NULOVÁ HYPOTÉZA VERSUS INTERVALY SPOLEHLIVOSTI

NOCAR David – ZDRÁHAL Tomáš, CZ

Resumé

Příspěvek se zabývá důvodem, proč by se výzkumní pracovníci a vědci měli při statistickém zpracování svých výzkumů raději zaměřit na odhady založené na velikosti účinku a intervalech spolehlivosti než na testování nulových hypotéz. Základním problémem testování nulové hypotézy je to, že podporuje dichotomických úvahy - odmítnout, nebo neodmítnout tuto hypotézu. V článku je vysvětleno, proč jsou úvahy založené na odhadech intervalů spolehlivosti vhodnější - že poskytují více informací než testování nulové hypotézy. Na druhé straně však mohou být tyto intervaly příliš dlouhé, a to proto, že nemáme v praxi možnost provést měření na dostatečně velkém výběrovém souboru. Dá se tomu ale vyhnout opakovanými šetřeními.

Klíčová slova: testování nulové hypotézy, velikost účinku, interval spolehlivosti.

NULL HYPOTHESIS VERSUS CONFIDENCE INTERVALS

Abstract

The paper deals with the reason why researchers and scientists should shift emphasis in the statistical treatment of their research as much as possible from NHST (Null Hypothesis Significance Testing) to estimation based on ESaCI (Effect Sizes and Confidence Intervals). A basic problem of NHST is that it encourages dichotomous reasoning, which refers to the reject-or-do not-reject mindset that is widespread within the NHST theories. We would like to explain why a shift to estimation thinking is likely to be valuable, and provide evidence that confidence ESaCI can be more informative than NHST; ESaCI provide more complete information. On the other hand confidence intervals could be disappointingly long. They are long because it is often neither practical nor possible to use sufficiently large samples to get shorter confidence intervals. A way to avoid this disadvantage is to combine results from multiple studies to get better estimates.

Key words: null hypothesis significance testing, effect sizes, confidence interval.

Úvod

Testování hypotéz a intervaly spolehlivosti představují technicky v podstatě totéž, avšak každá metoda sleduje něco jiného: Zatímco se test nulové hypotézy týká hodnocení platnosti pravděpodobnostního modelu, který popisuje chování náhodné veličiny, tak konstrukce intervalů spolehlivosti má za cíl charakterizovat přesnost bodového odhadu neznámého parametru. Vždy je však nutné, aby v každé výzkumné zprávě byla uvedena vedle výsledku testu (rozhodnutí o platnosti *nulové hypotézy*) i velikost dosaženého rozdílu (efektu) s příslušným intervalem spolehlivosti. Ze samotné *p*-hodnoty zvoleného testu nebo rozhodnutí o zamítnutí či nezamítnutí nulové hypotézy totiž není zřejmé, v jakých mezích se pozorovaná velikost rozdílu (účinku) pohybuje. Je svým způsobem paradoxní, že ačkoli stále více a více vědeckých pracovníků poukazuje na tyto omezující limity procesu testování nulové hypotézy a na výhody práce s intervaly spolehlivosti, pořád v mnoha vědních oborech, jmenovitě se to týká především psychologie, pedagogiky, sociologie a medicíny, mají vědci pocit, že je jejich povinností testovat nulovou hypotézu. Velikost účinku a intervaly spolehlivosti však poskytují, jak již bylo uvedeno výše, úplnější informace a jejich použití (umožněné především díky účinnému statistickému softwaru) pomůže výzkumným pracovníkům lépe prezentovat výsledky svých bádání. Protože však není často praktické ani možné použít dostatečně velký výběr a intervaly spolehlivosti jsou pak příliš dlouhé

(a tím se výhoda jejich použití eliminuje), je nutné intervaly zkrátit, a to kombinací výsledků z několika studií. Tento postup je v praxi velmi nákladný, takže se většinou neprovádí.

1 Teorie náhodných výběrů

Díky speciálně vytvořenému programu *Kombinatorika vyšší dimenze* (autorem je externí řešitel projektu Univerzity Palackého v Olomouci *IGA PdF UPOL 2015 005 Student Grant Competition* R. Baran z UJEP v Ústí nad Labem) lze vygenerovat vždy celou populaci (základní soubor) a také tolik výběrů, kolik potřebujeme - tj. dostatečný počet studií), a proto můžeme poznat, jak přesné závěry z různých výběrů jsme schopni udělat a toto porovnat s výsledky testování nulové hypotézy. Tímto vlastně zopakujeme postup, jak teorie náhodných výběrů vznikala:

Ze známé celé populace bylo vybráno velké množství náhodných výběrů (experimentů) a zkoumalo se, jak z výběrových statistik můžeme usuzovat na charakteristiky základního souboru. A protože tato populace byla zcela známa, byly známé i tyto charakteristiky a mohla být tak změřena účinnost testování nulové hypotézy a intervalů spolehlivosti.

Ujasněme si nyní, jaký je (teoretický) vztah mezi výsledkem statistického testu a intervalem spolehlivosti. Předně, statistický test zamítne nulovou hypotézu právě tehdy, když testovaný parametr nenáleží do příslušného intervalu spolehlivosti. Dále, interval spolehlivosti opravdu uvádí více informací než statistický test, neboť test nulové hypotézy nám na rozdíl od intervalu spolehlivosti nedává informaci o rozsahu studovaného vlivu. Výsledkem testování je pouze pravděpodobnost výsledku experimentu za předpokladu platnosti nulové hypotézy (test zdůrazňuje chybu prvního druhu, ale říká velmi málo o chybě druhého druhu); přitom chceme znát míru platnosti alternativní hypotézy, je-li dán výsledek experimentu. Výsledkem odhadu intervalů spolehlivosti je i odhad chyby druhého druhu - pokud totiž necháme chybu prvního druhu pevnou a zmenšujeme chybu druhého druhu, tak tak interval spolehlivosti zmenšuje svou délku.

2 Chyba prvního a chyba druhého druhu

Začněme ilustrativním příkladem.

V dílně Deset se vyrobilo celkem 20 výrobků a z nich bylo 10 vadných, v dílně Pět bylo z celkem 20 výrobků 5 vadných. Závadnost se dá zjistit jen časově velmi náročnými zkouškami - dejme tomu, že je únosné testovat jen 2 výrobky. Dá se nějak určit, z které dílny dodaná série 20 výrobků pochází? A když ano, jakého omylu, jaké chyby se dopustíme, když na základě vybrané dvojice rozhodneme o celé 20 prvkové sérii?

Řešení

Uvažujeme takto: Vybereme náhodně ze série, kterou jsme obdrželi, 2 výrobky. Když mezi nimi nebude ani 1, budeme usuzovat, že série pochází z dílny Pět. Bude-li vadný výrobek 1 či budou-li vadné oba 2, přisoudíme výrobky dílně Deset. Tato úvaha se zdá docela logická, neboť v dílně Deset je každý druhý výrobek vadný a vybíráme jen dvojici výrobků.

Tím jsme odpověděli na první otázku: Ano, dá se to (nějak) určit. To se dá ale vlastně vždy - takové kritérium si může vymyslet každý. Hlavně nás tedy zajímá nás, jaká je pravděpodobnost omylu, chyby.

Je třeba si pořádně uvědomit, že mohou nastat hned dva různé omyly, a to že

1. série pochází z dílny Pět, ale my se rozhodneme pro dílnu Deset a
2. série pochází z dílny Deset, ale my ji přisoudíme dílně Pět.

Podívejme se detailněji nejprve na případ 1.

Prvního omylu (chyby prvního druhu) se tedy dopustíme, když usoudíme, že výrobky jsou ze série vyrobené ve firmě Deset, ačkoliv pocházejí z dílny Pět. Podle naší úvahy se tak stane tehdy, když počet vadných výrobků v bude větší nebo roven 1, a když parametry hypergeometrického rozdělení (o které se zde zřejmě jedná) budou dány sérií výrobků z dílny Pět.

Pravděpodobnost, že ze dvou náhodně „vytažených“ výrobků bude vadný 1, je rovna

$$\frac{\binom{5}{0}\binom{15}{2}}{\binom{20}{2}};$$

použitím následující „excelovské“ funkce dostáváme:
HYPGEOM.DIST(1;2;5;20;0) ≐ 0,3947. Pravděpodobnost, že ze dvou náhodně „vytažených“ výrobků budou vadné oba 2 je rovna analogicky HYPGEOM.DIST(2;2;5;20;0) ≐ 0,0526.
Pravděpodobnost, že počet vadných výrobků v sérii bude $v \geq 1$ je tedy rovna asi 0,3947 + 0, 0526 = 0,4473 ≐ 0,447 = 44,7%; tutéž hodnotu získáme pochopitelně i takto:
1 - HYPGEOM.DIST(0;2;5;20;0) ≐ 0,447.

Co tato hodnota vlastně znamená, tedy co si máme představit pod formulací „chyba prvního druhu je α“? (Zde je α = 0,447.) Znamená, že s pravděpodobností α zamítneme správné tvrzení. V našem případě je správným tvrzením skutečnost, že naše série pochází z dílny Pět; my na základě přezkoumání 2 výrobků a zjištění, že aspoň 1 je vadný, rozhodneme, že z dílny Pět není a dopustíme se tím chyby 0,447. Kdybychom tedy prozkoumali úplně všechny možné dvojice výrobků z naší série z dílny Pět, byl by podíl počtu těch dvojic, které obsahují alespoň 1 vadný výrobek, k počtu všech dvojic roven 0,447. Problém je v tom, že ty všechny možné dvojice musíme znát. V tomto ilustrativním příkladu to takový problém není a později to uděláme, ale v praktických situacích kombinace bez opakování *k*- té třídy z *n* prvků nejrůznějšího charakteru pro větší *k* (větší než 30) a velká *n* (řádově stovky, tisíce, desetitisíce,…) určit všechny možné *k*- tice (ve statistické terminologii výběrové soubory) nedokážeme. Smyslem projektu IGA je (jak již bylo shora naznačeno) toto pro rozumně velké hodnoty provést a ukázat tak, jak dobře se na výsledky testování nulových hypotéz můžeme spolehnout.

Pokračujme případem 2.
Druhého omylu (chyby druhého druhu) se dopustíme, když usoudíme, že výrobky jsou ze série vyrobené ve firmě Pět, ačkoliv pocházejí z dílny Deset. To se stane tehdy, když počet vadných výrobků *v* bude roven 0, $v = 0$.
Pravděpodobnost, že ze dvou náhodně „vytažených“ výrobků nebude vadný žádný, je rovna

$$\frac{\binom{10}{0}\binom{10}{2}}{\binom{20}{2}};$$

jde o hodnotu HYPGEOM.DIST(0;2;10;20;0) ≐ 0,237 = 23,7 %.
Opět otázka: Co tato hodnota vlastně znamená, tedy co si máme představit pod formulací „chyba druhého druhu je β“? (Zde je β = 0,237.) Znamená, že s pravděpodobností β přijmeme špatné tvrzení. V našem případě je špatným tvrzením skutečnost, že naše série pochází z dílny Pět (správným tvrzení je fakt, že série pochází z dílny Deset); my na základě přezkoumání 2 výrobků a zjištění, že žádný není vadný, rozhodneme, že je z dílny Deset a dopustíme se tím chyby 0,237. Opět kdybychom tedy prozkoumali úplně všechny možné dvojice výrobků z naší série z dílny Deset, byl by podíl počtu těch dvojic, které neobsahují žádný vadný výrobek, k počtu všech dvojic roven 0,237.

Uvedený příklad byl záměrně velmi jednoduchý. Chceme na něm pouze „ověřit“, že náš postup při testování hypotéz (o nic jiného se tady prakticky nejednalo) a při určování intervalů spolehlivosti, založený na tom, že máme k dispozici celý základní soubor a všechny výběrové soubory, vede a musí vést ke stejným výsledkům, které získáme „teoretickým“ postupem. (V praxi však teoretický postup použít pochopitelně nemůžeme, takže musíme postupovat tak, jak je uvedeno dále; je zde však omezení pro větší základní soubory.) Již bylo shora uvedeno, že takto celá teorie náhodných výběrů vznikala: Ze známého základního souboru bylo vybráno velké množství výběrových souborů a zkoumalo se, jak můžeme z výběrových statistik usuzovat na

charakteristiky základního souboru. Pochopitelně ideální by bylo mít k dispozici VŠECHNY výběrové soubory, ale to je problém velmi těžce řešitelný i v současné době superpočítačů - stačí, aby základní soubor byl řádově roven desetitisícům a počet výběrů obvyklého rozsahu (20 - 100) bude…zkrátka VELIKÝ.

Vyřešme tedy teď náš příklad „experimentálně“, a to tak, že využijeme výstupů k tomuto účelu speciálně vytvořeného programu *Kombinatorika vyšší dimenze*, o němž byla řeč v úvodu tohoto článku. Tímto programem lze totiž vytvářet kombinace a kombinace s opakováním k-té třídy z n prvků pro n rovné řádově stovkám. Dostáváme tak *k*-členné výběry (experimenty) z *n*-členné populace (základního souboru). Výsledkem je obrovský počet možných výběrů ze základního souboru při vyšších hodnotách *k* a *n* (např. počet kombinací (tedy výběru bez vracení, což se v praxi většinou děje), resp. kombinací s opakováním (tedy výběru s vracením) je pro $k = 10$ a $n = 100$ je roven číslu 17 310 309 456 440, resp. číslu 42 634 215 112 710.

Velice krátce o programu *Kombinatorika vyšší dimenze*:

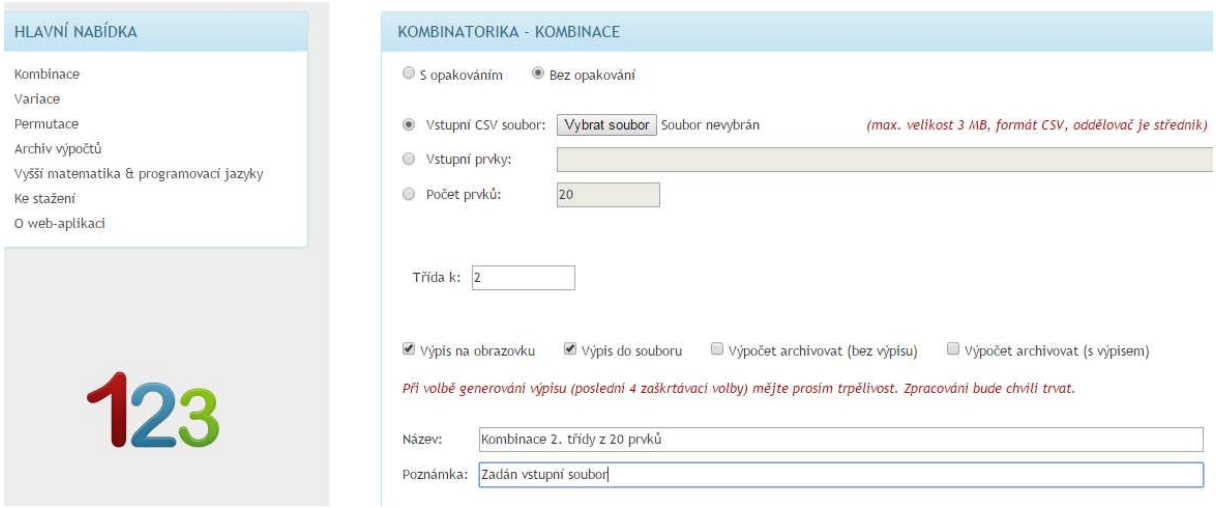
Pro výpočet hodnot byla použita standardní knihovna GMP (GNU Multiple Precision Arithmetic Library). Tato knihovna zajišťuje pouze úkony aritmetické, bitové a porovnávací. Z tohoto důvodu je výpis omezen. Výpis lze provést:

1) cyklem

2) rekurzivním voláním funkce.

V prvním případě je výpis omezen hodnotou počtu opakování dané sekvence. V případě druhém jde o omezení z hlediska hardwarových prostředků (převážně pak operační paměti, případně velikost disku - při zápisu do souboru). Z tohoto důvodu lze provést generování výpisu vždy jen do určité hodnoty.

Vygenerujme si tedy tímto programem potřebné výběrové soubory (nebo použijme soubory už vytvořené a v programu archivované - ta možnost je zde proto, že pro větší hodnoty *n* a *k* blízké *n*/2 trvá výpis všech výběrových souborů hodně dlouho) a vložíme je např. do MS Excel.



Obr. 1 – Ukázka prostředí programu Kombinatorika vyšší dimenze (printscreen)

	A	B	C
1	0	0	
2	0	0	
3	0	0	
4	0	0	
5	0	0	
6	0	0	
7	0	0	
8	0	0	
9	0	0	
10	0	0	
11	0	0	
12	0	0	
13	0	0	
14	0	0	

	A	B	C
178	0	1	
179	0	1	
180	0	1	
181	1	1	
182	1	1	
183	1	1	
184	1	1	
185	1	1	
186	1	1	
187	1	1	
188	1	1	
189	1	1	
190	1	1	
191			

Obr. 2 – Části listu MS Excel s výpisem některých kombinací 2. třídy

V Excelu zjistíme příslušným příkazem

1. počet dvojic tvořených alespoň jednou jedničkou pro případ 15 nul a 5 jedniček (odpovídá to případu 15 dobrých a 5 vadných výrobků)
2. počet dvojic tvořených jen nulami pro případ 10 nul a 10 jedniček (odpovídá to případu 10 dobrých a 10 vadných výrobků).

Ad 1. Takových dvojic nám Excel napočítal 85. Celkem je kombinací druhé třídy z 20 prvků 190. Podíl 85/190 ≐ 0,447 je tedy shora rozebíraná pravděpodobnost chyby prvního druhu.

Ad 2. Takových dvojic nám Excel napočítal 45. Proto dostáváme 45/190 ≐ 0,237, což je pravděpodobnost chyby druhého druhu.

Poznamenejme, že není účelem tohoto článku hledat jiné vhodnější kritérium pro zjišťování, odkud výrobky máme, pro něž by byla chyba prvního druhu „rozumně“ malá (v praxi to bývá 5 %).

This publication was supported by project IGA_PdF_2015_005 Student Grant Competition of Palacký University in Olomouc.

Literatura

1. CUMMING, G. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis.* New York: Routledge, 2012. ISBN 978-0-415-87967-5.
2. RUPPERT, D., MATTESON, D. S. *Statistics and Data Analysis for Financial Engineering.* Berlin: Springer Texts in Statistics, 2015. ISBN 978-1-4939-2614-5.

Kontaktní adresa:

David Nocar, Mgr. Ph.D.,
Katedra matematiky, Pedagogická fakulta UP, Žižkovo nám. 5, 771 40 Olomouc, ČR, tel.: 00420 585 635 5709, fax +420 585 231 400, e-mail: david.nocar@upol.cz

Tomáš Zdráhal, doc. RNDr. CSc.,
Katedra matematiky, Pedagogická fakulta UP, Žižkovo nám. 5, 771 40 Olomouc, ČR, tel.: 00420 585 635 5710, fax +420 585 231 400, e-mail: tomas.zdrahal@upol.cz