Iodny: Human sexuality. 5th Ed. New York, Harpers Collins College Publishers 1995.
McConaghy, N.: Sexual behavior. New York, Plenum Press 1993.
Mellan, J., Šípová, I.: Mladé manželství. Praha, Avicenum 1991.
Morris, D.: The naked ape. New York, Mc Graw Hill 1967.
Pondělíček, I., Pondělíčková-Mašlová, J.: Lidská sexualita. Praha, Avicenum 1971.
Pondělíček, I., Pondělíčková-Mašlová, J.: Jak (se) lidé milují. Praha, SNTL 1990.
Raboch, J.: Lékařská sexuologie. Praha, Státní pedagogické nakladatelství 1984.
Raboch, J., Hubálek, S.: Příspěvek ke studiu koitálního orgasmu u ženy. Čs. gynek., 1984, 49: 257-261.
Szilágyi, V.: Teórie o ženskom orgazme. „Forenzné aspekty sexuality, sexualita ženy." Zborník referátov prednesených na V. a VI. Košických sexuologických dňoch, 1980 a 1982.
Weiss, P., Zvěřina, J.: Sexuální chování obyvatel České republiky. Výsledky národního průzkumu. Praha, Alberta Plus 1999.
Zvěřina, J.: Sexuologie (nejen) pro lékaře. Brno, CERM 2003.

SÚHRN

Autori sa pokúsili o shrnutie základných poznatkov o charakteristikách ženského orgazmu. Uvádzajú vývoj skúmania orgazmu u žien, jeho psychologické i fyziologické aspekty, mechanizmy jeho prejavov. Charakterizujú základné štyri štádiá sexuálneho vzrušenia a ich anatomicko-fyziologické prejavy. Záverom sa zameriavajú aj na základné charakteristiky orgazmu vaginálneho a klitoridálneho typu.

UPOZORNĚNÍ PRO PŘISPĚVATELE ČASOPISU ČESKOSLOVENSKÁ PSYCHOLOGIE

Metodická studie Tomáše Urbánka „K prezentaci výsledků statistických analýz – 1. část", vznikla – právě tak, jako druhá část, která bude publikována v dalším ročníku našeho časopisu – na podnět redakční rady. Rukopisy zasílané do redakce často bohužel nesplňují kritéria pro publikaci v časopise, který je excerpován v Current Contents a má nezanedbatelný impakt faktor. Statistické zpracování výzkumných studií je důležitou částí publikovaných příspěvků a metodická studie Tomáše Urbánka „Poznámky k prezentaci výsledků statistických analýz" prezentuje standardy, které bude redakce při posuzování příspěvků nadále uplatňovat.

Redakční rada časopisu Československá psychologie

Československá psychologie 2007 / ročník LI / číslo 6

Metodické studie

K PREZENTACI VÝSLEDKŮ STATISTICKÝCH ANALÝZ – 1. ČÁST

TOMÁŠ URBÁNEK
Psychologický ústav AV ČR, Veveří 97, 602 00 Brno

ABSTRACT

To the presentation of results of statistical analyses – part one

T. Urbánek

The article presents some general principles of publishing the results of the statistical analyses of empirical data. In the first part it deals with rather general rules and principles, in the second one with several particular statistical tests, methods and approaches, explanation of their pursuing and listing the results that should be known to reader so as to be able to understand author's conclusions.

key words:
statistical analysis,
hypotheses testing,
presentation of results

klíčová slova:
statistická analýza,
testování hypotéz,
prezentace výsledků

1. ÚVOD

Začínající autoři připravující publikaci výsledků analýz svých empirických dat, ale i někteří zkušení autoři, kteří začali používat novou metodu analýzy, si nejsou jisti, jaká je „správná" forma prezentace výsledků analýz – které hodnoty ze záplavy výstupů poskytovaných statistickými programy uvést v článku, aby byla podpora výzkumných zjištění přesvědčivá a rozsah článku přitom nepřekročil meze stanovené redakcí. Důsledkem této nejistoty pak může být rukopis článku, který je na základě posudků recenzentů opakovaně vracen k přepracování, pokud není rovnou odmítnut.

Tento článek je pokusem vysvětlit podstatné aspekty prezentace takových výsledků ve vědeckých publikacích. Úmyslně přitom nepostupuji způsobem, který se nabízí, totiž uvedením různých („kanonických") seznamů pravidel pro autory publikovaných ve formě pokynů redakčních rad nebo publikačních manuálů a jejich sumarizací. Takový postup je sice možný, ale „pocit jistoty" by článek připravený tímto způsobem případným čtenářům, kteří by podle jeho doporučení chtěli publikovat, stejně nedal, protože by neposkytl vhled do důvodů, proč jsou tyto požadavky rozumné. (Kromě toho by byl takový článek pravděpodobně nudný jak pro autora, tak pro případného čtenáře.)[1]

Došlo: 31. 5. 2006; T. U., Psychologický ústav AV ČR, Veveří 97, 602 00 Brno; e-mail: tour@psu.cas.cz

Příspěvek byl napsán v rámci výzkumného záměru Psychologického ústavu AV ČR reg. č. AV0Z70250504 s názvem „Člověk v kontextech celoživotního vývoje".

[1] To ovšem neznamená, že by takové editorialy a diskuse o publikačních standardech nebyly samy o sobě zajímavé. Např. v APA byla v 90. letech vytvořena pracovní skupina zabývající se publikováním výsledků statistických analýz. Z jejich jednání vyplynuly určité závěry, které byly publikovány (Wilkinson, Task Force on Statistical Inference, 1999) a později zapracovány do

Z toho důvodu se pokusím postupovat spíše „logicky" a pokusím se vysvětlit základní principy statistické analýzy dat a prezentace jejich výsledků. Je také nutné připomenout si základní cíl vědeckého článku, kterým je prezentace a komunikace vědeckých poznatků. Stejně jako v běžné mezilidské komunikaci, i ve vědecké komunikaci je možné (a často nutné) leccos zamlčet. Ale přesto, že v běžné komunikaci neříkáme vše, ceníme si toho, když je komunikace co nejotevřenější, což můžeme zjednodušeně chápat tak, že v ní není zamlčeno nic *podstatného*. A vědecká komunikace by měla být ještě otevřenější v tom smyslu, že by měla dávat čtenáři možnost odhalit i ty slabiny našeho výzkumu, kterých si my sami nejsme vědomi. Proto by se měla publikace výzkumných výsledků týkat všeho, co *by mohlo být podstatné*.

Je přitom zřejmé, že hranice toho, co komu připadá podstatné, jsou neostré, a autor musí kromě těchto hledisek dbát také omezení týkajících se rozsahu článku. V některých případech pak existuje jediná možnost, jak situaci řešit – autor může nabídnout, že případným zájemcům poskytne další detaily, aby se mohli jeho prací zabývat hlouběji – např. podrobnější výsledky nebo dokonce data apod. Záleží samozřejmě na tom, jak se autor popř. instituce, kde pracuje, s případným zájemcem dohodne.

Mnoho nejasností týkajících se způsobů prezentace výsledků lze vyřešit tak, že se autor pokusí podívat na svůj článek z hlediska čtenáře nebo že ranou verzi svého článku dá přečíst kolegům, kteří se orientují v oboru, ale ne nutně v jeho výzkumech. Jejich otázky a připomínky mohou předejít mnoha kritickým poznámkám a požadavkům recenzenta a značně usnadnit přijetí článku redakcí časopisu.

Důležitá poznámka, kterou je nutné zdůraznit před hlubším ponořením se do statistických témat, obsahuje varování, že statistika je ve vědě pouze jedním z několika důležitých nástrojů. To znamená, že „pouhá" statistika výzkum nespasí. Hodnota každé výzkumné aktivity je závislá na jednotě teorie, použitých metod, výzkumného plánu, (nejen statistického) zpracování dat a interpretace výzkumných zjištění z hlediska teoretických, ale i praktických důsledků. Je však zbytečné, aby byla hodnota výzkumu snížena v důsledku toho, že byly statistické metody nesprávně použity nebo že byly získané výsledky publikovány neadekvátním způsobem.

2. STATISTICKÉ METODY V KOSTCE

Ať se to výzkumníkům v jakémkoli oboru líbí nebo ne, nikomu nemůže uškodit, když si prostuduje některou z učebnic aplikovaných statistických metod (např. v češtině Hebák, Hustopecký, 1987; Hendl, 2004; v dalších jazycích např. Tabachnik, Fidell, 1996; Zöfel, 2003). Ty tento článek nemůže ani v nejmenším nahradit. Přesto stručně uvedu aspoň základní principy, které je nutné brát v úvahu při statistickém zpracování dat a následné prezentaci výsledků.

Většina výsledků statistických analýz se uvádí v *tabulkách* a současně prezentuje v *grafech*. Všechny tabulky by měly obsahovat srozumitelné záhlaví a případné vysvětlivky použitých zkratek, aby se čtenář mohl snadno zorientovat v jejich obsahu. Tabulka je základní formou prezentace výsledků statistických analýz, v mnoha případech ji ale doplňujeme grafem, který umožňuje tabelované výsledky zobrazit ve vzájemných souvislostech tak, že čtenář mnohem snáze pochopí sdělovanou informaci. I zde je nutné dbát na přehlednost a srozumitelnost zobrazených výsledků. Existuje velké množství různých grafů a je nutné zvolit ten, který je pro prezentované výsledky vhodný. Obecná pravidla týkající se volby grafu se obtížně formulují; inspiraci je nutné hledat u autorů publikujících podobné výsledky, jaké máme v úmyslu publikovat my.

Publikace výsledků začíná zpravidla *deskripcí*, popisem získaných dat. Jedná se o popis zkoumaného výběrového souboru a popis zkoumaných proměnných z hlediska centrální tendence[2], variability[3] a tvaru rozložení[4], a to v celém souboru a případných podskupinách, pokud

5. verze Publikačního manuálu APA. Další články se zabývají dopady tohoto počinu na publikace v časopisech APA (např. Fidler et al., 2005). Tyto články a další prameny v nich uváděné nelze případným zájemcům než doporučit.

[2] Centrální tendenci vyjadřujeme pomocí nějaké střední hodnoty, kterou může být např. aritmetický průměr nebo jiný index podle úrovně měření a typu popisované proměnné, např. další formy průměrů (geometrický nebo harmonický) nebo medián, popř. modus.

[3] Variabilitu popisujeme např. pomocí entropie, variačního rozpětí, kvartilového rozpětí, rozptylu, směrodatné odchylky apod.

[4] Jedno z nejdůležitějších rozložení náhodných proměnných je rozložení normální, důležité je ale také rozložení rovnoměrné, exponenciální, chí-kvadrát, rozložení t atd. (viz např. Hendl, 2004; Zöfel, 2003).

je předmětem výzkumu jejich srovnávání. Tím ale získáme představu pouze o charakteristikách výběrového souboru (tzv. *statistikách*). Cílem kvantitativního výzkumu je snaha o zobecnění získaných výsledků vzhledem k relevantnímu základnímu souboru, tzv. populaci (a jejím parametrům). Ponechme zde stranou otázky pořizování výběrových souborů v psychologii a otázky jejich reprezentativnosti např. vůči obecné populaci[5] a soustřeďme se pouze na otázky tzv. inferenční statistiky, nazývané také *statistická indukce* nebo *statistické testování hypotéz*. Ty budou podrobněji zmíněny v následující části článku. Zde se spokojíme s konstatováním, že výsledky statistického testu jsou jiným typem výsledku než prostý popis dat.

Další rozlišení statistických metod je založeno na jejich předpokladech. Tzv. *parametrické metody* už svým názvem naznačují, že jsou založeny na parametrech, obvykle průměru, rozptylu a dalších charakteristikách rozložení dat. To znamená, že právě rozložení dat, a to obvykle *normální rozložení*, je jedním ze základních předpokladů, které musí být splněny, abychom mohli použít parametrické metody. Další předpoklady použití parametrických statistických metod v podstatě logicky souvisejí s předpokladem normálního rozložení – abychom mohli otestovat předpoklad normálního rozložení, potřebujeme dostatečně rozsáhlý výběrový soubor[6]. A abychom mohli vůbec uvažovat o základních parametrech normálního rozložení (kterými jsou průměr a směrodatná odchylka), naše data musí být měřena aspoň na *intervalové úrovni*. To znamená, že v případech, kdy nejsou splněny předpoklady (1) normálního rozložení, (2) dostatečného rozsahu výběrového souboru nebo (3) dostatečné úrovně měření proměnných, používáme *neparametrické metody*, které jsou založeny na mírnějších předpokladech, za což do jisté míry platíme tzv. silou testu (viz např. Cohen, 1969; Urbánek, 1999; Hendl, 2004) neboli pravděpodobností, že test odhalí skutečně existující účinek (efekt). To je problematika, které bude věnována následující část článku.

Posledním rozlišením, které zčásti souvisí s rozlišením metod na popisné a inferenční, je rozlišení *analýz* na explorační a konfirmační. V podstatě jakoukoli metodu lze v kontextu konkrétního výzkumného projektu použít pro hledání teoreticky zajímavých vztahů (neboli pro explorační účely), nebo pro testování předem zformulovaných hypotéz, popř. pro ověřování nějakého teoretického modelu (neboli pro konfirmační účely).

3. STATISTICKÉ TESTOVÁNÍ

Jak píše Cohen v několika svých stěžejních pracích (např. Cohen, 1969, 1994), praxe statistického testování nulových hypotéz (H_0), která se značně rozšířila ve společenských vědách a psychologii, má často podobu jakési kvazi-náboženské aktivity, jejímž účelem a cílem je „posvětit" data tak, že se prokáže statistická významnost výsledků určitých statistických testů. Výsledky prohlášené za „statisticky významné" jsou pak „posvěceny" řadou jedné až tří hvězdiček a považovány za vědecky relevantní (česky např. Urbánek, 1999).

Ironické výroky parafrázované v předchozím odstavci naznačují, že praxe by měla probíhat jiným než popsaným způsobem. První krok k rozpoznání důležitých aspektů této problematiky by mohl následovat na základě uvědomění si faktu, že statistické testy používáme k tomu, abychom získali podporu svých tvrzení.[7] Tuto podporu můžeme získat dvěma způsoby: (1) H_0 se nám díky použití vhodného statistického testu podaří zamítnout – tzv. *podpora zamítnutím*, nebo (2) H_0 se nám ani při použití vhodného statistického testu zamítnout nepodaří – *podpora nezamítnutím*.

Tyto dva způsoby jsou ale zcela odlišné z hlediska logiky hodnocení statistické významnosti výsledků testu. V případě, kdy je naším cílem H_0 zamítnout (což platí ve většině případů), bu-

[5] Což neznamená, že by otázka reprezentativnosti výběrových souborů v psychologii nebyla důležitým tématem. Reprezentativnost je důležitým předpokladem pro většinu statistických metod. Je ale diskutabilní, vzhledem k jakým znakům by měly být výběrové soubory v psychologii reprezentativní.

[6] Na často kladenou otázku, jaký rozsah výběrového souboru je dostatečný, neexistuje jednoduchá odpověď. Pro posouzení této otázky je nutné znát statistický test, který bude prováděn, předpokládanou velikost účinku, kterou se budeme snažit pomocí tohoto testu zjistit, a hodnotu statistické významnosti. Tato hlediska budou stručně objasněna v části 3.

[7] Domnívám se, že toto tvrzení platí bez ohledu na Popperův princip falzifikace, podle kterého je možné tvrzení pouze falzifikovat. To, co je možné na základě statistického testu zamítnout, je pouze nulová hypotéza, a ta je nástrojem sloužícím teorii. Diskuse na toto téma však s tématem tohoto článku souvisí jen okrajově.

deme se snažit udělat vše pro to, aby se tak stalo. Možností máme několik – mezi ty „vědecky čisté" patří např. získání dostatečně rozsáhlého výběrového souboru nebo použití přesnějšího (reliabilnějšího) nástroje měření, posílení vnitřní validity výzkumu apod. V případě, kdy podporu našich tvrzení bude představovat závěr, že H_0 nebylo možné zamítnout, je teoreticky možné připravit z metodologického hlediska problematický projekt a při prezentaci výsledků založených na nemožnosti zamítnout H_0 argumentovat okřídleným (obecně sice pravdivým, v tomto případě však alibistickým) tvrzením, že „i negativní výsledek je výsledkem".

Jediným lékem proti nepoučenému používání statistických testů H_0 je dostatečné pochopení toho, o co se při těchto postupech snažíme. Teprve pochopení souvislostí, které se pokusím objasnit v následujících odstavcích, umožňuje adekvátně reagovat na požadavky a připomínky recenzentů týkající se otázek síly testu, jejichž zohlednění se zejména v renomovaných světových časopisech stává naprostou samozřejmostí.

Jak se může čtenář přesvědčit z mnoha zdrojů a v širších souvislostech, na které v tomto článku není prostor (Cohen, 1969; Urbánek, 1999; Hendl, 2004; StatSoft, Inc., 2006), na praxi statistického testování H_0 se v současné době pohlíží jako na dichotomické rozhodování týkající se našeho názoru na to, v jakém stavu je zkoumaný jev (můžeme ho s trochou nadsázky nazvat „stav světa") v závislosti na dichotomickém výsledku vhodně zvoleného statistického testu. Tuto situaci lze popsat pomocí tab. 1, ve které jsou čtyři políčka obsahující všechny varianty tohoto rozhodnutí.

Tab. 1 Pravděpodobnosti a pojmy u statistického testu (vysvětlení v textu)

		Stav světa	
		H_0	H_A
Rozhodnutí	H_0	Správné přijetí $1 - \alpha$	Chyba II. druhu β
	H_A	Chyba I. druhu Statistická významnost α	Správné zamítnutí Síla testu $1 - \beta$

Je zřejmé, že políčka v hlavní diagonále představují žádoucí rozhodnutí – v případě, že H_0 platí, a my rozhodneme, že platí, je to správné rozhodnutí, a v případě, že H_0 neplatí, a my rozhodneme, že neplatí, je to také správné rozhodnutí. V případě, že H_0 platí a my rozhodneme, že neplatí, dopouštíme se tzv. chyby I. druhu, jejíž pravděpodobnost[8] je označena α, a maximální riziko vzniku této chyby, které jsme ochotni podstoupit, se tradičně nazývá hladina významnosti. Při provádění nějakého statistického testu je získaná hodnota α rovna pravděpodobnosti toho, že za podmínky nebo předpokladu platnosti H_0 získáme taková data, jaká právě analyzujeme. Hladina významnosti tedy neříká nic o platnosti H_0, ale o pravděpodobnosti výskytu analyzovaných dat za předpokladu H_0. Praktický význam této chyby a její pravděpodobnosti tedy spočívá v tom, že „ve světě" neexistuje nějaký vztah nebo rozdíl, např. terapeutický účinek, ale my učiníme závěr, že existuje. Lze tedy říci, že chyba I. druhu představuje svého druhu planý poplach.

Poslední dosud nekomentované políčko tab. 1 představuje druhý nežádoucí případ rozhodnutí na základě výsledků statistického testu – H_0 neplatí (tzn. platí nějaká alternativní hypotéza H_A), ale použitým statistickým testem se nepodaří ji vyvrátit. Pravděpodobnost této chyby se označuje β a její doplněk do hodnoty 1 (nebo 100 %, podle toho, jak jsme zvyklí o pravděpodobnosti uvažovat) neboli $(1 - \beta)$ se nazývá *síla testu*. Tato síla testu je velmi důležitá hodnota pravděpodobnosti[9] jevu, že za předpokladu neplatnosti H_0 *bude* tato hypotéza skutečně zamítnuta. Z praktického hlediska se tedy jedná o pravděpodobnost toho, že zjistíme reálně existující rozdíl nebo vztah, např. terapeutický účinek. A to je zřejmě ta nejdůležitější a vysoce žádoucí vlastnost statistického testu – že jsme schopni s jeho pomocí detekovat účinek, který skutečně existuje.

Pokud uvážíme hodnoty pravděpodobností chyb I. a II. druhu z hlediska dříve uvedeného rozlišení účelu statistického testování na podporu zamítnutím a podporu nezamítnutím, je zjevné,

[8] Přesně řečeno se jedná o podmíněnou pravděpodobnost zamítnutí H_0 za podmínky, že platí.

[9] V případě β se opět jedná o podmíněnou pravděpodobnost – tentokrát jevu, že H_0 nebude zamítnuta, za podmínky, že neplatí. Hodnota $(1 - \beta)$ je pak podmíněná pravděpodobnost jevu, že H_0 bude zamítnuta za podmínky, že neplatí.

že v případě podpory zamítnutím se zdá být důležitější udržet nízkou pravděpodobnost chyby I. druhu (tzn. hodnotu α), aby se předešlo „planému poplachu", zatímco v případě podpory nezamítnutím je zcela zásadní dosáhnout nízké hodnoty rizika chyby II. druhu (hodnoty β), abychom neplýtvali prostředky na zbytečný výzkum.

Z uvedených principů plyne, že jestliže se budeme snažit minimalizovat riziko chyby jednoho druhu, poroste riziko chyby druhého druhu, podobně jako se může stát, že v případě podezření na vážnou chorobu budeme léčit zdravého člověka (chyba I. druhu), nebo naopak ve snaze ušetřit za zdravotní péči (nebo neztrácet čas nebo peníze maroděním) nebude léčen skutečně nemocný člověk (chyba II. druhu).

V zásadě neexistuje důvod pro použití nějakých předem stanovených mezních hodnot pro velikosti chyb obou druhů. V praxi se však ustálil požadavek, aby bylo riziko chyby I. druhu menší než 0,05 (tzv. 5% hladina významnosti) nebo v případech, kdy chceme být přísnější, 0,01. Je ale možné v odůvodněných případech slevit i na hodnotu 0,10 – např. pokud máme zajištěnou dostatečnou sílu testu, naše data jsou vzácná a obtížně jsme je získávali atd. Maximální přípustná velikost rizika chyby II. druhu se uvádí od hodnoty 0,20 (tzn. síla testu je 80 %) nebo přísnější 0,10 (síla testu 90 %). I zde ale záleží spíše na zvážení rizik než na nějaké ostře definované hranici a je možné akceptovat i vyšší hodnoty.

Dosažení nízkých úrovní rizika obou typů chyb je specifické pro každý statistický test a závisí na dvou důležitých veličinách: Jsou jimi jednak rozsah výběrového souboru (N), pro který obecně platí, že s růstem N se zvyšuje síla testu, takže roste šance zamítnutí H_0, a jednak tzv. velikost efektu nebo účinku (*effect size*), což je jeden z nejjednodušších ukazatelů „skutečné", „věcné" nebo „teoretické" významnosti získaného výsledku, v kontrastu k „pouhé" statistické významnosti[10]. Pro různé statistické testy se tato velikost účinku počítá různě – např. pro rozdíl dvou výběrových průměrů se počítá jako tento rozdíl dělený společnou variabilitou obou podsouborů vyjádřenou pomocí směrodatné odchylky.

I na základě laického posouzení by nám mělo být zřejmé, že je poměrně snadné prokázat existenci výrazného, velkého účinku. K prokázání takového účinku nám často stačí i poměrně malý výběrový soubor – samozřejmě, pokud byl výběr proveden ve shodě s předpoklady prováděného statistického testu. Oproti tomu, pokud je velikost účinku, který se pomocí statistického testu snažíme prokázat, malá, musíme se snažit získat rozsáhlý výběrový soubor a provést výzkum vyznačující se značnou mírou vnitřní validity pomocí vysoce reliabilních a validních měřících nástrojů.

Zásadní problém s velikostí účinku přitom v mnoha případech tkví v tom, že o ní víme velmi málo. Zvláště v případech, kdy používáme statistické testy k účelům explorační analýzy dat, nemůžeme mít představu o tom, jak velké účinky se pomocí nich pokoušíme zjišťovat. Takový výzkum je pak poměrně obtížné naplánovat, protože pro detekci výrazných účinků stačí malé rozsahy výběrových souborů, ale pro detekci slabých účinků jsou obecně potřebné velké rozsahy výběrových souborů. Jako vodítko navrhnul Cohen (1969) tři úrovně velikosti účinku pro různé statistické testy – např. pro porovnání dvou výběrových průměrů t-testem jsou to velikosti účinku[11] s hodnotami 0,2, 0,5 a 0,8. Tyto úrovně jsou orientační a lze je využít právě při plánování výzkumů[12].

Na závěr této části je třeba zmínit se o intervalech spolehlivosti, které jsou kromě velikostí účinků další vysoce žádoucí formou prezentace výsledků statistických analýz (a to jak v tabulkách, tak v grafech). Intervaly spolehlivosti vlastně uvádějí rozsah, v jakém se vyskytuje se stanovenou pravděpodobností testovaná statistika. Kromě toho, že z nich můžeme vyčíst výsledek testování H_0 (pokud např. interval spolehlivosti pro rozdíl dvou výběrových průměrů obsahuje 0, nelze na stanovené hladině významnosti zamítnout H_0 o rovnosti obou průměrů), můžeme také posoudit, jestli byly zkoumané statistiky změřeny s dostatečnou přesností. Pokud jsou tedy

[10] Nový Publikační manuál APA a některé časopisy APA v současné době prosazují také diskusi o věcné významnosti získaných výsledků (viz např. Fidler et al., 2005). Tu je možné posuzovat v jakémkoli typu výzkumu, např. v klinickém výzkumu se diskutuje o tzv. minimální klinicky významné změně (Mareš a Urbánek, 2006).

[11] Velikost účinku v případě t-testu se počítá podle vzorce $\delta = \frac{m_1 - m_2}{s}$, kde m_1 a m_2 jsou hypotetické průměry a s je hypotetická společná směrodatná odchylka obou skupin.

[12] V některých komerčních statistických programech už dnes najdeme nástroje na provádění různých forem analýzy síly testu (např. Statistica). V nekomerční sféře je možné vybavit se freewarovým programem z internetu jako je např. G*Power (Erdfelder, Faul, Buchner, 1996; www.psycho.uni-duesseldorf.de/aap/projects/gpower/).

velikosti účinků a intervaly spolehlivosti uváděny v článcích (nebo je možné je z publikovaných údajů aspoň vypočítat *ex post*), je možné takové údaje potom zpracovávat v meta-analytických studiích (viz např. Hendl, 2004).

4. SCHÉMA STATISTICKÉHO TESTU

Ale co je to vlastně statistický test? Opět bych rád zdůraznil, že na tuto otázku nám nejlépe odpoví různé učebnice statistiky citované v části 2 nebo specializované knihy (např. Kanji, 1993), ve kterých si uživatel může vybrat vhodný test podle toho, jakou statistiku, v jakých souborech a za jakých podmínek chce testovat.

Obecně je princip všech statistických testů obdobný a jejich uživatel by měl projít následující kroky, které se obvykle uvádějí v literatuře (i když jejich seznam není vždy stejný):

1. Ověření předpokladů testu
2. Formulace nulové hypotézy
3. Formulace alternativní hypotézy
4. Výpočet testového kritéria
5. Výpočet p-hodnoty
6. Rozhodnutí o platnosti/neplatnosti H_0

V minulosti se namísto výpočtu p-hodnoty počítala tzv. kritická hodnota testového kritéria odpovídající předem zvolené hodnotě hladiny významnosti. V mnoha učebnicích se popisuje, jak tuto kritickou hodnotu najít ve speciálních tabulkách. Pokud by vypočtená hodnota testového kritéria byla vyšší než tato kritická hodnota, H_0 by byla na základě výsledků testu zamítnuta na zvolené hladině významnosti. Výpočet p-hodnoty slouží témuž účelu, jen o něco přesněji vyjadřuje pravděpodobnost získání analyzovaných dat za předpokladu platnosti H_0.

V mnoha pramenech věnovaných tomuto tématu se píše, že hladina významnosti (tzn. povolená míra rizika chyby I. druhu) má být zvolena před provedením výzkumu. Současně by každý, kdo provádí výzkum, měl mít představu o síle testu (tzn. pravděpodobnosti zjištění nějakého účinku za předpokladu, že je v datech skutečně přítomen v podobě naměřených hodnot). K tomu ovšem potřebuje mít aspoň orientační představu o velikosti účinku, který se snaží prokázat (nebo jehož existenci se snaží vyloučit). Na základě těchto úvah pak stanoví rozsah výběrového souboru *N*. Tyto úvahy ale mají předcházet provádění výzkumu, který provedením statistického testu v jistém smyslu vrcholí. Z toho důvodu také tyto úvahy nezařazuji do seznamu kroků statistického testu jako takového – představují jakýsi širší soubor podmínek nebo kontext pro jeho provádění.

Na závěr tohoto oddílu je ještě nutné zmínit se o tom, jaký je rozdíl mezi statistickou metodou a statistickým testem. Pojem metody je širší. Součástí statistické metody může být kromě různých výpočtů, které nejsou statistickými testy, také jeden nebo více statistických testů sloužících pro určitá dílčí rozhodnutí týkající se volby dalšího postupu nebo verdiktu o dílčích výsledcích (a výsledku celkovém).

5. PŘÍKLADY

V dalších částech článku bych se chtěl věnovat příkladům statistických metod se zřetelem k uvádění výsledků analýz získaných s jejich pomocí. Tento seznam nemůže být úplný, protože statistických metod jsou desítky.

5.1. Porovnání průměrů t-testem

Dá se říci, že t-test je jednou z nejklasičtějších statistických metod, za jejíž vznik vděčíme pracovníku pivovaru Guinness Gossetovi (test publikoval pod pseudonymem Student), který se zabýval kontrolou jakosti piva (Zöfel, 2003). Jedná se o test, s jehož pomocí se pokoušíme porovnat dva výběrové průměry[13]. Nulová hypotéza má tedy v tomto případě podobu: H_0: $m_1 = m_2$ (alternativní: H_A: $m_1 \neq m_2$). Mezi předpoklady t-testu patří normalita rozložení v obou porovnávaných podskupinách. Rozsahy obou podskupin by měly být v ideálním případě zcela stejné nebo co nejpodobnější, ale můžeme se i dočíst, že přípustná je nerovnost rozsahů až do poměru

[13] Kromě t-testu pro porovnávání dvou průměrů nezávislých výběrů existuje také t-test pro porovnání jednoho výběrového průměru s nějakou vybranou hodnotou a také t-test pro porovnání párových hodnot – např. průměry hodnot získaných od manželských párů nebo dvojčat nebo před a po skončení terapie apod.

5:1 (Tabachnik a Fidell, 1996). Dalším důležitým předpokladem je homogenita rozptylů, která se ověřuje pomocí Levenova nebo Bartlettova testu. V případě zamítnutí H_0 o rovnosti rozptylů na základě některého z těchto testů se výpočty provádějí podle modifikovaných vzorců. Některý statistický software (např. SPSS) počítá výsledky oběma způsoby současně a uživatel se na základě výsledku Levenova testu rozhoduje, který výsledek použije.

K často řešeným otázkám patří, jak velký má být soubor pro provedení t-testu. Z předchozích odstavců by mělo být jasné, že odpověď na tuto otázku závisí na několika věcech. Výsledky analýzy síly testu nám napovídají, že v případě, že bychom chtěli pomocí oboustranného t-testu a při zvolených hodnotách $\alpha = 0{,}05$ a $\beta = 0{,}20$ prokázat existenci velkého účinku[14] (0,8), potřebovali bychom celkové $N = 52$, pro střední účinek (0,5) a stejné parametry projektu by to bylo $N = 128$ a pro malý účinek (0,2) a stejné parametry by to bylo $N = 788$.

T-test je založen na hodnotě *t* (a rozložení *t*) pro počet stupňů volnosti[15] $df = N - 2$. Při uvádění výsledků je tedy logicky nutné uvést ty hodnoty, které se vztahují k předpokladům a výsledkům testu. Jedná se o celkovou velikost výběrového souboru (N), velikost obou podskupin (N_1 a N_2), průměry (m_1 a m_2) a směrodatné odchylky (s_1 a s_2) v jednotlivých podskupinách, hodnota statistiky *F* pro Levenův test a příslušná p-hodnota, hodnota statistiky *t*, počet stupňů volnosti pro tento t-test ($N - 2$) a jeho p-hodnota.

Pokud je p-hodnota pro hodnotu *t* nižší než nějaká dostatečně nízko stanovená mez, říkáme, že výsledek t-testu je statisticky významný. Tento závěr činíme ve skutečnosti na základě úvahy, že pokud platí H_0 (tzn. průměry v obou skupinách se neliší), je pravděpodobnost toho, že bychom získali taková data, jaká právě analyzujeme, nižší než 5%[16].

Čtenář, který rozumí tomu, jak se provádí tento test, tedy může na základě výše uvedených údajů posoudit např. to, jaká byla aktuální síla použitého t-testu (pokud to autor sám neuvede).

5.2 Analýza rozptylu

Analýza rozptylu (ANOVA – zkráceně z anglického *analysis of variance*) je metoda pro porovnání více než dvou výběrových průměrů. V nejjednodušším případě tzv. jednorozměrné nebo jednofaktorové analýzy rozptylu, kdy se porovnávají průměry v několika skupinách definovaných jako klasifikace na základě jedné (nominální nebo ordinální) proměnné, se jedná o jednoduché rozšíření t-testu. Ve složitějších případech, kdy se používá více faktorů, které se vzájemně kombinují, už se v podstatě jedná o komplexní zobecnění t-testu – o nástroj pro posuzování statistické významnosti rozdílů mezi úrovněmi jednotlivých faktorů a jejich vzájemných interakcí.

Dá se říci, že o analýze rozptylu a jejím zobecnění, tzv. obecném lineárním modelu, byly napsány celé knihy, ve kterých je možné dozvědět se o možných variantách metody a postupech použitelných ve speciálních případech např. porušení určitých předpokladů. Zde je tedy možné zmínit se pouze o základních předpokladech a principech této metody.

V ideálním případě by měla být analýza rozptylu prováděna u normálně rozložených dat jednotlivých podskupin, které by měly být na sobě vzájemně nezávislé a měly by mít stejnou variabilitu. Podskupiny v každém z faktorů by měly být stejně velké. To vyplývá z faktu, že původně byla tato metoda vytvořena pro analýzu experimentálních dat, kde se tyto podmínky dají relativně snadno kontrolovat. Protože mnoho dat (nejen) v psychologii nepochází z experimentálních výzkumů, byly vyvinuty i další postupy a testy.

Základní myšlenka analýzy rozptylu je ale společná všem různým přístupům a spočívá v úvaze, že jestliže se mezi sebou jednotlivé analyzované podskupiny neliší z hlediska průměrných hodnot (H_0 má tedy podobu $m_1 = m_2 = \ldots = m_k$) a variability, pak variabilita v rámci jednotlivých podskupin je stejná jako variabilita mezi těmito podskupinami. Tyto dva odhady variability se provádí pomocí součtu čtverců odchylek jednotlivých hodnot od skupinového průměru v každé z podskupin a součtu čtverců odchylek skupinových průměrů od celkového průměru (přičemž odchylka průměru každé podskupiny je v tomto výpočtu zastoupena tolikrát, kolika hodnotami je daná podskupina tvořena).

Výsledkem těchto analýz pro jeden faktor je hodnota *F* s dvěma počty stupňů volnosti – jeden se vypočítá jako počet podskupin bez jedné a druhý jako počet respondentů bez počtu skupin.

[14] Jak již bylo uvedeno, v případě t-testu se velikost účinku počítá jako absolutní hodnota rozdílu hypotetických výběrových průměrů dělená celkovou směrodatnou odchylkou.

[15] Tento počet stupňů volnosti se používá, pokud platí předpoklad homogenity rozptylu. Za podmínky nehomogenity rozptylu se používá modifikovaný vzorec (viz např. Kanji, 1993; Hendl, 2004).

[16] Přesněji řečeno, tato pravděpodobnost je rovna p-hodnotě pro tento t-test.

Pro tuto hodnotu se vypočítá hladina významnosti na základě rozložení F s příslušnými počty stupňů volnosti – neboli provede se tzv. F-test.

Pochopitelně i v případě analýzy rozptylu je vysoce relevantní otázka síly testu. Protože ale není možné dát jednoduchá vodítka nebo pravidla týkající se např. nutné velikosti podsouborů (úvahy, které je nutné provést, jsou aspoň stručně naznačeny v případě t-testu spolu s představou o tom, o jak velké počty osob se jedná v případě různých velikostí účinku), je nutné zájemce o tuto problematiku odkázat na specializovaný software[17].

Jestliže je p-hodnota pro hodnotu F (F-test) menší nebo rovna 0,05 (pokud zvolíme hladinu významnosti 5 %), znamená to, že za předpokladu platnosti H_0 (tzn. hypotézy, že průměry ve všech podskupinách jsou stejné) máme méně než 5% pravděpodobnost, že bychom získali taková data, jaká právě analyzujeme. Na základě tohoto zjištění je zvykem (podobně jako v případě t-testu) učinit závěr, že H_0 se zamítá na 5% hladině významnosti.

Protože výsledky tohoto tzv. celkového testu nejsou bez přihlédnutí k detailům interpretovatelné, nutným doplňkem výsledků jsou hodnoty průměrů v jednotlivých podskupinách definovaných kombinacemi úrovní jednotlivých faktorů. Počet těchto podskupin roste s počtem faktorů a s počtem jejich úrovní. Zpravidla bývá obtížné říci, které rozdíly mezi průměry konkrétních podskupin jsou důležité a které ne. V takových případech je pak možné provést tzv. „*post hoc* testy", které lze použít k otestování statistické významnosti rozdílů mezi průměry jednotlivých dvojic podskupin. Těchto testů existuje celá řada, jejich použití je však spojeno[18] s tzv. „inflací hladiny významnosti". Tento problém je způsoben faktem, že pokud provedeme velký počet testů s týmiž daty (dejme tomu 10), pravděpodobnost chyby I. druhu roste. Pokud by tedy riziko chyby I. druhu bylo v případě každého testu rovno hodnotě α, pak riziko (pravděpodobnost) toho, že chyba I. druhu nastane aspoň u jednoho z těchto deseti testů se dá vypočítat na základě binomické věty jako $1-(1-\alpha)^{10}$. V případě běžně používané hladiny významnosti 0,05 to dělá asi 0,40, v případě „přísnější" nebo „konzervativnější" hladiny 0,01 je to necelých 0,10. Pokud první z výsledků vyjádříme slovně, znamená to, že při deseti testech na hladině významnosti 0,05 máme pravděpodobnost 0,4, že aspoň jeden z těchto testů bude statisticky významný, přestože ve skutečnosti se žádná dvojice průměrů mezi sebou neliší[19].

Pokus o řešení tohoto problému představuje např. tzv. Bonferroniho korekce, která u každého testu dělí vypočtenou p-hodnotu počtem provedených testů. Tato, ale ani žádná z metod však nejsou z metodologického a statistického hlediska ideální, a proto Cohen a Cohen(ová) (Cohen a Cohen, 1983) doporučují provést pouze tzv. Fischerův „chráněný t-test", což je právě postup naznačený výše – pokud bude výsledek F-testu statisticky významný, je možné provést *post hoc* testy bez jakékoli korekce, ale s vědomím, že výskyt statisticky významných výsledků čistě náhodně, bez toho, že by skutečně existovaly rozdíly mezi jednotlivými průměry (tzn. výskyt chybných rozhodnutí I. druhu) je pravděpodobnější než obvykle. Pokud bude výsledek celkového F-testu statisticky nevýznamný, *post hoc* testy se na základě tohoto principu provádět nemají.

U výsledků analýzy rozptylu je nezbytně nutné uvádět popisné statistiky (průměry a směrodatné odchylky v jednotlivých podskupinách spolu s údaji o jejich rozsahu), aby bylo možné posoudit, jak se liší hodnoty v jednotlivých podskupinách. Dále by měl být uveden výsledek testu homogenity rozptylu (podobně jako v případě t-testu) a výsledek F-testu pro celý model a případně pro jednotlivé faktory (tzn. vždy hodnota F, obě hodnoty stupňů volnosti a příslušná p-hodnota). Pokud by to bylo žádoucí, je možné provést tzv. *post hoc* testy a uvést výsledky těch z nich, které byly statisticky významné (tzn. hodnotu t a její p-hodnotu).

5.3 Pearsonův korelační koeficient

Pearsonův korelační koeficient je v psychologii velmi často používaným indexem lineárního vztahu mezi dvěma proměnnými. Je to index, který vyjadřuje těsnost tohoto vztahu na škále ‹-1; 1›, kde extrémní hodnoty znamenají dokonalý záporný (r = -1) nebo kladný (r = 1) vztah, kdy by v tzv. scatterplotu neboli korelačním grafu všechny body ležely na přímce, a r = 0 znamená naprostou absenci vztahu, kdy by všechny body ležely na ploše kruhu (nebo elipsy).

V literatuře (např. Cohen a Cohen, 1983) se uvádí, že korelační koeficient není příliš citlivý k porušení předpokladu normality rozložení. Je ale nutné dát pozor na výskyt tzv. odlehlých případů (*outliers*), které by mohly hodnotu korelace značně vychýlit.

U korelačního koeficientu se nejčastěji provádí test významnosti odchylky od 0, který je založený na rozložení t (Cohen a Cohen, 1983) s počtem stupňů volnosti $df = N-2$. Testovaná H_0 má tedy v tomto případě podobu r = 0. Výpočet hodnoty t se provádí na základě vzorce (1).

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \qquad (1)$$

Pro tento t-test se stanoví p-hodnota a zjistí se tak pravděpodobnost chyby I. druhu. Velikost souboru, kterou je nutné zajistit pro dostatečnou sílu tohoto testu (aspoň 0,8), pro $\alpha = 0{,}05$ a pro střední velikost účinku, kterou představuje hodnota r = 0,3, je rovna N = 82.

U korelačního koeficientu je to s uváděním výsledků poměrně jednoduché – stačí velikost korelačního koeficientu, velikost souboru a p-hodnota. Je ale třeba mít na paměti, že v případě exploračního použití korelační analýzy, podobně jako v případě s post *hoc testy* u analýzy rozptylu, velký počet provedených testů (vypočtených korelací a jejich p-hodnot) vede k „inflaci" statistické významnosti.

Další statistické metody jsou uvedeny v druhé části tohoto článku.

[17] Viz pozn. pod čarou 12.

[18] Tento problém se netýká pouze post hoc testů u analýzy rozptylu, ale obecně situace, kdy provádíme velký počet testů u stejných dat, např. výpočtu velkého počtu korelačních koeficientů v exploračním výzkumu.

[19] Tyto otázky jsou velmi zajímavé i z hlediska síly testu. Případným zájemcům lze doporučit např. Cohen a Cohen (1983) nebo Cohen (1969).

LITERATURA

Cohen, J. (1969): Statistical Power Analysis for the Behavioral Sciences. New York, Academic Press.

Cohen, J. (1994): The Earth is Round (p < .05). American Psychologist, 49, 12, 997-1003.

Cohen, J., Cohen, P. (1983): Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Erdfelder, E., Faul, F., Buchner, A. (1996): GPOWER: A General Power Analysis Program. Behavior Research Methods, Instruments, & Computers, 28, 1, 1-11.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C., Schmitt, R. (2005): Toward Improved Statistical Reporting in the Journal of Consulting and Clinical Psychology. Journal of Consulting and Clinical Psychology, 73, 1, 136-143.

Hebák, P., Hustopecký, J. (1987): Vícerozměrné statistické metody s aplikacemi. Praha, SNTL, Alfa.

Hendl, J. (2004): Přehled statistických metod zpracování dat. Praha, Portál.

Kanji, G. K. (1993): 100 Statistical Tests. London, Sage.

Mareš, J., Urbánek, T. (2006): Minimální věcně významné změny v diagnostikované kvalitě života. Československá psychologie, 50, 6, 557-568.

StatSoft, Inc. (2006): Electronic Statistics Textbook. Tulsa, OK: StatSoft. WEB: http://www.statsoft.com/textbook/stathome.html. Vyhledáno 30.5.2006.

Tabachnik, B. G., Fidell, L. S. (Eds.) (1996): Using Multivariate Statistics. New York, Harper Collins College Publisher.

Urbánek, T. (1999): Problematika síly testu v kontextu kvantitativního výzkumu v psychologii. In: Heller, D., Sedláková, M., Vodičková, L. (Eds.): Kvantitativní a kvalitativní výzkum v psychologii. Praha, Psychologický ústav AV ČR, 63-71.

Wilkinson, L., Task Force on Statistical Inference (1999): American Psychologist, 54, 8, 594-604.

Zöfel, P. (2003): Statistik für Psychologen im Klartext. München, Pearson Studium.

SOUHRN

Tento článek uvádí některé obecné zásady týkající se publikace výsledků statistických analýz empirických dat. V první části se věnuje spíše obecným zásadám a principům, v druhé části několika konkrétním statistickým testům, metodám a přístupům, vysvětlení jejich provádění a uvedení výsledků, se kterými by měl být čtenář seznámen, aby mohl pochopit autorovy závěry.