, n = 1 210. Poznámka: Uvedena jsou sloupcová procenta. Výpočet byl proveden v systému SPSS.

zorovaná data. [Euromise, kapitola 7; obdobně Zvára, Štěpán 2002: 167]. Jde tedy o podmíněnou pravděpodobnost získání dat s testovou statistikou stejnou, jako je naše, nebo „horší" při platnosti nulové hypotézy v základním souboru **P (*D* / H_0) a nikoliv** o pravděpodobnost platnosti nulové hypotézy při existenci našich dat **P (H_0 / *D*)** [Cohen 1994, Loftus 1996]. Naše uvažování je běžně řízeno následující logikou: je-li statistická významnost nízká (většinou menší než 5 %), nulová hypotéza pro základní soubor nejspíš neplatí[6], protože získat náš výběr z takového základního souboru je velmi nepravděpodobné (ale ne nemožné!).

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

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

V příkladu vycházíme z dat získaných v České republice v rámci výzkumu ISSP 2007, který byl tematicky zaměřen na volný čas. Respondenty byly osoby starší 17 let. Pro popis jednoho z výsledků získaných ve výběrovém souboru (čítajícím 1 222 respondentů, z toho 545 mužů a 677 žen) byla zvolena kontingenční tabulka zobrazující frekvenci konzumace alkoholu u mužů a žen (viz tabulka 2).

Z popisné statistky (sloupcových procent) plyne, že v námi sledovaném výběru platí, že muži pijí alkohol častěji než ženy. Za pomoci postupů vycházejících z testování statistických hypotéz zkusíme vyřešit otázku, zda je tento závěr možné zobecnit na celou dospělou populaci ČR[7]. Konkrétně

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

7 S ohledem na to, že naše data jsou pouze z výběru, nelze automaticky závěry zjištěné ve výběru zobecňovat na populaci. Důvodem, proč toto zobecnění nelze přímo

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

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

Zdroj: ISSP 2007, n = 1 210.

užijeme chi-kvadrát test o nezávislosti znaků[8]. Nulová hypotéza (H_0) tohoto testu tvrdí, že mezi zkoumanými proměnnými (v našem případě frekvence konzumace alkoholu a pohlaví) neexistuje v základním souboru (dospělé populaci) souvislost. V našem případě lze formulaci nulové hypotézy upravit, neexistence souvislosti znamená věcně, že muži i ženy v ČR konzumují alkohol stejně často (tedy rozložení frekvence konzumace alkoholu u mužů i žen je v populaci totožné). Alternativní hypotéza (H_1) naopak tvrdí, že zkoumané veličiny spolu v základním souboru souvisí, v našem konkrétním případě lze interpretovat souvislost v tom smyslu, že rozložení frekvence konzumace alkoholů u mužů a žen je odlišná. Výsledek statistického testování uvedených hypotéz shrnuje tabulka 3.

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

---

provádět, je existence výběrové chyby. I když tuto chybu pro naše data neznáme, lze ji za pomoci známých vzorců odhadnout.

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

9 Stupně volnosti pro uvedený test jsou dány součinem počtu řádků a sloupců vždy zmenšeným o jednotku u příslušné kontingenční tabulky. V našem případě máme pět kategorií frekvence konzumace alkoholu a dvě kategorie pohlaví, počet stupňů volnosti je tedy dán výpočtem $4 \times 1 = 4$.

Běžně se namísto srovnávání testového kritéria s kvantily statistických rozdělení využívá vypočtené pravděpodobnosti chyby prvního druhu (většinou značené Sig., P, P-value či P-level). Postup, který se uplatňuje je následující: **v případě, že vypočtená pravděpodobnost chyby prvního druhu je menší než námi předem stanovená hranice** (nejčastěji 0,05 – viz dále v textu)**, zamítáme nulovou hypotézu, v případě opačném nulovou hypotézu nezamítáme**. Pokud postup použijeme v našem případě, zjišťujeme, že softwarem vypočtená pravděpodobnost chyby prvního druhu[10] je menší než 0,05, a tudíž zamítneme nulovou hypotézu a přijmeme alternativu, tj. muži a ženy v ČR se z hlediska frekvence konzumace alkoholu odlišují.

„Nedostatky" statistické významnosti

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

a) nedostatečná výpověď o základním souboru,
b) nereálnost nulových hypotéz,
c) mechanická práce s klasickou 5% hladinou (hvězdičky, stepwise, nejlepší modely apod.),
d) statisticky významné neznamená důležité,
e) nepublikování statisticky nevýznamných výsledků.

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

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

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

null hypothesis [Cohen 1994, Loftus 1996], česky bychom mohli užít výraz nicotné nulové hypotézy. Nemělo by být naším cílem namísto triviálního vyvracení nereálných hypotéz formulovat hypotézy, které mají reálný základ? Například namísto hypotézy o nulové korelaci v základním souboru ($R = 0$) formulovat hypotézu o slabé závislosti (řekněme $R \leq 0{,}2$). Obdobně místo hypotézy o nulovém rozdílu průměrů příjmů mužů a žen ($\mu_1 - \mu_2 = 0$), formulovat hypotézu o tom, že rozdíl je 2 000 Kč nebo nižší ($|\mu_1 - \mu_2| \leq 2000$)? Proč toto neděláme? Předně na to nejsme zvyklí a za druhé v softwarech je zpravidla možné testovat jen výše uvedené nicotné nulové hypotézy. Bylo by zřejmě dobré změnit naše zvyky a případně pátrat v našich softwarech, zda není možné uživatelsky nastavit reálnější nulové hypotézy. Ne vždy je toto nastavení možné, pak nezbývá než užít ruční výpočty a případně apelovat na tvůrce softwaru, aby příslušné procedury upravili k potřebě uživatelů. Dodejme, že proti autorům, kteří tvrdě kritizují nereálnost běžných nulových hypotéz [Cohen 1994, Loftus 1996], vystoupili jiní, kteří naopak nulové hypotézy hájí [Např. Biskin 1998].

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

Prostý uživatel často místo statistické významnosti používá jen oblíbené hvězdičky. Jedna hvězdička znamená významnost na 5% hladině, dvě hvězdičky pak 1% a tři hvězdičky 0,1%[11]. Z uvedeného příkladu (0,051 vs. 0,049) je zřejmé, jak pouhé čtení hvězdiček může být ošidné [Leahey 2005, Selvin 1957]. Počítač samozřejmě není moudrý analytik, ale uplatňuje striktně fisherovské doporučení. Na tomto místě je důležité upozornit i na skutečnost, že čím je větší výběrový soubor, tím je ceteris paribus pravděpodobnější, že se hvězdičky (příp. více hvězdiček) objeví. Tato skutečnost plyne ze zákona velkých čísel, podle něhož střední chyba odhadu při rostoucí velikosti výběrového souboru klesá [více o fenoménu velkých souborů a dopadu na statistickou významnost viz Soukup Rabuši 2007: 389–390].

Podobně jako hvězdičky se chovají i počítačové procedury hledající „nejlepší" model. Velice známá a oblíbená je například procedura stupňovité regrese (Stepwise regression), nicméně obdoby jsou známy i pro oblast lo-

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

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

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

Ad d) S výše uvedenou kritikou striktního dodržování 5% hladiny významnosti a případného čtení hvězdiček souvisí další varování. **Neplatí tvrzení, že čím více hvězdiček, tím je výsledek důležitější nebo kvalitnější.** Správně je pouze tvrzení: **nižší vypočtená hladina významnosti značí vyšší statistickou významnost. Ale nic více.** Tříhvězdičkový výsledek není hodnotnější než dvouhvězdičkový[12] (je jen méně pravděpodobné, že náš výběr je ze základního souboru, kde platí nulová hypotéza).

Ad e) S častým omylem statisticky významné = důležité souvisí i výzkumná praxe spočívající v publikaci pouze „důležitých” (správně: jen statisticky

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

významných) výsledků. V abstraktech vědeckých textů se povětšinou objevují výsledky, které byly statisticky významné, i když v samotném textu se dost často uvádí i poznatky, které se nepodařilo prokázat. Pozoruhodná je i strategie, kdy autor uvede v abstraktu, že mu některé výsledky nevyšly statisticky významné, ale přesto si myslí, že rozdíly u daného fenoménu existují. Věda se tak podivuhodně reprodukuje převážně statisticky významnými výsledky, nevýznamné lépe neuvádět, zejména pak v abstraktu. Tento hon za statisticky významnými výsledky lze vypozorovat jak v zahraničí, tak samozřejmě u nás. Prvním, kdo na problém upozornil, byl zřejmě Rosenthal [1979], který fenomén pojmenoval jako **problém hromadění pouze statisticky významných výsledků** (file drawer problem). Základní problém honby za statisticky významnými výsledky je zejména v tom, že v případě metaanalytických postupů jsou závěry dělány jen na základě statisticky významných výsledků, a kvůli tomu jsou výsledky zkreslené. Studie se statisticky nevýznamnými výsledky se následkem jejich nepublikování do metaanalýz nezahrnou. Lze odhadovat, že tento proces má multiplikativní efekt, protože čím více se v dané oblasti publikuje statisticky významných výsledků, tím spíše se ten, kdo takového výsledku nedosáhne, neodváží své výsledky publikovat[13].

II. Jak si poradit s „nedostatky" statistické významnosti?

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

Intervaly spolehlivosti (confidence intervals)

Pokud užíváme nebo publikujeme pouze vypočtenou hladinu statistické významnosti, jsou možnosti posouzení výsledků poměrně omezené. Ještě menší jsou, pokud uvedeme pouze, zda je výsledek statisticky významný či nikoliv. V této situaci lze říci, zda rozdíl (závislost) je statisticky významný, a toť vše. Nás ale spíše zajímá, jak moc je velký nebo malý (statisticky významný) rozdíl dvou skupin nebo velká či malá závislost proměnných. Bodovým odhadem velikosti rozdílu nebo závislosti je hodnota vypočtená z výběro-

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

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

	Nestandardizované koeficienty		Standardizované		
	B	chyba odhadu	Beta	t	Sig.
konstanta	6885,2	1024,0		6,72	3,27E−11
počet let školní docházky	615,2	61,7	0,31	9,97	3,16E−22
pohlaví	−3590,2	376,5	−0,30	−9,54	1,53E−20
Závislá proměnná: čistý příjem jedince					
R2 = 0,193, F-test (Sig < 0,0005)					

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

vých dat. Tato hodnota trpí všemi nedostatky bodového odhadu: jedná se o jediné číslo platné pouze pro náš výběr a je téměř vyloučeno, aby stejná hodnota byla platná i pro základní soubor. Z tohoto důvodu přistupují statistici ke konstrukci intervalů spolehlivosti. Tyto intervaly zahrnují s určitou pravděpodobností (nejčastěji se opět užívá fisherovských 95 %) hodnotu odhadované charakteristiky v základním souboru. Lze získat nejen informaci o tom, zda je výsledek statisticky významný (1)[14], ale **navíc lze získat i představu o tom, v jakém rozpětí se může hodnota příslušného parametru[15] pohybovat v celé populaci** (2). Ekvivalentní postup ke srovnání vypočtené hladiny statistické významnosti s hodnotou 0,05 lze u intervalů spolehlivosti aplikovat za pomoci této otázky: obsahuje interval spolehlivosti hodnotu platnou dle nulové hypotézy?[16] Ukažme srovnání obou postupů na jednoduchém příkladu regresní analýzy.

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

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

Z uvedené tabulky můžeme konstatovat, že léta vzdělání i pohlaví jsou v České republice statisticky významnými prediktory příjmu jedince (Sig. v posledním sloupci je zcela jistě menší než běžně užívaných 0,05). Interpretujeme-li klasicky hodnoty regresních koeficientů (sloupec nadepsaný B), lze říci, že české ženy měly v průměru v roce 1999 o 3 590 Kč méně než muži

14 Tedy informaci, kterou lze získat i statistickým testováním (viz výše).

15 V souvislosti s diskusemi o věcné významnosti se většinou hovoří o tzv. velikosti efektu (rozdílu, koeficientu apod.).

16 Pokud ji neobsahuje, zamítáme H_0 a přijímáme H_1, v opačném případě pak nezamítáme H_0.

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

	Koefi-cienty	Chyba odhadu	t	Sig.	Dolní mez	Horní mez
konstanta	6885,2	1024,0	6,72	3,27E−11	4875,3	8895,1
počet let školní docházky	615,2	61,7	9,97	3,28E−22	494,1	736,2
pohlaví	−3590,2	376,5	−9,54	1,55E−20	−4329,1	−2851,3
R2 = 0,193, F-test (Sig < 0,0005)						

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

(protože kodování proměnné bylo 1 = muž, 2 = žena) a že s každým rokem vzdělání si Čech či Češka průměrně přilepší o 615 Kč. Tyto bodové odhady mohou být ale poměrně vzdálené od skutečné hodnoty těchto parametrů v základním souboru. Proveďme proto ještě jednou analýzu s tím, že vypočteme navíc intervalové odhady regresních koeficientů (tabulka 5).

V tabulce 5 jsou nové dva poslední sloupce. Naznačují nám, že rok školy navíc přinese (s 95% pravděpodobností) v ČR něco mezi 494 a 736 Kč čistého příjmu navíc. Rozdíl mezi českými ženami a muži se s velkou pravděpodobností (opět 95 %) pohybuje mezi 2 851 a 4 329 Kč. Vidíme, že výpověď je mnohem nejednoznačnější, zároveň ale mnohem bližší realitě. Statistika (aspoň ta inferenční) neumí sdělovat své závěry s jistotou, ale jen s určitou mírou pravděpodobnosti. Připomeňme, že v intervalu spolehlivosti je skryta nejen informace o statistické významnosti (pokud by interval obsahoval nulu, nebyl by regresní koeficient statisticky významně odlišný od nuly), ale navíc i informace o možné hodnotě koeficientu v celém základním souboru.

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

Míry asociace a problematičnost intervalových odhadů

Zatímco s intervaly spolehlivosti regresních parametrů a rozdílů v průměrech si lze většinou bez problémů poradit, u měr asociace je situace mnohem složitější. Vypočítáme-li například z výběrových dat hodnotu korelačního koeficientu dle Pearsona 0,7, jaký je jeho intervalový odhad? Jaká je jeho pravděpodobná hodnota v celém základním souboru. Standardně umíme

provést test o nulové hodnotě příslušné míry asociace a rozhodnout, zda proměnné na sobě závisí (většinou jen lineárně) či nikoliv. Nicméně daleko zajímavější může být zjištění, jak moc spolu proměnné souvisí v základním souboru, a o tom nám bodový odhad z výběrových dat opět moc nepoví. Problém intervalového odhadu korelačního koeficientu tkví v tom, že tento nemá při neznámé hodnotě jeho skutečné hodnoty žádné známé statistické rozdělení, a tudíž nelze snadno užít kvantity těchto rozdělení. Nicméně je známo, že po fisherově transformaci má Pearsonův korelační koeficient přibližně normální rozdělení a tohoto lze využít pro konstrukci intervalu spolehlivosti. Postup lze popsat takto [Fan, Thompson. 2001: 525]:

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

Tento postup je relativně jednoduchý, ale kdybychom ho chtěli následovat pokaždé, při práci by nás zdržoval. Statistici naštěstí vyvinuli pomůcky, které výpočty provedou za nás[17]. Stačí pouze dosadit hodnotu korelačního koeficientu vypočítaného z výběru a počet respondentů. Pomůcky mají podobu on-line kalkulátorů na webu nebo tabulkových kalkulátorů dostupných přes web[18]. Problémem těchto kalkulátorů je omezení intervalů spolehlivosti na 90 % nebo 95 %. Pro zájemce jsem vytvořil obecnější pomůcku, která je k dispozici na mých webových stránkách[19]. Nicméně protože v sociologii častěji používáme korelační koeficienty pro ordinální data (Spearman, Kendallovo tau), bylo by vhodnější počítat intervaly spolehlivosti pro tyto koeficienty. Zde nelze užít postupu transformace na veličinu s přibližně normálním rozdělením, ale nejvhodnější je nejspíše technika bootstrapu (ta by byla použitelná i pro Pearsonův koeficient). Detailnější popis tohoto přístupu přesahuje možnosti tohoto textu, zájemce lze odkázat na freewarový statistický program R, kde je tento přístup implementován. Uvedení intervalu spolehlivosti (a jeho následná interpretace) pro korelační koeficient je určitě

17 Z běžně užívaných statistických paketů umí intervaly spolehlivosti pro korelační koeficienty počítat SAS a STATISTICA.

18 Příklad prvního typu lze nalézt na http://faculty.vassar.edu/lowry/rho.html a druhého typu na např. na http://www.childrens-mercy.org/stats/weblog2005/CorrelationCoefficient.asp.

19 Viz http://samba.fsv.cuni.cz/~soukup/pomucky.

daleko vhodnější než test o nulové hodnotě koeficientu někdy doprovázený jeho bodovým odhadem.

Shrňme, že intervaly spolehlivosti jsou slibnou alternativou k testům statistické významnosti nulových nerálných hypotéz. Jejich hlavní výhodou oproti testům je, že dopředu **není třeba žádné hypotézy formulovat**[20]. Další výhodou je **možnost používání metaanalytických postupů ze získaných intervalů spolehlivosti**. Intervaly spolehlivosti jsou zejména zastánci věcné významnosti považovány za vhodný nástroj [Rozeboom 1960, Cohen 1988, 1994, Thompson 2002], nicméně jejich užívání doporučují i mnozí jiní autoři [Tversky, Kahneman 1971, Jones 1984, Jones, Matloff 1986, Perry 1986, Hunter 1990, Robinson, 2003 Brandstätter 1999, **Denis** 2003].

Síla testu (Power analysis)

Oponenti uvádění statistické významnosti oprávněně namítají, že problém statistického testování může spočívat v tom, že sice držíme pod kontrolou chybu prvního druhu, ale ztrácíme ze zřetele, jaká je síla testu [Tversky, Kahneman, 1971 Jennions, Miller 2003, Denis 2003, Cohen, 1988, 1994, Borkowski, Welsh 2001, McCloskey 1985]. Připomeňme, že **síla testu** (viz tabulka 1 a komentář k ní) **je pravděpodobnost** (hodnota pohybující se mezi 0 a 1) **správného přijetí alternativní hypotézy za předpokladu, že je tato v základním souboru platná**. U mnohých výzkumů je tato síla testu velmi malá, ale protože ji výzkumník nezná, nemůže toto posoudit.

Výpočet síly testu je záležitostí matematiků-statistiků, v jejichž textech lze nalézt příslušné vzorce [česky např. Anděl 2003, 2005, Kalounová, 2000]. Pro sociální vědce je daleko vhodnější užívání statistického softwaru, který má příslušné vzorce implementovány. Již v roce 1989 napočítal Goldstein [1989] třináct produktů, které umí tyto výpočty provést. V dnešní době jich jsou jich desítky a navíc mají tyto procedury implementovány i některé obecné statistické produkty (např. SPSS, STATA, Statistica). Mnohé programy umožňují samostatné výpočty síly testu, jiné produkty jsou doplňkem existujících statistických produktů (SAS) nebo i tabulkových kalkulátorů (Excel). Detailnější přehledy a srovnání produktů lze nalézt v článcích [Goldstein, 1989, Thomas, Krebs, 1997] a samozřejmě na internetu po zadání sousloví „power analysis“. Sílu testu je vhodné počítat orientačně ještě před provedením výzkumu a spojit s ní úvahu o velikosti výběrového souboru (viz další část článku o velikosti výběrového souboru).

Platí pravidlo, že pokud je síla testu malá, není vhodné výzkum vůbec provádět, protože přijetí alternativní hypotézy není pravděpodobné. V literatuře bývá doporučována hodnota 0,8 jako minimální pro sílu testu [Vacha-Haase; Nilsson 1998: 49]. Dodejme, že při této hodnotě je pravděpodobnost chyb-

20 Dodejme, že je samozřejmě žádoucí před prováděním analýz (i celého výzkumu) hypotézy formulovat. Děkuji za upozornění na toto doplnění anonymnímu recenzentovi.

ného nezamítnutí nulové hypotézy při její neplatnosti 0,2, a to není málo. Pro zájemce dodejme, že síla testu závisí na pravděpodobnosti chyby prvního druhu (čím je vyšší, tím je vyšší síla testu), na velikosti výběrového souboru (u větších výběrových souborů je síla testu vyšší) a na rozdílu sledovaného ukazatele v porovnávaných skupinách (vyšší rozdíl implikuje ceteris paribus vyšší sílu testu). Samozřejmě, že síla testu závisí i na reliabilitě měřeného ukazatele.

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

Pro možnost detailnějšího pochopení síly testu demonstrujme na jednoduchém příkladu koncept síly testu a podstatu výpočtu.

Příklad 3: Výpočet síly testu a ukázka souvislosti testu s velikostí výběrového souboru a velikostí efektu

Pro jednoduchost uvažujme, že měříme například počet dětí ve třídách v mateřské škole. Předpokládejme, že v loňském roce bylo zjištěno, že průměrný počet dětí ve třídě v mateřské škole v ČR byl 20 dětí (na základě údajů ze školské statistiky). Letos jsme provedli dotazování ve 25 (n) náhodně vybraných mateřských školách a zjistili jsme, že v těchto vybraných školkách je průměrně 22 dětí ve třídě ($\Delta = 2$)[21], směrodatná odchylka (s) byla 4 děti. Testované hypotézy plynoucí z našeho výzkumu jsou následující:

H_0: Průměrný počet dětí ve třídách v ČR se meziročně nezměnil (činí tedy i letos 20 dětí), tj. $\mu = 20$.[22]

H_1: Průměrný počet dětí ve třídách se meziročně změnil (zvětšil či zmenšil, použijeme tzv. oboustranný test), tj. $\mu \neq 20$.

Postup řešení: Pro řešení úlohy se nabízí využít jednovýběrový t-test, nejdříve vypočítáme test a dosaženou chybu prvního druhu (hladinu statistické významnosti). Poté provedeme výpočet síly testu a ukážeme graficky sílu testu pro jiné velikosti výběrových souborů a jiná naměřená data.

Testové kritérium $t = \Delta / s / \sqrt{n} = 4 / 2 / 5 = 10$.

Odpovídající chyba prvního druhu se zjistí za pomoci inverzní funkce hus-

21 Symbol Δ používáme pro napozorovaný efekt, tj. pro rozdíl mezi průměrným počtem dětí skutečně napozorovaným (22) a průměrem zjištěným loni (20), který udává naši nulovou hypotézu (viz dále).

22 Symbol μ je vyhrazen pro neznámou střední hodnotu sledovaného ukazatele v celé populaci, tj. průměrný počet dětí ve třídě ve všech školkách v ČR.

toty pravděpodobnosti t-rozdělení s 24 stupni volnosti[23]. Vypočtená hodnota chyby prvního druhu činí 0,006 (pravděpodobnosti zamítnutí nulové hypotézy na základě našich dat za předpokladu, že v populaci nulová hypotéza platí) a vede nás k zamítnutí nulové hypotézy a přijetí hypotézy alternativní. Budeme tedy tvrdit, že průměrný počet dětí ve třídách se změnil, s ohledem na zjištění plynoucí z našeho výzkumu (s ohledem na průměr vyplývající z našich dat uzavřeme, že došlo ke zvýšení průměrného počtu dětí ve třídě).

Bez znalosti síly testu resp. jejího výpočtu, ale vůbec nevíme nakolik „silné" je naše tvrzení o zvýšení průměrného počtu dětí ve školkách. Je tedy namístě spočítat pravděpodobnost přijetí hypotézy o změně počtu dětí ve třídě, za předpokladu, že tato hypotéza v populaci opravdu platí (síla testu, 1 – β). Doplňkově také zjistíme, jaké chyby II. druhu (β) se dopouštíme.

Postup pro výpočet síly testu v našem konkrétním případě je následující [Zvára, Štěpán 2000: 169–171]: Vypočítáme si charakteristiky t-rozdělení za předpokladu platnosti alternativní hypotézy, tj. nalezneme kritický obor testového kritéria t při platnosti nulové hypotézy za pomoci intervalu spolehlivosti: [24]

$$\mu + t_{0,95}(n-1) \times s / \sqrt{n} = 20 + 1{,}71 \times 4 / 5 = 21{,}37 \quad (25)$$

Pro výpočet chyby druhého druhu budeme využívat inverzní funkci k hustotě pravděpodobnosti t-rozdělení se 24 stupni volnosti pro hodnotu danou výpočtem: (22 – 21,37) / 4 / 5 = 0,032.[(26)]

Chyba druhého druhu (β) činí v našem případě 0,44 a odečtením od jedné získáme sílu testu 0,56 = 1 – 0,44. Můžeme tedy říci, že v našem případě jsme sice zamítli nulovou hypotézu (za předpokladu, že tato v populaci platí) o nezměněném průměrném počtu dětí ve třídě s velmi malou pravděpodobností (0,006), ale **síla našeho tvrzení o přijetí alternativní hypotézy**

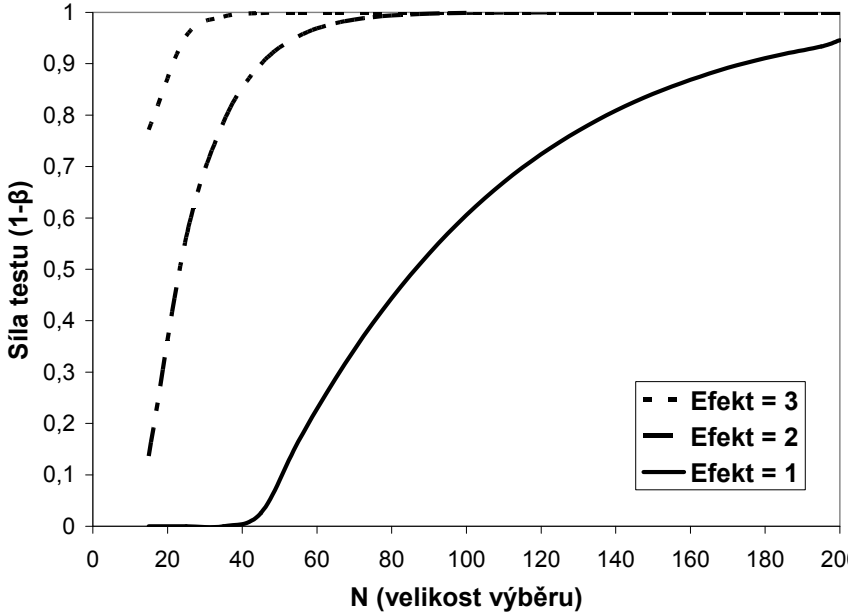
23 Počet stupňů volnosti jednovýběrového t-testu je dán vzorcem n – 1, tedy o jednotku zmenšenou velikostí výběru.

24 Při výpočtu síly testu je třeba předem stanovit pravděpodobnost chyby prvního druhu. Vycházíme z konvenčně užívané hodnoty 0,05. Zvýšení této pravděpodobnosti povede ceteris paribus ke zvýšení síly testu a naopak.

25 Na tomto místě je nutno upozornit, že výpočet byl zjednodušen. Pro výpočet síly oboustranného testu by měla být počítána dolní a horní mez 95% intervalu spolehlivosti (využilo by se tedy nikoliv 95%, ale 97,5% kvantilu) a síla testu by se vypočetla jako součet dvou hodnot. Namísto toho jde zde počítána jen horní mez intervalu (tak, jako kdyby se prováděl jednostranný test) a nebylo poté třeba síly testu (resp. chyby II. druhu) sčítat. Numericky jsou oba postupy ekvivalentní, použitý postup je při ručním výpočtu jednodušší.

26 Hodnota 22 odpovídá napozorovanému průměru v našich datech, hodnota 21,37 je hodnota horní meze intervalu spolehlivosti z kroku 1. Rozdíl mezi těmito hodnotami ještě dělíme směrodatnou odchylkou (4) a odmocninou s počtu pozorování. Výsledná veličina má t-rozdělení, a proto lze užívat její kvantily či hustoty pravděpodobnosti.

Graf 1. Velikost síly testu pro jednovýběrový oboustranný t-test pro různé velikosti výběrového souboru a různé velikosti efektů (α = 0,05)



je poměrně malá. Konkrétně je **pravděpodobnost přijetí alternativní hypotézy o změně počtu dětí ve třídách** (za předpokladu, že v populaci ke změně došlo) jen o **velikosti 0,56!** Tato hodnota je výrazně nižší než v literatuře doporučovaná mez 0,8 (srov. výše). Zároveň je pravděpodobnost chybného nezamítnutí nulové hypotézy při její neplatnosti 0,44 (β).

Pro ilustraci vlivu velikosti výběrového souboru a velikosti naměřené charakteristiky ve výběrovém souboru použijeme graf 1.[27] Jednotlivé křivky ukazují vývoj velikosti síly testu s ohledem na velikosti výběrového souboru. Jednotlivé křivky jsou zkonstruovány pro náš příklad, za předpokladu, že ve výběrovém souboru byl naměřen průměrný počet dětí ve třídách o velikosti 21 (efekt = 1)[28], 22 (efekt = 2), resp. 23 (efekt = 3).

Graf potvrzuje empiricky na našich datech, že s rostoucí velikostí výběrového souboru roste síla testu (nelineárně), síla testu roste též s velikostí napozorovaného efektu (srov. křivky pro různé velikosti efektu). Pokud se podíváme na nejnižší křivku (odpovídá průměrně napozorovaným 21 dětem

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

28 Připomeňme, že efektem rozumíme rozdíl mezi skutečně napozorovanou hodnotou příslušné charakteristiky a hodnotou předpokládanou dle nulové hypotézy (v našem případě 20 dětí ve třídě).

ve třídě), vidíme, že potřebné síly dosahuje t-test až pro cca 140 pozorování. Pro úplnost dodejme, že pro výše uvedené křivky platí s výjimkou nejnižší (efekt = 1) do cca 45 pozorování (viz graf 2), že výsledky t-testu by byly statisticky významné minimálně na 5% hladině a zamítali bychom tedy nulovou hypotézu. Síla těchto zamítání by ovšem byla v mnohých případech poměrně malá, jak plyne z grafu 1.

Minimální velikost výběru (n)

Další doporučovanou pomůckou pro lepší užívání statistických procedur ve výzkumech je plánování velikosti výběru [Snyder, Lawson 1993]. Každý výzkumník, který chce provádět výběrové šetření, by měl dobře zvážit, jak veliký soubor je pro něj vhodný. Nemělo by záležet na libovůli ani na dostupných finančních prostředcích, ale na statistickém zdůvodnění velikosti výběrového souboru. Klasický vzorec pro minimální velikost výběru záleží na velikosti očekávaného rozdílu, tj. efektu (Δ), směrodatné odchylce měřené charakteristiky (σ) a pravděpodobnosti, se kterou budeme chtít odhadovat průměrnou velikost hodnoty měřené charakteristiky pro celou populaci (nejčastěji opět klasických 95 %). Můžeme použít vzorec 1.1.

$$n \geq (u_{1-\alpha/2})^2 \times \sigma^2 / \Delta^2 \quad (1.1)$$

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

Pro ilustraci vlivu velikosti výběrového souboru na chybu prvního druhu se vraťme k příkladu 3. Předpokládáme velikost napozorovaného efektu 1, velikosti chyby prvního druhu pro různě velké výběrové soubory uvádí graf 2.[29] Graf potvrzuje klesající pravděpodobnost chyby prvního druhu s ohledem na velikost výběrového souboru (ceteris paribus), pokles je ovšem nelineární. Na tuto skutečnost (souvislost statistické významnosti a velikosti výběrového souboru) se často poukazuje jako na negativum. Pro velké datové soubory (řádově tisícové a větší) pak nemá testování statistické významnosti valný význam [Soukup, Rabušic 2007].

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

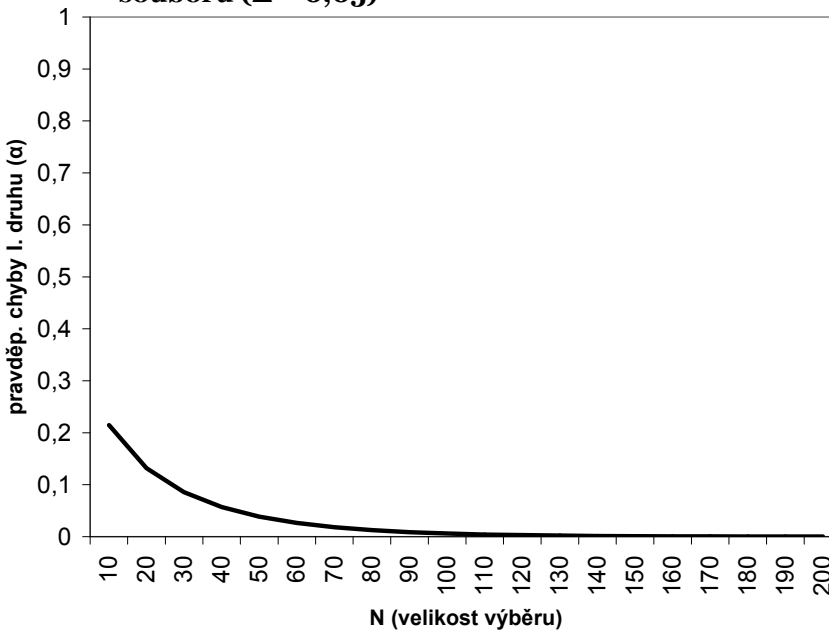
$$n \geq (u_{1-\alpha/2})^2 \times (p \times (1-p))^2 / d^2, \quad (1.2),$$

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

V sociologických šetřeních je situace vhodné velkosti výběru ještě komplikovaná tím, že většinou není cílem výzkumu měřit jednu charakteristiku. Proto velikost výběrového souboru by měla být větší nebo rovna největší z hodnot, které vypočteme dle uvedených vzorců pro jednotlivé rozdíly a proporce, jež hodláme ve výzkumu měřit. Dodejme navíc, že výše uvedené

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

Graf 2. Velikost chyby prvního druhu pro jednovýběrový oboustranný t-test pro různé velikosti výběrového souboru (Δ = 0,05)



vzorce platí stricto sensu pro prostý náhodný výběr, pro jiné designy výzkumů existují vzorce obdobné a výše uvedené platí s určitou mírou nepřesnosti. Dobrou zprávou pro navrhovatele výzkumů je, že výpočet minimální velikosti vzorku je již implementován do mnohých statistických produktů, které zpravidla umí odhadovat i sílu testu (více viz v předchozí části věnované tomuto tématu). Kromě apriorního odhadu velikosti výběrového souboru je některými autory [Snyder 2000] doporučováno a posteriori vypočítat pro každé zjištění z výzkumu minimální velikost výběrového souboru k prokázání statisticky významného rozdílu nebo závislosti (tzv. **What if strategy**).

Porovnávání více modelů za pomoci informačních kritérií

Jako relativně slibné řešení slabin statistické významnosti se jeví také užívání postupu, kdy není testován jeden model a přijímán nebo odmítán za pomoci hodnot typu 0,09 nebo 0,03. Postup spočívá v tom, že výzkumník negeneruje jednu hypotézu o jednom modelu, ale na základě teoretických předpokladů (a to je pro tento přístup, ale nejen pro něj, klíčové) několik vzájemně si konkurujících modelů, tzv. multiple working hypothesis [Anderson, Burnham, Thompson 2000]. Pro jednotlivé modely je poté vypočtena věrohodnostní funkce a na základě ní hodnota informačního kritéria. Nejčastěji je užíváno

Akaikeho informační kritérium (AIC, [Akaike, 1972]) a bayesovské Schwarzovo informační kritérium (BIC, [Raftery, 1995]). Tato kritéria jsou uváděna zpravidla ve formě „čím menší, tím lepší model" a analytik pak velice jednoduše vybere model, který je nejvhodnější pro jeho data. Tento postup je hojně užíván v rámci strukturních [Urbánek 2000], regresních a loglineárních [Hebák a kol. 2005a: 111, resp. Hebák a kol. 2005b: 35] a víceúrovňových modelů [Soukup 2006: 1007]. U jednodušších procedur se zpravidla neužívá, ale to samozřejmě neznamená, že jej nelze užít. Bez problémů lze počítat hodnoty informačních kritérií i pro jiné než uvedené procedury, jako například regresní analýzu (lineární i logistickou).

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

- princip **parsimonie** – tj. snaha vybrat co nejjednodušší model obsahující minimální počet parametrů postačujících k popisu reality na základě našich dat,
- princip **srovnávání více modelů namísto dichotomického rozhodování** o nulové vs. alternativní hypotéze (tento princip je založen na představě, že zpravidla existuje několik teoretických koncepcí, které si vzájemně konkurují a tuto pluralitu by měla odrážet i analýza dat) a
- princip **získání silného důkazu** – vychází z problému nicotných nulových hypotéz (viz část Nedostatky statistické významnosti) a preferuje výběr mezi silnými hypotézami o několika modelech (oproti rozhodování nicotná nulová hypotéza vs. smysluplná alternativní hypotéza).

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

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

Pro úplnost dodejme vzorec užívaný pro výpočet neznámějšího Akaikeho kritéria:

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

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

Ze vzorce je patrné, že se zohledňuje věrohodnost příslušného modelu (tedy pravděpodobnost, že model je vhodný pro naše data, de facto informační hodnota modelu) a dále složitost modelu (tj. princip parsimonie); složitější modely jsou penalizovány navyšováním AIC za každý další parametr o dvě

jednotky. Pro vyhodnocení se užívá pravidla, že modely s nižší hodnotou AIC jsou vhodnější. Kromě základního AIC kritéria existuje i modifikované AIC kritérium (AIC_c) pro malé výběry (malý výběr je zde definován jako výběr, kde počet pozorování je méně než čtyřicetinásobkem počtu parametrů modelu).

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

Z pohledu sociologie vědy je poměrně zajímavé, že kritérium AIC, i když vzniklo dříve, je v sociologii téměř neznámé a dominuje užívání kriteria BIC. Weakliem [2004: 170] uvádí, že v databázi JSTOR nalezl v sociologických článcích 16 užití AIC a 100× bylo použito kritérium BIC. U článků z oboru ekonomie bylo použito AIC ve 120 případech, BIC ve 100. Vysvětlení pro tuto disproporci je poměrně jednoduché. Zatímco kritérium AIC bylo sociologům detailněji představeno až v roce 2004 (viz dále), kritérium BIC zavedl do sociologie již v polovině 80. let Raftery [1986, 1995]. Od té doby se kritérium BIC stalo standardním sociologickým nástrojem, často problematicky a nesprávně užívaným, obdobně jako statistická významnost [např. Weakliem 1999, Firth Kuha 1999]. Zájemce o detailní informace o kritériu AIC (a srovnání s kritériem BIC) lze odkázat na monotematické číslo Social Methods and Research z roku 2004 (vol. 33).

Přejdeme nyní k druhému informačnímu kritériu, tzv. bayesovskému Schwarzovu informačnímu kritériu. Kritérium bylo prvně představeno Schwarzem [1978], do sociologie jej zavedl již zmíněný Raftery jako alternativu k napadanému statistickému testování. BIC na rozdíl od AIC vychází explicite z bayesovské statistiky a cílem kritéria je porovnávat poměrově věrohodnost (aposteriorní pravděpodobnost[30]) našeho modelu a modelu saturovaného (tj. modelu obsahujícího všechny myslitelné parametry). Pro úplnost uveďme i vzorec bayesovského Schwarzova informačního kritéria:

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

kde N je velikost výběrového souboru.

S ohledem na uvedenou logiku **platí** i zde, **že se preferují modely s nižším BIC a prvotně se srovnává se saturovaným modelem,**

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

jehož BIC je nulové. **Modely ze záporným BIC jsou vnímány jako modely lepší než saturovaný** (jsou jednodušší a přesto nemají výrazně menší aposteriorní pravděpodobnost).

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

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

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

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

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

Dodejme ještě několik poznámek o problémech při praktickém používání informačních kritérií. Jedním z problémů může být situace, kdy dle jednoho kritéria je nejlepší například model *A* a dle druhého jiný model, například *B* [Soukup 2006: 1007]. Situace však může být ještě komplikovanější, protože informačních kritérií je více (cca 10) a tudíž i „nejlepších" modelů může být více. Zde analytikovi nezbývá než přihlédnout i k jiným pomůckám, jako jsou celkové testy modelů, diagnostika reziduí apod. Nelze doporučit, aby se ana-

lytik spolehl pouze na užívání informačních kritérií a užíval je jako doplňkový nástroj[31]. Doplňme, že v rámci strukturních modelů se nabízí i jiná kritéria ke srovnávání modelů, jako je AGFI, RMSEA apod. [Urbánek 2000]. I tato kritéria jsou mnohými autory doporučována jako alternativa ke statistické významnosti [Onwuegbuzie, Levin, Leech, 2003: 1042].

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

Slovní řešení nedostatků v rámci statistické významnosti

Někteří autoři navrhují, aby došlo ke změně výrazu statistická významnost, případně aby se výrazu striktně užívalo takto a nebylo používáno pouze samotného slova významnost. Jaké návrhy lze nalézt v literatuře? První relativně mírný návrh formuloval Thompson [1996]. Jeho doporučení zní: „**Vždy užívejte** sousloví **statistická významnost**, nikdy pouze slova významnost“ (tučně autor článku). Toto doporučení plyne z častého omylu významný = důležitý. Aby k tomuto pomýlení docházelo pokud možno co nejméně, navrhuje Thompson výše uvedené pravidlo.

Jiní autoři s ním ale polemizují, že tato praxe je jazykově problematická a nejspíš zbytečná [Levin, Robinson 1997]. Navrhují naopak nahradit sousloví statistická významnost jinými slovy, jako vhodný adept se jim jeví slovo **nenáhodnost** (non-chance). Tento výraz má vyjádřit, že statisticky významný výsledek (dle alternativní hypotézy) není získán náhodně, ale je poměrně pravděpodobné, že tento výsledek platí i pro celou populaci. S tímto názorem ovšem vyslovil polemiku opět Thompson [1998b].

Na citátech si ukažme problematičnost (nesprávnost) užívání sousloví statistická významnost v české vědě. „Navíc jsme dosáhli **statistické významnosti všech čtyř parametrů**.“ [Trilobyte 2004]. „**Statistické významnosti bylo dosaženo**.“ [Widimský 1999]. Oba citáty nepřímo naznačují, že hlavním cílem vědy (resp. výzkumu) je dosahovat statistické významnosti. To je nutno razantně odmítnout. Statistická významnost pouze hovoří o možnosti zobecnit data z výběrů (resp. experimentů) na populaci, ale nijak neměří významnost věcnou (viz počátek článku). Uveďme s komentářem další dva citáty: „**Rozdíl nedosáhl statistické významnosti**“.

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

[Ballantyne, M. B.; Olsson, A. G.; Cook, T. J.; Mercuri, M. F.; Pedersen, T. R.; Kjekshus, J. 2001]. „**A tento pokles (bez statistické významnosti**) nastal i ve všech sledovaných okresech". [Státní zdravotní ústav 1999]. Pokles či rozdíl může být statisticky významný na nějaké hladině, kterou musí výzkumník explicitně uvést (a pokud možno i dopředu stanovenou), jinak je tvrzení značně nepřesné. Namístě je ovšem komentovat zejména věcnou velikost rozdílu a statistickou významnost případně pouze zmínit jako doplněk. Užívání neúplného výrazu dokumentuje citát učebnice biomedicínské statistiky: „Statistické programy nám umožňují testovat významnost parciálních korelačních koeficientů." [Euromise] Dodejme, že statistické pakety umožňují testovat právě statistickou významnost, nikoliv jinou, slušelo by se tedy doplnit do citované věty slovo „statistickou".

S ohledem na uvedené citace je namístě podpořit Thompsonovo doporučení **užívat vždy plného sousloví statistická významnost a navíc ji užívat jen tam, kde je to vhodné**. Česky by bylo možné používat místo sousloví statistická významnost v oblasti výpočtů založených na výběrech též slova **zobecnitelnost**, které nemá zavádějící konotace. Výraz nenáhodnost není pro český kontext zřejmě vhodný.

Závěr

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

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

Literatura

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

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

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

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

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

Arbuthnott, J. 1710. An argument for divine providence taken from the regularity in the births of both sexes. *Philosophical Transactions of the Royal Society*. Vol. 27: 186–190.

Ballantyne, M. B.; Olsson, A. G.; Cook, T. J.; Mercuri, M. F.; Pedersen, T. R.; Kjekshus, J. 2001. Vliv nízkého HDL cholesterolu a zvýšené hladiny triglyceridů na riziko kardiovaskulárních příhod a účinek terapie simvastatinem ve studii 4S. Dostupné z http:// www.zdravcentra.cz/cps/rde/xchg/zc/xsl/79_1720.html.

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

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

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

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

Burnham, K. P.; Anderson, D. R. 2004. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*. Vol. 33: 261–304.

Cohen, J. 1994. The earth is round ($p < .05$). *American Psychologist*. Vol. 49: 997–1003.

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

Cox, D. R. 1982. Statistical significance tests. *British Journal of Clinical Pharmacology*. Vol. 14 : 325–331.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

King, G. 1986. How Not to Lie With Statistics: Avoiding Common Mistakes in Quantitative Political Science. *American Journal of Political Science*, Vol. 30 (No. 3): 666–687.

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

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

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

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

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

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

Raftery, A., E. 1995. Bayesian model selection in social research. In Mardsen, P., V. Sociological Metodology. MA: Cambridge. Blackwell.

Raftery, A., E. 1999. Bayes Factors and BIC: Comment on "A Critique of the Bayesian Information Criterion for Model Selection". *Sociological Methods & Research February*. Vol. 27: 411–427.

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

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

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

Rosenthal, R. 1979. The "file drawer problem" and the tolerance for null results. *Psychological bulletin*. Vol. 86: 638–641.

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

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

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

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

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

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

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

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

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

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

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

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

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

Trilobyte. 2004. Stránky statistického software Trilobite. Dostupné z http:// www.trilobyte.cz/qlin.html.

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

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

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

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

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

Widimský, J. 1999. Hypertenze starších osob. *Časopis české společnosti pro hypertenzi*. Vol. 2. Dostupný z http://www.hypertension.cz/casopis/2_99/6.html.

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

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