NEPARAMETRICKÉ STATISTICKÉ TESTY A JEJICH SOFTWAROVÁ PODPORA

NON-PARAMETRIC STATISTICAL TESTS AND THEIR SOFTWARE REALISATION

L. Ličman, K. Langová, J. Zapletalová

Ústav lékařské biofyziky, Lékařská fakulta, Univerzita Palackého v Olomouci, ČR

Souhrn

Cílem tohoto článku je představit pokročilé možnosti zpracování dat pomocí statistického SW SPSS s využitím programátorského přístupu k této aplikaci. Článek má statistickou a informatickou část. Ve statistické části jsou popsány neparametrické analýzy dat a post hoc testy, které jsou i s pokročilým statistickým SW SPSS uživatelsky obtížněji realizovatelné. V informatické části je popsána možnost realizace post hoc testů pomocí .NET komponenty s COM rozhraním, což pak umožňuje mnohem širší využití komponenty nejen pro rozšíření statistických testů. Dalším využitím může být například tvorba celého statistického postupu složeného z více statistických testů a výstupů podle individuálních potřeb uživatele.

Klíčová slova

statistika, informatika, neparametrické testy, post hoc testy, kontingenční tabulky, Kruskal-Wallisův test, Mann-Whitneyho test, SPSS, .NET

Abstract

The purpose of this article is to introduce advanced data processing with SPSS statistical software and use of a programming approach for this application. The article has a statistical and informatics part. The statistical section describes the non-parametric data analysis and post hoc tests, which are difficult to realize for common SPSS user. The informatics section describes the realisation of this post hoc test with the help of the .NET component with COM interface. This approach allows much wider use of the component, not only for the extension of the statistical tests. Another usage may be, for example, formation of the statistical procedure which is composed of several statistical tests and outputs according to the individual needs of the user.

Keywords

statistics, Informatics, nonparametric tests, post hoc tests, cross tables, Kruskal-Wallis test, Mann-Whitney test, SPSS, .NET

Úvod

V biomedicíně má mnoho úloh klinických studií komparativní charakter. Posuzujeme např. účinky různých způsobů léčby nebo vliv různých rizikových i protektivních faktorů na úspěšnost léčby. Toto vede k úlohám porovnat několik různých vybraných skupin pacientů. Při statistickém usuzování nás zajímá, zda jsou rozdíly mezi jednotlivými výběry statisticky významné.

Ve 30. letech 20. století vytvořil R. A. Fisher metodu analýzy rozptylu (ANOVA), která se používá k porovnání tří a více populačních průměrů. Použití této parametrické metody je vázáno na splnění několika předpokladů, především předpokladu normálního rozdělení dat. Pro medicínská data není tento předpoklad často splněn (např. kvůli existenci extrémně vysokých nebo nízkých hodnot). V tomto případě je nutné použít pro porovnání výběrových souborů neparametrické metody. Neparametrickým ekvivalentem analýzy rozptylu

je Kruskal-Wallisův test. Tento test ověřuje nulovou hypotézu, že porovnávané výběry pochází ze stejné populace, tj. distribuční funkce všech výběrů jsou shodné, proti alternativě, že alespoň jedna distribuční funkce se liší. Pokud nulovou hypotézu zamítneme, je třeba provést další analýzu, abychom zjistili, které výběry se od sebe liší. Jednou z možností je porovnat každou dvojici výběrů, nebo porovnat všechny výběry s kontrolním výběrem. Porovnání po dvojicích je možné provést pomocí Mann-Whitneyho U-testu. Problémem je skutečnost, že mnohonásobné testování zvyšuje pravděpodobnost výskytu chyby prvního druhu, to je, že zamítáme nulovou hypotézu, která ve skutečnosti platí. Chyba prvního druhu by při testování neměla překročit hranici 5 %.

Pro řešení problému mnohonásobného porovnávání existuje několik metod, jako například Bonferroniho. Bonferroniho metoda spočívá v úpravě hladiny signifikance při testování. Buď hladinu významnosti, na které provádíme testování (většinou α=0,05), podělíme počtem prováděných testů nebo ekvivalentně dosažené hladiny významnosti získané při mnohonásobném porovnávání násobíme počtem provedených testů.

Kruskal-Wallisův test je možné použít pro data kvantitativní nebo ordinální. Při porovnání výběrů v nominálních znacích (např. pohlaví pacientů) jsou data uspořádána do kontingenčních tabulek a nulová hypotéza je ověřena obvykle chí-kvadrát testem nebo Fisherovým přesným testem. Pokud máme porovnat více než dva výběry, dostáváme se do podobné situace jako u Kruskal-Wallisova testu. Po zamítnutí nulové hypotézy přistupujeme k mnohonásobnému porovnání s opětovným využitím chí-kvadrát testu nebo Fisherova přesného testu. Hladinu signifikance opět můžeme korigovat pomocí Bonferroniho metody.

Na našem pracovišti používáme pro statistické analýzy biomedicínských dat software SPSS, který, bohužel, nepodporuje testy mnohonásobného porovnání u neparametrických metod.

V programu SPSS, který je jedním ze standardů na poli zpracování statistických dat, můžeme kombinovat ovládání jak pomocí menu, tak skriptovacího jazyka. Podporu chybějící post hoc analýzy jsme se rozhodli do SPSS přidat, Bonferroniho metodu korekce signifikance jsme použili z důvodu snadné implementace.

Materiál a metody

Automatizace v SPSS

SPSS nabízí dva způsoby automatizace. První způsob je zápis statistického testu pomocí své syntaxe. Tento zápis nahrazuje výběr a vyplnění formulářů z menu. Značně tak ulehčuje práci, protože se nemusí pokaždé vyplňovat potřebné parametry testu pomocí menu. Syntaxe je prostý text, který se případně dá vygenerovat programem a spustit. Zápis syntaxe se dá kombinovat i s programovacím jazykem Python. Druhý způsob automatizace je zcela programátorský. Pomocí krátkého programu (skriptu) buď v programovacím jazyku Visual Basic nebo Python se dají přímo ovládat výstupní objekty (tvořit tak například výstupní tabulku) nebo přímo spouštět testy, zapsané pomocí syntaxe z prvního způsobu.

Ani jeden ze způsobů automatizace sám nestačí na automatizaci neparametrické post hoc analýzy. Například pro neparametrický Kruskal-Wallisův test s post hoc analýzou potřebujeme nejdříve zjistit výsledek Kruskal-Wallisova testu a pokud je zamítnuta nulová hypotéza, musíme provést porovnání všech dvojic pomocí Mann-Whitney U-testu a výsledky všech porovnání přehledně uspořádat do jediné výsledné tabulky. Toto se musí navíc udělat v prostředí, kde se dají statistické analýzy a výsledné výstupy rozdělit (pomocí funkce Split) podle libovolně zvolené proměnné. Neprovádíme tak na datovém souboru jedinou post hoc analýzu dat, ale několik, v závislosti na zadaném rozdělení podle další proměnné. Výsledek tedy není pouze jeden, počet výsledků závisí na vstupním rozdělení dat. I tyto výsledky je potřeba ještě dále uspořádat do přehledných tabulek.

Nejdůležitější požadavky na rozšíření SPSS jsou:

1. Komponenta musí být ovladatelná a musí se spouštět přímo z SPSS, protože musí reagovat na nastavené filtry v datovém souboru, jeho rozdělení atp.
2. Všechny parametry statistického testu musí být vybrány z formuláře, který bude specifický pro tento test. Musí se tak vytvořit nový formulář, nestačí nám formulář už obsažený v SPSS.
3. Komponenta musí postupně spustit několik statistických testů.
4. Komponenta musí parsovat výsledky statistických testů a na základě těchto výsledků spouštět další statistické testy.
5. Komponenta musí výsledky uspořádat do nových tabulek a umožnit SPSS tuto novou tabulku včlenit do výstupu.

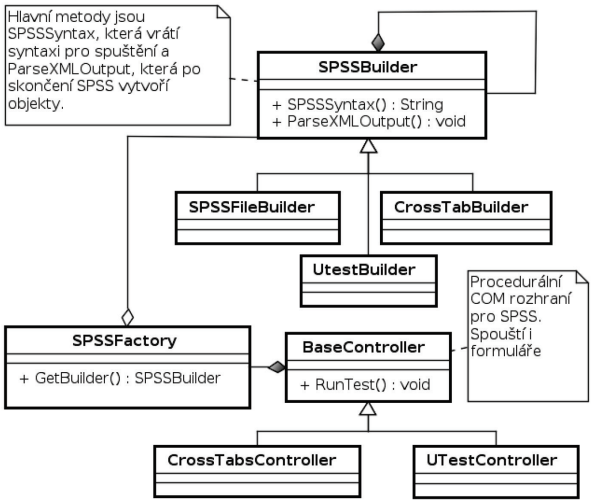
Výsledek spuštěného statistického testu (tabulka nebo tabulky) se nemůže zobrazovat jako výstup v SPSS, ale jsou to pouze mezivýsledky pro další zpracování. SPSS umožňuje takový výsledek spuštěného testu (zapsaného pomocí SPSS syntaxe) uložit do textového souboru ve značkovacím jazyce XML a nezobrazovat ho pouze jako tabulkový výstup. Soubor XLM se pak dá parsovat pomocí různých knihoven v různých jazycích. Po rozparsování se ve zpracovávaném XML souboru dají najít potřebné výsledky například pomocí XPath nebo LINQ.

Vlastní rozšíření pro SPSS se tak skládá ze dvou částí. První část je skript ve Visual Basicu, který se spouští přímo z prostředí SPSS a přes COM rozhraní komunikuje s .NET komponentou. Tento skript spouští pomocí syntaxe SPPS potřebné testy. Na závěr načte z .NET komponenty zformátovanou tabulku a zobrazí ji v SPSS výstupu stejně, jako se zobrazují nativní SPSS výsledky.

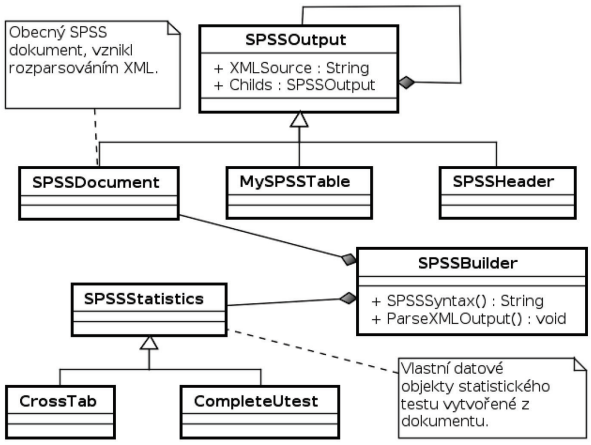
Druhá část rozšíření pro SPSS je .NET komponenta, která je popsána v následujícím odstavci.

Návrh komponenty

Komponentu je potřeba navrhnout co nejobecněji. Prvním základním předpokladem je zjistit informace o proměnných. Pro konstrukci porovnávaných dvojic je třeba znát všechny jedinečné hodnoty dané proměnné a jejich popisky. Pro tento účel používáme knihovnu přímo od SPSS (IBM SPSS Statistics Input/Output Module). Dalším úkolem je zajistit komunikaci SPSS s komponentou. K tomuto účelu stačí mít komponentu přístupnou pomocí COM rozhraní, tím se dá využívat a ovládat přímo ze skriptu SPSS.



Obr. 1: Class diagram komponenty. Základní rozdělení.



Obr. 2: Class diagram objektu SPSSBuilder podrobně.

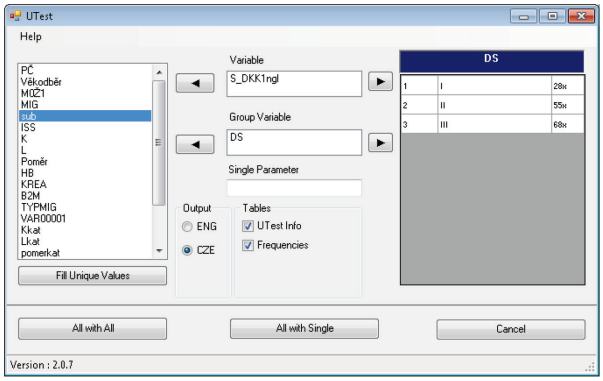
Další částí je jádro komponenty. Je vytvořeno pomocí objektové metodiky, s častým použitím návrhových vzorů a se snahou o co největší obecnost a znovupoužitelnost. Základem je návrhový vzor Builder, který postupně buduje výsledný složený statistický test. Základní chování tohoto Builderu je jednoduché. SPSS Builder vydá potřebnou textovou syntaxi pro vytvoření jednoho XML souboru. SPSS tuto syntaxi spustí a Builderu oznámí, že analýza skončila. Na tyto činnosti stačí dvě metody každého Builderu - SPSSSyntax a ParseXMLOutput. Na základě zpracovaného mezivýsledku Builder může připravit další syntaxi, takže tento proces probíhá v cyklu, dokud není vrácená syntaxe prázdná. Celá problematika je samozřejmě složitější, protože Builder buduje složené rekurzivní objekty (vzor Composite), případně se Builder může sám skládat z více Builderů (opět návrhový vzor Composite, v tomto případě použitý i na Builder). Class diagram návrhu komponenty je vidět na obrázku 1 a obrázku 2. Po vytvoření kompletního statistického testu už jenom komponenta vrátí přes COM rozhraní textový zápis vytvořených tabulek, které je třeba zobrazit. Jednoduchý klient komponenty v SPSS se na závěr pouze zobrazí do statistického výstupu.

Uživatelské rozhraní komponenty

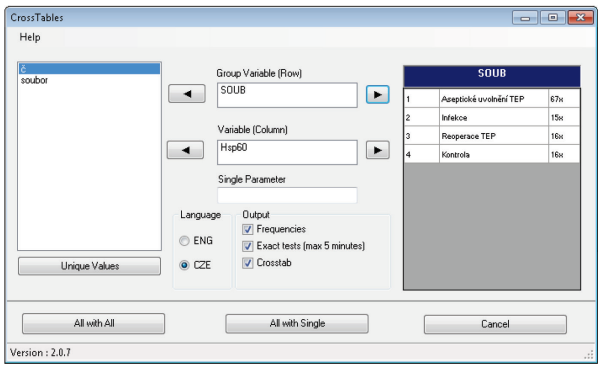
Komponenta obsahuje dva formuláře. Jeden pro post hoc analýzu Kruskal-Wallisova testu a druhý pro stejné post hoc analýzy v kontingenčních tabulkách. Navíc oproti formulářům v SPSS naše formuláře umožňují i jednoduchou frekvenční analýzu přímo ve formuláři. Už při výběru proměnných je možné se podívat, jaké kategorie proměnná obsahuje. Ukázka formulářů je na obrázku 2 a obrázku 3. Kromě klasické post hoc analýzy dvou skupin, kdy se každá skupina testuje s každou, komponenta umožňuje navíc post hoc analýzu, kde lze vybrat jednu skupinu jako kontrolní a pouze tuto srovnat se zbývajícími. Místo všech kombinací druhé třídy z n prvků bez opakování

$$C_2(n) = \frac{n!}{2!(n-2)!} \qquad (1)$$

tak máme při výběru kontrolní skupiny pouze (n - 1) dvojic.



Obr. 3: Formulář pro post hoc analýzu po Kruskal-Wallisově testu.



Obr. 4: Formulář pro post hoc analýzu kontingenčních tabulek.

V analýze medicínských dat například může být kontrolní skupina tvořena výsledky pacientů léčených konzervativně a srovnávat se může se skupinami léčenými jinými postupy.

Výsledky

Funkčnost navržené komponenty byla testována na datech klinických studií. Post hoc analýza u Kruskal-Wallisova testu byla například použita u klinické studie zabývající se studiem pacientů s mnohočetnými myelomy

Byla studována závislost stavu nemoci (třístupňová ordinální proměnná DS udávající stádium nemoci, I znamená nejlehčí stádium) a hladiny rekombinantního proteinu (DKK1 - Dickkopf-related protein 1). Jeho hladina je zvýšená právě při mnohačetném myelomu. Byla testována nulová hypotéza, která předpokládala, že hodnoty DKK nejsou závislé na stavu nemoci. Nejdříve byl proveden Kruskal-Wallisův test (tab. 1a). Dosažená hladina významnosti byla nižší než 0,05. Nulovou hypotézu jsme zamítli a následně bylo provedeno mnohonásobné porovnání s Bonferroniho korekcí. Jak je z výsledků v tabulce 1b patrné, tak statisticky významný rozdíl hladin tohoto proteinu je pouze mezi stádiem DS I a DS III (p=0,0004). U pacientů ve třetím stádiu nemoci byla zjištěna statisticky významně vyšší hladina DKK1 než u pacientů v 1. stádiu nemoci.

	Kruskal-Wallis Statistics
	Value
Kruskal-Wallis Chi-Square	15,061
Kruskal-Wallis Asymp. Sig.	**0,0005**
Kruskal-Wallis df	2

Tab. 1a: Výsledky Kruskal-Wallisova testu u studie mnohočetného myelomu.

DS	UTest Statistics				
	U	W	Z	Sign 2 tail	Bonf. corr.
I:II	591	997	-1,725	0,085	0,254
I:III	483	889	-3,795	**0,0001**	**0,0004**
II:III	1423,500	2963,500	-2,285	**0,022**	0,067

Tab. 1b: Post hoc testy u studie mnohočetného myelomu.

Post hoc analýza u kontingenčních tabulek byla použita u klinické studie zabývající se možnostmi detekce zánětu v periprotetických tkáních u pacientů s implantací totální endoprotézy (TEP) kyčle nebo kolene. V rámci studie byla mj. sledována pozitivita markeru Hsp-60. Hsp-60 je protein, který je jedním z markerů poškození tkání (nebo jejich stresu), v tomto případě zánětem, který souvisí s přítomností endoprotézy. Cílem studie bylo porovnat výskyt pozitivity markeru Hsp-60 u čtyř skupin pacientů. Tabulka 2a uvádí zjištěný výskyt pozitivity ve všech sledovaných skupinách pacientů a v tabulce 2b jsou shrnuty výsledky testu homogenity chí-kvadrát a výsledky Fisherova přesného testu na kontingenční tabulce. Výsledky Fisherova přesného testu jsou spolehlivější v případě, kdy očekávané četnosti za platnosti nulové hypotézy (tj. distribuce jsou stejné ve všech porovnávaných skupinách) jsou menší než 5.

SOUBOR		Hsp60		
		Neg.	Poz.	Total
Aseptické uvolnění TEP (1)	N	46	18	64
	%	71,9%	28,1%	100,0%
Infekce (2)	N	4	11	15
	%	26,7%	73,3%	100,0%
Reoperace TEP (3)	N	16	0	16
	%	100,0%	0%	100,0%
Kontrola (4)	N	9	7	16
	%	56,3%	43,8%	100,0%
Total	N	75	36	111
	%	67,6%	32,4%	100,0%

Tab. 2a: Kontingenční tabulka u studie pacientů s implantací totální endoprotézy (TEP) kyčle nebo kolene.

	Chi Square Test (Complete Table)		
	Chi Square-Asymp. Sig. (2-sided)	Chi Square-df	Fisher's Exact Test-Exact. Sig. (2-sided)
Hsp60:SOUBOR	**0,0001**	3	**<0,0001**

Tab. 2b: Výsledky analýzy kontingenční tabulky. Chí kvadrát test a Fisherův přesný test.

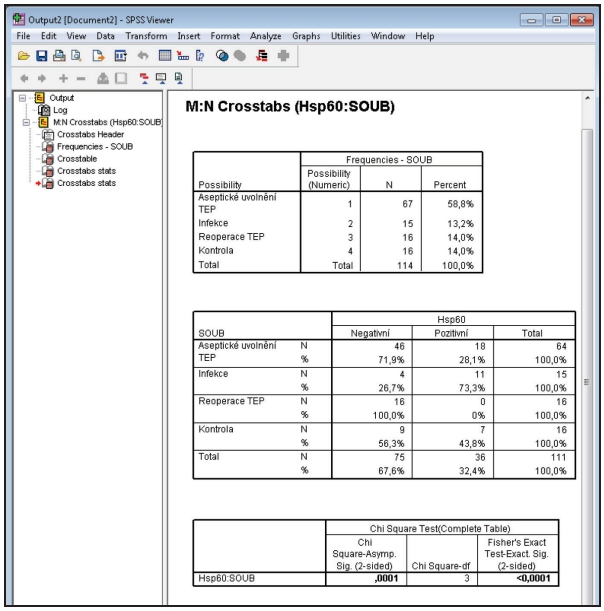
Z výsledků v tabulce 2b je zřejmé, že sledované skupiny pacientů se signifikantně liší ve výskytu pozitivity markeru Hsp-60. Hladina signifikance chí-kvadrát testu homogenity je p = 0,0001 a hladina signifikance Fisherova přesného testu p < 0,0001. Kontingenční tabulku je potřeba dále analyzovat - porovnat skupiny pacientů po dvojicích a zjistit které skupiny se liší statisticky významně. Tabulka 2c uvádí výsledky porovnání skupin po dvojicích, tj. post hoc analýzy.

SOUB	Chi Square-Asymp. Sig. (2-sided)	Bonf. corr.	Fisher's Exact Test-Exact. Sig. (2-sided)	Bonf. corr.
1:2	**0,001**	**0,006**	**0,002**	**0,013**
1:3	**0,016**	0,096	**0,017**	0,100
1:4	0,228	1	0,242	1
2:3	**<0,0001**	**0,0001**	**<0,0001**	**<0,0001**
2:4	0,095	0,572	0,149	0,893
3:4	**0,003**	**0,017**	**0,007**	**0,041**

Tab. 2c: Výsledky post hoc analýzy u kontingenční tabulky. Popis jednotlivých skupin je u tab. 2a.

Signifikantně vyšší výskyt pozitivity Hsp-60 byl prokázán v souboru 2 Infekce (73%) ve srovnání se souborem 1 Aseptické uvolnění (28%, p = 0,013), resp. ve srovnání se souborem 3 Reoperace TEP (0%, p<0,0001). V souboru 3 Reoperace TEP byl prokázán signifikantně nižší výskyt pozitivity Hsp-60 než v souboru 4 Kontrola (0% vs. 44%, p = 0,041).

Na obrázku 5 je zobrazena ukázka statistických výstupů naší komponenty v prostředí SPSS. Jak je z obrázku patrné, výstup je podobný nativním výstupům z SPSS. Navíc zvýrazní signifikantní výsledek a umožňuje zvolit popisy tabulek podle potřeby buď v angličtině, nebo v češtině.



Obr. 5: Ukázka výstupu komponenty v prostředí SPSS.

Diskuze

Požadavky na post hoc testy mnohonásobného porovnání jsou natolik obecné, že se dají využít v podstatě na konstrukci jakékoli posloupnosti za sebou prováděných statistických testů. Bez rekurzivní zpětné vazby (spuštění statistického testu na základě předchozího výsledku) je každý pokus o automatizaci redukován na jednoduchý formát dotaz - odpověď podobně jako například v SQL. Proto nestačí jednoduché možnosti automatizace obsažené v SPSS a vznikla tak potřeba doprogramovat vlastní řešení.

Pro komponentu v .NETu místo programování v jazyku Python, který SPSS podporuje nativně, jsme se rozhodli z několika důvodů. Prvním důvodem je komplexnost .NET frameworku, který je mnohem častěji používán v korporátním prostředí oproti Pythonu. Složitější a rozsáhlejší aplikace jsou v tomto prostředí snadněji udržovatelné díky modifikátorům přístupů k metodám tříd a díky statickým typům oproti dynamickým typům v Pythonu. Kód a knihovny Pythonu se dají začlenit do prostředí .NET pomocí IronPythonu, opačný postup není možný. Z prostředí .NETu je tak dostupných víc knihoven a dokonce je z něj přístupný i kód Pythonu. Výhodou prostředí .NET je také integrovaný dotazovací jazyk LINQ, který velmi zjednodušuje práci s objekty a s XML textovým výstupem SPSS.

Další výhodou je samo vývojové prostředí Visual Studia, kde se rozsáhlejší kód velice dobře spravuje a programuje.

Závěr

Mnoho studií komparativního charakteru se neomezuje pouze na dvě skupiny. Při porovnání více skupin (například Kruskal-Wallisův test, chí kvadrát test atd.) nám výsledek poskytne pouze informaci o tom, zda jsou skupiny stejné nebo různé. Ke zjištění které skupiny jsou stejné a které se od sebe liší, můžeme dále použít post hoc analýzu a srovnávat jednotlivé skupiny mezi sebou po dvojicích. Zdánlivě by stačilo provést tyto testy na stejné hladině α jako původní test. Kdyby byl ovšem každý z těchto testů mnohonásobného porovnání proveden na původní hladině α, byla by výsledná hladina, na které testujeme, vyšší než původní α. Proto se volí postupy, které udrží hladinu α na obvyklé úrovni 0,05. Při post hoc testech je tedy nutné korigovat hladinu α například pomocí Bonferroniho korekce.

Podařilo se vytvořit velmi potřebnou komponentu pro automatické zpracování neparametrických post hoc testů v SPSS. Komponenta je nyní využívána a dále se může rozšiřovat (přidávat do uživatelských výstupů další možnosti), je natolik obecná, že popsaná knihovna a postupy se dají využít i na konstrukci jakéhokoli statistického postupu, který se sestává z více

kroků a samozřejmě na konstrukci plně uživatelských tabulkových výstupů v SPSS.

Poděkování

Autoři děkují za poskytnutá data lékařům Fakultní nemocnice v Olomouci. Tato práce byla zpracována za podpory projektu CZ.1.07/2.4.00/17.0058 Prohloubení odborné spolupráce a propojení ústavů lékařské biofyziky na lékařských fakultách a projektu CZ.1.05/2.1.00/01.0030.

Literatura

[1] Peter Armitage, Geoffrey Berry *Statistical Methods in Medical Research*. 2002, John Wiley & Sons.

[2] Hendl, J. *Přehled statistických metod zpracování dat.* 2004, Praha: Portál.

[3] Jana Zvárová, *Základy statistiky pro biomedicínské obory.* 2006 EuroMISE centrum.

[4] SPSS Inc. Released 2006. SPSS for Windows, Version 15.0. Chicago, SPSS Inc.

[5] Raynald Levesque and SPSS Inc., *SPSS Programming and Data Management*. 2007, SPSS Inc.

[6] *SPSS 14.0 Developer's Guide*. 2005 by SPSS Inc.

[7] *IBM SPSS Statistics Input/Output Module*. 2011, IBM.

[8] *SPSS 15.0 Command Syntax Reference.*

[9] Ilja Kraval *Design Patterns v OOP*. 2002.

[10] Gamma, Erich; Richard Helm, Ralph Johnson, and John Vlissides *Design Patterns: Elements of Reusable Object-Oriented Software.* 1995, Addison-Wesley. ISBN 0-201-63361-2.

[11] Charlie Calvert, Dinesh Kulkarni *Essential LINQ*, 2009, Pearson Education, Inc.

Mgr. Libor Ličman
Ústav lékařské biofyziky
Lékařská fakulta
Univerzita Palackého v Olomouci
Hněvotínská 3, CZ-775 15 Olomouc

E-mail: libor.licman@gmail.com
tel.: +420 723 513 539