

Statistical Data Processing in STAT1 Application

Abstrakt:

Analýzy datových souborů užitím metod exploratorní, zejména induktivní statistiky je v dnešní době nedílnou součástí řady oblastí lidské činnosti. Článek se věnuje problematice statistického zpracování dat pomocí nástroje STAT1, který pracuje pod Microsoft Office Excel. Je podáno vysvětlení, uveden zdroj a provedeno srovnání zmíněné aplikace s alternativními softwarovými produkty. Součástí příspěvku je ukázka praktického použití STAT1 při řešení reálných úloh.

Abstract:

Data analyses using methods of exploratory and inductive statistics nowadays form an integral part of many areas of human activities. The paper is focused on the statistical processing of data using a new application STAT1 that works under Microsoft Office Excel. The explanation is given, the source is stated, and the comparison with alternative application software tools is mentioned. Moreover, the examples of practical utilization of STAT1 in the military area are presented.

Klíčová slova:

Analýza datových souborů, exploratorní metody, induktivní statistika, aplikace STAT1, srovnání alternativních softwarových produktů.

Key words:

Data sets analysis, exploratory methods, inductive statistics, STAT1 application, alternative software tools comparison.

Úvod

Statistika tvoří nedílnou součást nejenom ekonomického, ale i technického a humanitně zaměřeného vzdělání. Se statistikou se setkává každý, kdo potřebuje zpracovávat a vyhodnocovat datové soubory, ať v souvislosti s řešením rezortních úkolů či mimo něj.

Většina laiků si statistiku spojuje s matematikou, považuje ji za jakousi její odrůdu. Ve srovnání s matematikou má však statistika zcela odlišnou filozofii – užívá zcela odlišný způsob myšlení. Statistika se nezabývá výpočty veličin, ale jejich odhady. Je tvořena třemi základními pilíři: pravděpodobností, náhodnou veličinou a popisnou

statistikou. Pochopit jakou roli ve statistice hrají a jaké jsou souvislosti mezi těmito součástmi navzájem a směrem k odhadům, je pro začátečníka často ne příliš snadné. Na druhé straně pozitivně působí skutečnost, že statistika řeší zejména praktické problémy ze života a odpovídá na reálné otázky.

Cílem tohoto příspěvku je právě v této praktické rovině seznámit potenciálního uživatele s nově vzniklým statistickým nástrojem „Aplikace STAT1“, provést jeho základní srovnání s alternativními softwarovými produkty, a rovněž nastínit možnosti a způsob jeho praktického využití.

Statistický software

Výpočetní technika vstoupila do statistiky již před mnoha lety, pro většinu moderních statistických metod je počítačová podpora nezbytností. Na trhu je k dispozici řada softwarových produktů, jako např. Statistica, SPSS, Statgraphics, QCExpert, Minitab, S plus či Matlab (statistics toolbox), pomocí nichž je možné statistické analýzy provádět. Uvedené produkty se liší rozsahem nabízených metod a analýz, grafickým rozhraním, uživatelskou přístupností, univerzálností apod. Co však mají všechny tyto produkty společné, to je skutečnost, že jsou komerční, je tedy nutné si zaplatit licenci. Jednou z mála výjimek je jazyk a statistické prostředí R. Jedná se tzv. GNU (z anglického „GNU's Not Unix!“) projekt, je tedy k dispozici zdarma. Položíme-li si otázku, který z uvedeného software tedy používat, jsme většinou limitováni dostupností daného software na pracovišti, tedy potažmo cenou.

Autoři příspěvku už vyzkoušeli odborné přednosti programů Statistica, QCExpert, Matlabu a R. První dva jmenované softwarové produkty nejsou nikterak náročné na ovládnání a orientaci v nabízené struktuře metod resp. jednotlivých technik. Mají intuitivní grafické rozhraní a oba jsou k dispozici i v českém jazyce.

Matlab je známý a rozšířený produkt s širokým záběrem v různých odvětvích exaktních disciplín. Vyžaduje však speciální toolboxy a neobejde se bez znalosti speciálního jazyka. Vhodný je spíše pro pokročilejší statistické metody náročné na výpočetní stránku. Pro řadu uživatelů může být překážkou, že je pouze anglickém jazyce, i když literaturu popisující práci s tímto programem je samozřejmě možné najít i v češtině. Program R se v poslední době stává velice oblíbeným nástrojem pro statistické zpracování dat a modelování. Pro začátečníka, či jedince s minimálním povědomím o statistice jako takové, se může tento software jevit jako příliš náročný na ovládnání (je třeba se naučit základy jazyka R). Navíc je prostředí R k dispozici, stejně jako Matlab, pouze v anglické mutaci.

Na základě těchto zkušeností vytvořili autoři excelovskou aplikaci STAT1, pomocí níž mohou provádět základní statistické výpočty a zpracování dat i začátečníci či lidé s minimálním statistickým vzděláním. Z uživatelského hlediska se tedy jedná o nenáročný, dostupný a aplikačně široce využitelný nástroj.

Statistika v Excelu

Základní zpracování dat je možné provádět pomocí známého produktu Excel, který je součástí kancelářského balíku MS Office. Značnou výhodou tohoto balíku je právě jeho masová rozšířenost. Nejedná se pochopitelně o speciální statistický software,

ale tabulkový charakter Excelu umožňuje využít řady jeho zabudovaných prostředků a zajímavých vlastností.

Především se jedná o široké možnosti užití vlastních numerických výpočtů pomocí rovnic vytvořených uživatelem. Excel zvládá i maticové výpočty. Různé numerické výstupy lze velmi jednoduchým způsobem uspořádat do tabulek, kterých statistika hojně využívá. Výhodou je zejména to, že uživatel pracuje interaktivně a uspořádá výstupy podle svých potřeb. Zanedbatelné nejsou ani možnosti grafické, Excel nabízí pestrou škálu typů grafů, každý ještě v několika modifikacích. Uživatel se podle svých zkušeností může rozhodnout, který z grafů bude pro zobrazení statistické vlastnosti nejvhodnější.

V řadě situací je možné využít vlastní excelovské procedury z různých oblastí statistiky, označené jako analytické nástroje. Využít je možné balíčky popisné a pořadové statistiky, rozdělení četností, dvouvýběrových testů, analýzy rozptylu, regrese a korelace a dalších. Jedná se o pevné procedury, které dávají stále stejné typy výsledků, a je možné je proto snadno komentovat. Nejširší nabídku poskytuje Excel v kategorii statistických funkcí. Nespornou výhodou Excelu je jeho schopnost provést automaticky přepočítání všech definovaných funkcí při změně vstupních hodnot.

Je však třeba upozornit také na jisté nedostatky Excelu. Ti, kteří se již se statistickými funkcemi v Excelu setkali, si jistě povšimli poněkud nejasné a místy i zavádějící terminologie v popisu funkcí i v nápovědě k jednotlivým funkcím. Mimo těchto nejasností, nad kterými by mohl mnohý uživatel mávnout rukou, se zde objevují i jiné nešvary. Jedná se zejména o nejednotnost při určování hodnot inverzních funkcí hustot některých pravděpodobnostních rozdělení (především ve starších verzích Excelu) apod. Tato nejednotnost v zadávání parametrů excelovských funkcí je matoucí a může způsobit uživateli problémy, někdy i chyby ve výpočtech. Podobných nepříjemností by bylo možno uvést více.

Aplikace STAT1

Vedle využití zabudovaných excelovských prostředků – zejména grafů, analytických nástrojů a funkcí – lze v Excelu vytvořit prostředí podle vlastních požadavků. Autoři příspěvku vytvořili v tomto prostředí aplikaci s názvem STAT1, viz [1]. Je určena především pro základní zpracování dat v podobě popisné statistiky a dále jsou zde implementované metody jednorozměrné induktivní statistiky, dvouvýběrové testy a chí-kvadrát test nezávislosti v kontingenční tabulce. Aplikace je konstruovaná tak, aby s naprosto minimálními vstupy poskytovala snadno řadu užitečných výstupů, které již stačí „jen“ správně interpretovat. Z tohoto pohledu se chová jako profesionální programy založené na výběru procedur z menu a stanovení parametrů prováděné analýzy.

Právě komentáře a interpretace získaných výsledků považují autoři za naprosto klíčové, protože umožňují uživateli postupně budovat správnou představu o statistické filozofii. Prostor k tomu je vytvořen zejména tím, jak jednoduché je ovládání této aplikace. Uživatel se tedy soustředí na problémy skutečně statistické a nemusí svojí pozornost věnovat samotnému ovládání programu.

První list aplikace STAT1 s označením data slouží k vložení datových souborů určených k statistické analýze (viz obr. 1 na str. 118). Při prvním setkání s tímto programem zjistí uživatel, že list již obsahuje datové soubory z publikace *Základy statistiky*. [2]

Použitelnost aplikace není však omezena jen na příklady z uvedené knihy. Je také samozřejmě možné, aby si uživatel do listu data vložil svoje vlastní hodnoty, které v záhlaví sloupce označí – pojmenuje. Výběr konkrétních dat (proměnných) pro další zpracování se pak provádí přímo ve zvoleném výpočetním listu podle toho, jakou analýzu se uživatel rozhodne provádět. Dále se vloží požadované parametry úlohy (jsou zvýrazněné červeně), např. hladina významnosti, velikost přípustné chyby apod. Statistické výstupy – výsledky jednotlivých analýz – jsou zobrazené v zelených polích.

Další tři listy – popisné charakteristiky, bodové rozdělení a intervalové rozdělení, viz obr. 2 (na str. 118) – umožňují provést exploratorní analýzu dat, pomocí které lze posoudit důležité vlastnosti pozorovaného datového souboru. Z tabulkového a grafického vyjádření rozdělení četností dokážeme orientačně posoudit, z jakého rozdělení výběr pochází, zda je výběr homogenní a zda neobsahuje odlehlé hodnoty. Získáme také nej-používanější výběrové charakteristiky polohy, variability a koncentrace. Ty jsou všechny počítané pomocí excelovských funkcí z původních dat (uvedených v listu data).

Tyto tři listy navíc shodně obsahují procedury představující testy o nulové šikmosti a nulové špičatosti resp. kombinovaný C-test o šikmosti a špičatosti a jejich modifikované verze vhodné pro náhodné výběry menšího rozsahu. [3] Tyto testy se použijí na zvolené hladině významnosti k rozhodnutí, zda zkoumaný výběr pochází z normálního rozdělení. Spolu s tzv. Q-Q plotem představují názorný prostředek k posouzení normality. Pokud testy normálního rozdělení zamítnou, považujeme data za výběr z jiného (libovolného, neznámého) rozdělení. To má význam při výběru listů, na kterých jsou připravené další statistické metody. S implementací jiných testů tvaru rozdělení, které jsou standardně nabízeny statistickými programy, se počítá v dalším rozšíření aplikace STAT1.

Na dalších listech jsou zpracované jednovýběrové a dvouvýběrové metody (označení listů 1V a 2V) odhadů a testů pro střední hodnoty a rozptyly za předpokladu normality, a také pro libovolné rozdělení. Po výběru datových souborů se zobrazí potřebné výběrové charakteristiky (zpravidla rozsah souborů, výběrové průměry a výběrové odchylky) a na zvolené hladině významnosti se určí intervaly spolehlivosti. Po nastavení zvolené alternativy (u 1V-úloh i hodnoty testovaného parametru) se s ohledem na řešení konkrétního problému zobrazí kritická hodnota a p-hodnota pro dané nastavení – test. Veškeré výstupy a výsledky jsou zobrazené v zelených buňkách.

List kontingenční tabulka je určen pro testování nezávislosti v kontingenční tabulce užitím tzv. chí-kvadrát testu, viz např. [3], [4] nebo [5]. Poslední list aplikace tvoří elektronické statistické tabulky.

Ovládání jednotlivých listů je zcela intuitivní, uživatel má však také oporu v knize. [2] Jednoduchý návod je také možné najít na stránkách <http://k101.unob.cz/stat1/>, odkud lze zdarma získat popsanou aplikaci. Rovněž je nutné poznamenat, že autoři již pracují na její anglické mutaci.

Příklady řešené v aplikaci STAT1

Praktické užití aplikace STAT1 si ukažme na dvou konkrétních příkladech.

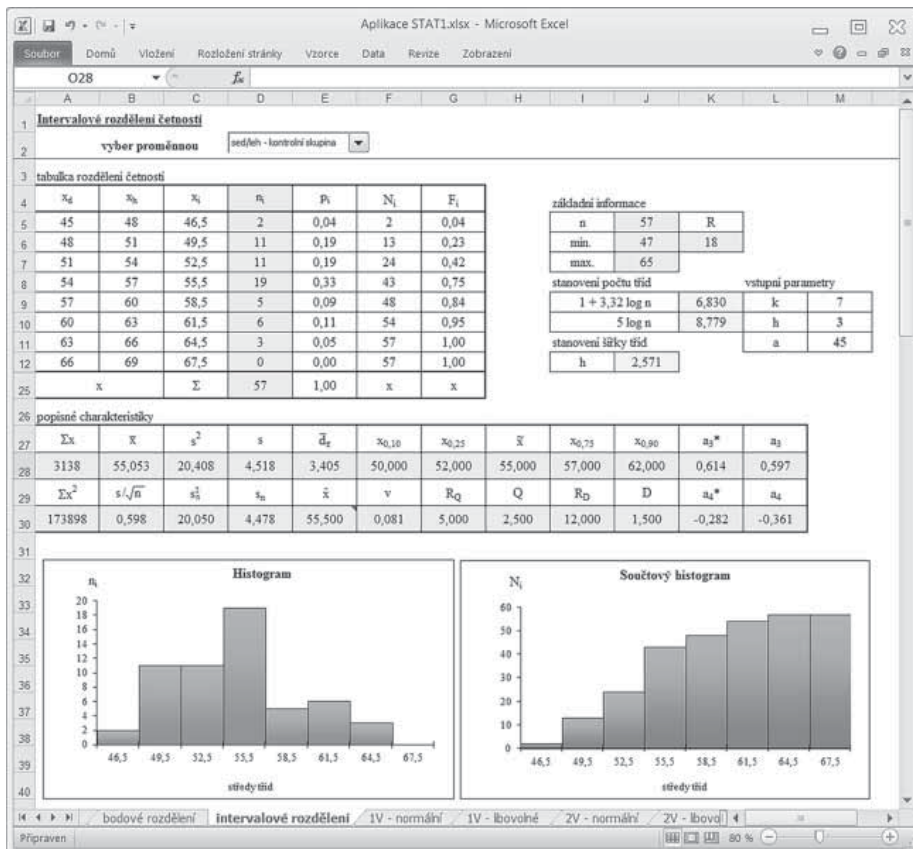
V rámci tělesné přípravy byla u jedné skupiny vojáků (experimentální skupina) zavedena inovovaná alternativní forma tělesné přípravy. Cílem bude zodpovědět otázku, zda tento nový přístup vede ke zlepšení fyzické výkonnosti vojáka, konkrétně se zaměříme

Applikace STAT1.xlsx - Microsoft Excel

sed/leh - experimentální skupina

	A	B	C	D	E	F	G
1	sed/leh - experimentální skupina	sed/leh - kontrolní skupina	moje data 3	tuk v mléku	výška chlapců	prach ve vzduchu	38-1 lačůvky
2	58	50		14,85	83	1,23	49,8
3	56	65		14,68	85	1,51	50,2
4	56	51		15,27	81	1,41	50,3
5	55	59		14,77	82	1,14	49,5
6	53	65		14,83	84	1,47	50,0
7	57	62		14,95	82	1,10	49,3
8	55	65		15,08	79	1,53	50,0
9	48	51		15,02	84	1,22	50,9
10	62	62		15,07	80	1,34	50,4
11	55	50		14,98	81	1,24	50,0
12	54	55		15,15	82	1,54	49,7
13	50	50		15,49	82	1,31	50,6
14	67	55		14,83	80	1,27	50,2
15	55	50		14,95	82	1,16	49,9
16	50	55		14,78	80	1,45	50,1
17	58	50			82	1,34	

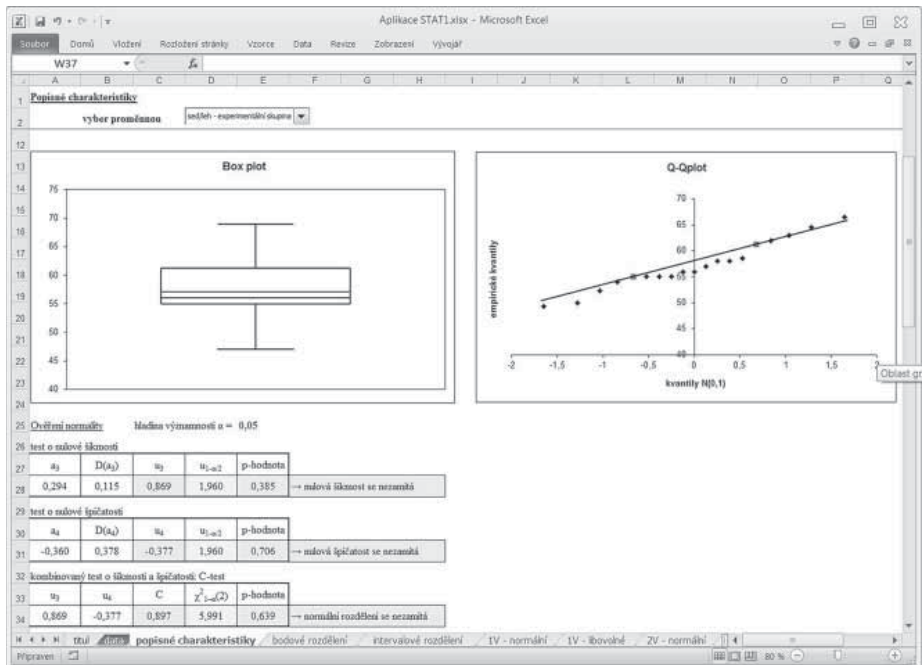
Obr. 1: Datové soubory



Obr. 2: Intervalové rozdělení četnosti a výběrové charakteristiky

na počet sedů-lehů za minutu. Máme k dispozici dva datové soubory obsahující výkony v dané disciplíně pro sledovanou experimentální skupinu (46 vojáků), u níž byla tělesná příprava prováděna novými alternativními postupy, a pro kontrolní skupinu (57 vojáků), kde probíhala tělesná příprava obvyklým způsobem. Vstupní výkonnostní úroveň obou skupin v uvedené disciplíně byla shodná.

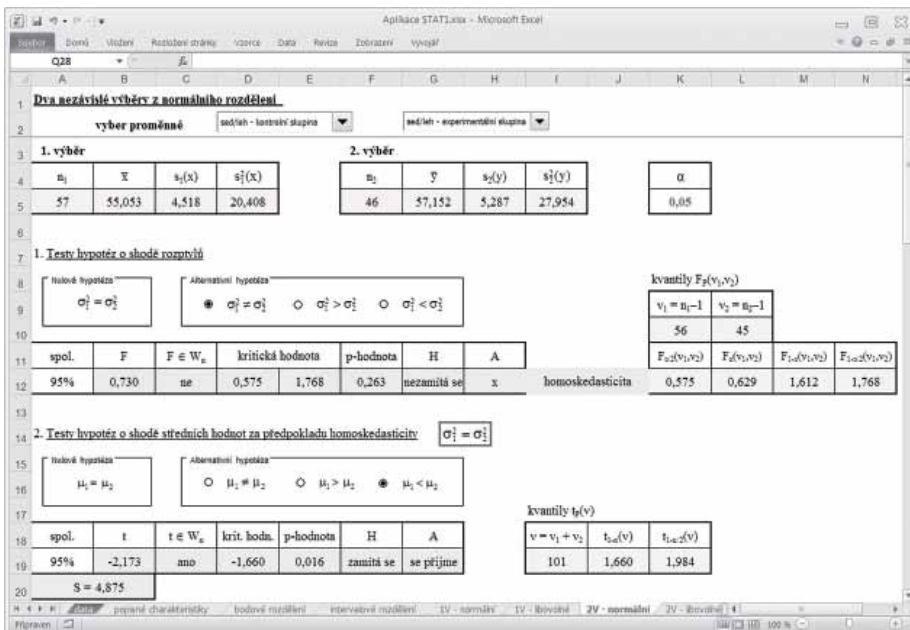
Ke statistickému řešení daného problému užitíme aplikaci STAT1. Oba datové soubory vložíme do datového listu pod názvy sed/leh – experimentální skupina a sed/leh – kontrolní skupina, viz obr. 1. Výhodiskem pro řešení úlohy bude exploratorní analýza dat, kterou provedeme na obou souborech. Vzhledem k povaze dat provedeme intervalové rozdělení četností (list intervalové rozdělení), ze kterého je patrné (viz obr. 2), že data jsou homogenní, rozdělení téměř symetrické a bez odlehlých hodnot, součástí jsou i výběrové charakteristiky (na obr. 2 je uvedeno intervalové rozdělení četností pro data kontrolní skupiny, pro intervalové rozdělení četností dat sed/leh – experimentální skupina platí totéž). Dále lze pro oba výběry ověřit, že pochází z normálního rozdělení (pro data sed/leh – experimentální skupina viz obr. 3).



Obr. 3: Ověření normality

Řešení dané úlohy má charakter dvouvýběrového problému, v rámci kterého je nutné porovnat střední hodnoty výkonů u experimentální a kontrolní skupiny. Zavedení inovované tělesné přípravy se reálně projevilo tak, že se zvýšila hodnota výběrového průměru u experimentální skupiny vzhledem ke kontrolní skupině (obr. 4, str. 120). Proto budeme testovat shodu středních hodnot $\mu_1 = \mu_2$ proti alternativní hypotéze $\mu_1 < \mu_2$. Obě proměnné mají normální rozdělení, použijeme tedy list 2V-normální, v jehož horní části provedeme výběr obou proměnných a zvolíme rovněž hladinu významnosti

testu $\alpha = 0,05$. Nejprve je nutné provést test na shodu rozptylů, např. v [2], a v závislosti na jeho výsledku pokračujeme testem o shodě středních hodnot (předpokládáme homoskedasticitu, tedy stejné rozptyly, neboť test tuto shodu nezamítá). Výsledkem je konstatování, že na hladině významnosti 5 % se hypotéza o shodě obou středních hodnot zamítá, viz obr. 4. Prakticky to tedy znamená, že s 95% spolehlivostí můžeme tvrdit, že inovovaná forma tělesné přípravy vede k lepším výkonům v disciplíně sed-leh.



Obr. 4: Dvouvýběrový test o shodě středních hodnot

Otevírá se samozřejmě prostor pro formulace dalších praktických problémů, které lze na základě našich obou měření řešit. Například je možné využít list 2V-párový test pro posouzení progresu ve výkonnosti vybrané skupiny za dané období apod.

Při řešení následující úlohy použijeme chí-kvadrát test nezávislosti v kontingenční tabulce, který lze v praxi využít např. při zpracování nejružnějších typů dotazníkových šetření. Jedná se o vyhodnocení odpovědí na jednu konkrétní otázku z dotazníku, viz [6], určeného pro sběr empirických dat, které budou podkladem pro vyhodnocení stavu dalšího profesního vzdělávání skupiny personalistů společných sil AČR.

V kontingenční tabulce (str. 121) jsou shrnuty odpovědi 112 náhodně vybraných respondentů na otázku „Je součástí plánování vašeho kariérního rozvoje plán dalšího profesního vzdělávání?“. Úkolem bude na hladině významnosti 5 % prokázat závislost mezi popsaným způsobem plánování kariérního rozvoje (respondenti vybrali jednu nabídnutou odpověď na škále určitě ano – spíše ano – spíše ne – určitě ne) a dosaženým stupněm vzdělání respondentů.

Nejprve je zapotřebí sloučit vhodné kategorie sledované veličiny tak, abychom dosáhli dostatečného počtu odpovědí ve všech buňkách kontingenční tabulky a splnili tak nutnou podmínku pro provedení chí-kvadrát testu v kontingenční tabulce, viz [4, 5].

Tab.: Kontingenční tabulka

vzdělání	určitě ano	spíše ano	spíš ne	určitě ne	celkem
SŠ	13	17	10	11	51
VŠ	36	14	9	2	61
Celkem	49	31	19	13	112

V našem případě je tato podmínka splněna. Hodnoty zapíšeme do listu Kontingenční tabulka – část Empirické četnosti (viz obr. 5). Závěrem je konstatování, že na zvolené hladině významnosti 5 % je závislost v kontingenční tabulce statisticky významná, neboli s 95% spolehlivostí můžeme tvrdit, že nejvyšší dosažené vzdělání personalistů AČR má vliv na způsob plánování kariérního rozvoje.

Aplicace STAT1.xlsx - Microsoft Excel

1 χ^2 - test nezávislosti v kontingenční tabulce

2

3 Empirické četnosti

	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	$n_{i\cdot}$
x_1	13	17	10	11							51
x_2	36	14	9	2							61
x_3											0
x_4											0
x_5											0
x_6											0
x_7											0
x_8											0
x_9											0
x_{10}											0
$n_{\cdot j}$	49	31	19	13	0	0	0	0	0	0	112

Nulová hypotéza: Znaky X a Y jsou nezávislé
 Alternativní hypotéza: Znaky X a Y jsou závislé

spolehlivost	χ^2	χ^2 e W_e	st. volnosti	krit. hodnota	p-hodnota	H	A
95%	16,609	ano	3	7,815	0,001	zamítna se	se přijme

hladina významnosti: 0,05

Závislost v tabulce je statisticky významná

Obr. 5: Test nezávislosti v kontingenční tabulce

Závěr

V současnosti se práce s reálnými daty stává každodenní záležitostí a statistické zpracování datových souborů je nezbytné při nejrůznějších činnostech lidského konání. Na trhu je k dispozici řada softwarových produktů, pomocí nichž je možné tyto statistické analýzy provádět. K nim lze řadit i nově vzniklou aplikaci STAT1, která pracuje pod Microsoft Office Excel. Slouží pro základní zpracování dat prostřednictvím exploratorní analýzy dat, metod jednorozměrné indukční statistiky, dále jsou zde implementovány dvouvýběrové testy a chí-kvadrát test nezávislosti v kontingenční tabulce. Aplikace poskytuje řadu užitečných výstupů v podobě tabulek, grafů a statistických závěrů. Prakticky jakoukoliv část výstupů je možné ihned přenášet do textového dokumentu, např. Wordu.

Přidanou hodnotou je nesporně fakt, že se jedná o uživatelsky nenáročný a volně dostupný nástroj. Stávající analýzy pokrývají základní statistické postupy. Aplikace STAT1 je otevřena k dalšímu rozšiřování své funkčnosti.

Literatura:

- [1] NEUBAUER, J. - SEDLAČÍK, M. - KRŽÍŽ, O. *Aplikace STATI*. 2012. Dostupná na [www: <http://k101.unob.cz/stat1/ >](http://k101.unob.cz/stat1/).
- [2] NEUBAUER, J.- SEDLAČÍK, M. - KRŽÍŽ, O. *Základy statistiky*. 1. vyd. Praha: Grada, 2012, 240 s. ISBN 978-80-247-4273-1.
- [3] ANDĚL, J. *Základy matematické statistiky*. 1. vyd. Praha: Matfyzpress, 2005, 358 s. ISBN 80-86732-40-1.
- [4] BISHOP, Y. M. M.- FIENBERG, S. E.- HOLLAND, P. W. *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, 1975, ISBN 978-0-262-02113-5.
- [5] MANN, P. S. *Introductory Statistics*. 6. vyd. Hoboken: JohnWiley & Sons, 2007, ISBN 978-0-471-75530-2.
- [6] KUBÍNYI, L. *Dotazník pro vyhodnocení stavu dalšího profesního vzdělávání skupiny personalistů společných sil AČR*. 2012.

Jaký máte názor na profesionalizaci armády?

Professionalizace armády byla nutná. Došlo ke snížení délky základní vojenské služby na jeden rok a objevovaly se tendence dalšího snižování až o šest měsíců. Dnes však jsou velmi složité zbraňové systémy, které vyžadují specialisty profesionály připravené pro jejich ovládnutí, což není možné zvládnout během jednoho roku základní vojenské služby. Navíc by bylo v té době nutné změnit zákon, neboť nám legislativní rámec neumožňoval vysílat vojáky základní služby do zahraničí, nesměli totiž opustit území republiky. Zejména po vstupu do NATO byl naplněn předpoklad, že by vojáci měli být v rámci plnění aliančních závazků součástí jednotlivých bojových uskupení v zahraničí. Na základě toho bylo nutno vytvořit podmínky pro naši aktivní účast v zahraničních misích. V neposlední řadě je potřeba zmínit finanční zdroje. Profesionální armáda rozhodně není levná, nevyžaduje však takový počet vojáků jako naše předlistopadová armáda a je nutná. Šlo o krok správným směrem.

Jak vnímáte pozici armády v současné společnosti v České republice?

Velmi pozitivně. Poslední nezávislé výzkumy ukazují, že armáda má 60% důvěryhodnost, což je nejvyšší číslo ze všech státních organizací, a zároveň důkaz dobré práce našich vojáků nejen doma, ale i v zahraničí. V souvislosti s plněním závazků vůči našim partnerům se aktivně zapojujeme do zahraničních operací, kde jsme velmi dobře hodnoceni. Dokonce pozorují, že situace se pomalu otáčí v tom smyslu, že před časem jsme byli velice pozitivně vnímáni v zahraničí, čemuž ovšem neodpovídal obraz armády doma – v poslední době se setkáváme s kladnými názory i u naší veřejnosti. Armáda dnes má co nabídnout a má kvalitní, profesionální vojáky.

Nakolik jsou důležité zmíněné zahraniční mise českých vojáků?

Musím říci, že mise tvoří velmi důležitou součást kariéry vojáka z povolání. Z hlediska celospolečenského je nutno si uvědomit, že jsme součástí NATO. Dnes vzhledem k počtům nejsme schopni republiku ubránit sami a spoléháme na to, že v případné nutnosti budeme součástí aliančního uskupení, které by působilo v blízkosti naší země, nebo přímo na našem území. I z hlediska závazků vůči NATO a EU je tedy velký předpoklad naší účasti na zahraničních misích.

Navíc v porovnání s obdobím studené války a tzv. bipolárního světa je dnes vyšší riziko válečného konfliktu, protože svět je multipolární a na scéně se objevují státy, které chtějí získat např. jadernou převahu. Prostředí je také asymetrické, povstalců nejsou nutně v uniformách apod., na což musíme vojáky připravovat. A zkušenosti se získávají pouze v praxi, při operacích, které jsou nezbytnou součástí kariéry vojáka.

**Z rozhovoru s genmjr. Ing. Miroslavem Žižkou, prvním zástupcem NGŠ AČR.
Zpravodaj 2013 [ročenka VoZP ČR 2013] MK ČR E 15576.**