NMST551 STATISTICKÝ PROJEKTOVÝ SEMINÁŘ

ZÁSADY PSANÍ VĚDECKÝCH ČLÁNKŮ A VÝZKUMNÝCH ZPRÁV

1 Úvod

Vytváření písemných a grafických výstupů, dokumentů, článků a zpráv je přirozenou součástí práce (nejen) každého vědce. U statistika, který se zabývá aplikacemi statistických metod pro řešení problémů z jiných věd, oborů a oblastí lidské činnosti, je schopnost srozumitelně a kvalitně presentovat výsledky své práce přímo klíčová. Nic totiž není platná sebelépe provedená analýza, když její závěry a jejich význam nikdo nepochopí.

Psaní článků je spíše umění než věda a jakékoli rady a pravidla týkající se tohoto tématu musí být nutně pouze rámcové. Každý článek má svůj zvláštní účel a je psán pro určitý typ čtenáře. Některé články se musí podřídit vnějším pravidlům jako jsou formátové požadavky vědeckých časopisů, omezení počtem stránek/slov, státní normy apod. Většina konvencí ohledně psaní článků je čistě zvyková a do značné míry subjektivní a vychází nejen z logiky věci, ale i z estetického cítění autora. Proto většina rad a zásad obsažených v tomto dokumentu nemá být chápána dogmaticky, ale spíše jako doporučení, která je možno přizpůsobovat situaci.

Nezkušený autor má většinou tendenci psát článek jako popis své vlastní zkušenosti. Vysvětluje, co udělal nejdříve a co na to navazovalo, jak provedl to či ono, zmíní to, co mu připadalo zajímavé, pomine to, co mu je zřejmé. Píše to prostě ze svého pohledu, svým vlastním jazykem. Výsledkem je výplod, který je pro kohokoli jiného nečitelný a nepochopitelný. Zkusme to ilustrovat výňatkem z fiktivní výzkumné zprávy, jakou by mohl napsat hodně nezkušený statistik:

> *Nejdříve jsem načetl data. Byla tam nějaká chybná pozorování, tak jsem je vyhodil. Všechny NA jsem taky vyhodil. Některé veličiny jsem převedl na faktory, zvláště* `rggrp`, `msmt` *a* `sex.prd`. *Udělal jsem si průměry a u veličiny* `wrcpt111` *to vyšlo 111.263345, u veličiny* `wrcck112` *bylo 2.8363E−03 a u veličiny* `wrcpt112` *to dalo 109.276200. Udělal jsem si plot, na kterém bylo vidět, že některá data jsou normální a jiná nejsou. Použil jsem funkci* `shapiro.test` *a vyšlo 0.9402, hypotézu jsem zamítl. Fitnul jsem model* $Y_{ijk} = a_j^{\gamma} + x_{ij}(c_{ij} - \delta)^{\beta} + \varepsilon$. *Parametry vyšly* $\gamma =$ *9.2376E−07,* $\delta =$ $-$`Inf` *a* $\beta = 0.14423$. *Na základě testu jsme hypotézu nemohli zamítnout. Namaloval jsem si vlivná pozorování a žádná jsem nenašel. Ale z obrázku jsem viděl, že výsledky nejspíš nemají Weibullovo rozdělení. Možná bychom se mohli domnívat, že to asi nebylo slabě stacionární.*

Tento odstrašující příklad je samozřejmě přehnaný, ale ne zcela nereálný. Cílem tohoto dílka je poradit, jak se vyhnout nejzávažnějším chybám při technickém psaní a jak vytvořit zprávu, která se snadno čte, je srozumitelná, budí dobrý dojem a slouží svému účelu.

2 Struktura článku

Články, které spočívají v presentaci statistických analýz konaných za určitým účelem a činí z jejich výsledků konkrétní závěry, mívají následující stavbu:

1. Abstrakt
2. Úvod
3. Metody
4. Výsledky
5. Závěr

Proberme si teď stručně, co by mělo být obsaženo v jednotlivých částech článku.

Abstrakt by měl stručně shrnout celý obsah článku. Obvyklý rozsah abstraktu je jeden odstavec až půl stránky. Abstrakt má jednou nebo dvěma větami vystihnout podstatu každé zbývající části článku, tj. zahrnout to nejdůležitější z úvodu, metod, výsledků a závěrů. Abstrakt se vždy píše nakonec.

Úvod má vysvětlit, jaký problém se řeší, podat ho v širším kontextu a zformulovat konkrétní otázky, které má článek vyřešit. Je zde obvykle vysvětlena studovaná problematika na úrovni odpovídající znalostem předpokládaného čtenáře. Je zde shrnuto, co je o daném problému již známo, kdo na něm pracoval a k čemu dospěl. Hojně se cituje publikovaná literatura. Autor by měl zdůvodnit, proč je studovaný problém důležitý a co jeho řešení přinese. V této části článku je třeba co nejpřesněji a co nejkonkrétněji specifikovat otázky, na něž se článek bude snažit odpovědět. Otázek může být i víc, ale měly by být rozděleny na 1–2 otázky primární, na nichž je soustředěn hlavní zájem, a otázky podružné, sekundární. Je velmi důležité, aby otázky specifikované v úvodu článku byly stanoveny před provedením analýzy. Pokud se teprve během analýzy ukáže zajímavá odpověď na otázku, kterou si předtím nikdo nekladl[1], nemůže být průkaznost takového výsledku kladena naroveň odpovědím na otázky předem plánované.

Metody vysvětlují veškerou metodiku studie. Jedná-li se o experiment, je v metodách popsáno, jak byl proveden. Jedná-li se o pozorovací studii, je objasněno, jak byli vybíráni její účastníci a z jaké populace pocházeli. Je třeba vysvětlit metody sběru dat a metody měření důležitých veličin. Obsahují-li data chybějící pozorování nebo jiné zásadní problémy, autor by měl objasnit, jak časté byly a jak s nimi bylo naloženo. Popisují se zde i statistické metody použité k analýze. Musí být uvedeno, na jakých podskupinách pozorování byly jednotlivé analýzy provedeny, jak byly kódovány veličiny, jaké veličiny byly zahrnuty v regresních modelech a jak byly ztransformovány, jaké testy byly provedeny a jak byly počítány intervaly spolehlivosti. Uvádí se tu též, v jakém softwaru byly analýzy implementovány. Metody by měly být popsány tak podrobně, aby podle tohoto popisu mohl někdo jiný celou studii, sběr dat a jejich analýzu zopakovat. Metody je třeba popisovat stručně a přesně, střídmým technickým jazykem,

[1] Typicky může nastat třeba toto: Provedeme-li plánovanou analýzu na celém souboru, nevidíme žadný efekt. Ale podíváme-li se na muže starší 55 let pracující v bankovním sektoru, zjeví se významný rozdíl ve zkoumané veličině.

s jednoznačně definovanými pojmy. Do metod nepatří úvahy nad alternativními způsoby provedení ani zdůvodňování proč bylo provedeno zrovna toto a ne něco jiného. Do metod též nepatří presentace jakýchkoli výsledků s výjimkou informace o velikosti populací, množství vynechaných pozorování, případně o velikosti analyzovaných podskupin.

Výsledky obsahují soupis výsledků vybraných k presentaci v článku. Obvykle se takové presentace dočká jen velmi malá část výsledků, které statistik během analýzy vyprodukuje. Výsledky by měly začínat popisem analyzované populace formou popisných statistik. Tato část je důležitá kvůli tomu, aby bylo zřejmé, na jaké populaci byla studie provedena a aby si čtenář mohl učinit představu o vhodnosti jejího zobecnění na populace jiné. Dále by měly následovat deskriptivní analýzy, které se snaží naznačit odpověď na zkoumané otázky pomocí jednoduchých tabulek a grafů. Interpretace deskriptivních analýz musí být opatrná; rozdíly v průměrech či odlišnost v grafu neznamenají nutně, že porovnávané skupiny se skutečně liší. Zde je na místě používat slova jako *„může být“*, *„zdá se“*, *„graf naznačuje, že“* apod. Definitivní potvrzení později poskytnou hlavní analýzy. Pokud ovšem hlavní analýzy dojdou k odlišným závěrům, než které lze vytušit z tabulek a grafů, a autor nenajde pro tento fakt přijatelné vysvětlení, mohou být výsledky hlavních analýz a celého článku nahlíženy s nedůvěrou. Výsledky hlavních analýz obvykle uvádíme po deskriptivních analýzách. Je třeba je ihned interpretovat v kontextu zkoumaného problému a objasnit, jakým způsobem tyto výsledky odpovídají na otázky specifikované v úvodu článku. Po hlavních analýzách mohou následovat výsledky vedlejších analýz, jsou-li jaké, případně různé neplánované doplňkové analýzy.

Účelem **diskuse** je shrnout odpovědi na otázky položené v úvodu, vyhodnotit spolehlivost těchto odpovědí a jejich praktický význam, uvést je do kontextu problému a porovnat je s dříve známými výsledky a fakty (včetně těch zmíněných v úvodu). Diskuse může probírat různé aspekty analýz, které mohly ovlivnit výsledky, zdůvodňovat vhodnost použitých metod, vysvětlovat silné a slabé stránky použitého přístupu, upozorňovat na potenciální skryté problémy, shrnovat výsledky vedlejších nebo alternativních analýz, které nebyly vybrány k presentaci v článku, vyslovovat hypotézy k ověření v budoucích studiích a naznačovat směry dalšího výzkumu. Diskuse je nesmírně důležitá, protože uvádí na správnou míru interpretaci obdržených výsledků v celém kontextu studovaného problému a formuluje jejich konkrétní praktické důsledky. Diskuse bývá často nejdelší částí celého článku. Diskuse je místo, kde můžeme vyslovit pochybnosti nad vhodností zvolených metod, nebo je naopak obhájovat, kde můžeme upozornit na nejistotu ohledně splnění předpokladů použitých metod a pokázat na možné důsledky jejich porušení, kde si můžeme dovolit i trochu zaspekulovat. Diskuse by však měla vždy obsahovat nějaký konkrétní závěr, a to buď na samém začátku nebo naopak na konci.

Je-li to možné a vhodné, můžeme k článku přiložit **dodatky**, obsahující podrobnosti, jež nejsou zajímavé pro všechny čtenáře a nehodí se pro zahrnutí do článku. Mohou to být podrobnější popisy metod, rovnice použitých modelů, matematické vzorce, na něž nelze odkázat do literatury, podrobnější tabulky a grafy výsledků, podrobnosti o softwarové implementaci metod, nebo výňatky ze softwarového kódu. Článek musí být v každém případě srozumitelný a úplný i bez studování dodatků.

3 Pravidla presentace výsledků

V zásadě máme tři možnosti, jak presentovat konkrétní výsledek statistické analýzy (odhad parametru, výsledek testu apod.): v textu článku, v tabulce anebo v grafu. Některé zásady bychom však měli dodržovat bez ohledu na formu presentace výsledku. Uveďme si nejdříve některé z oněch obecných zásad.

- Parametry modelů uvádíme v podobě, která má nějakou praktickou interpretaci, a tuto interpretaci slovně popíšeme. Nelze napsat *„odhad parametru δ byl 0.0012“*. Musíme si promyslet, co parametr δ znamená, převést ho na co nejsrozumitelnější škálu a napsat třeba *„odhadnutý relativní nárůst potenciálu za jednu sekundu byl 0.12%“*.
- Fysikální rozměr parametru nesmíme opomíjet. Pokud má např. odhad parametru u prediktoru *výška* v lineárním regresním modelu pro hmotnost hodnotu 0.877, pak jej vypisujeme jako 0.877 kg/m.
- Počet platných číslic či desetinných míst volíme podle kontextu, nikoli podle toho, co zobrazí počítač. Zřídkakdy je rozumné uvádět více než tři platné číslice. Počet desetinných míst však musí být konsistentní u výsledků, které se čtou společně (např. jeden sloupec tabulky). Píšeme tedy odhad a interval spolehlivosti takto: *1.24 (0.00 – 2.48)*, nikoli takto *1.2 (0.0006 – 2.5)*.
- Bodové odhady parametrů, zvláště těch, které jsou podstatné pro účel analýzy, doprovázíme intervalem spolehlivosti. Bodové odhady bez intervalů bychom měli zpravidla uvádět pouze v tabulkách deskriptivních statistik.

Jako kritérium výběru mezi presentací v textu, tabulce či v grafu poslouží nejlépe plošná úspornost jednotlivých variant. Mám-li na sloupcovém diagramu tři sloupce, spotřebuji tím vlastně půl stránky na presentaci tří čísel. V takovém případě je nejvhodnější uvést tato tři čísla přímo v textu článku. Tabulka s dvěma řádky a dvěma sloupci je také velmi neúsporná, čtyři čísla v textu zaberou mnohem méně místa. Tabulky se obecně hodí pro přehlednou systematickou presentaci většího počtu číselných výsledků, které je třeba porovnávat. Grafy se hodí tam, kde je třeba porovnat ještě větší počet číselných výsledků, jejichž konkrétní číselné hodnoty není třeba přímo uvádět a pro něž lze najít takový způsob grafické presentace, jenž usnadní jejich vizuální porovnání. Grafy se také hodí pro zobrazení složitějších funkcí, které nelze srozumitelně popsat uvedením jednoduše interpretovatelných parametrů, a pro ilustraci interakcí v regresních modelech. Používám-li například kvadratickou regresi, je lepší odhadnutou parabolu nakreslit do grafu, než uvádět odhadnuté hodnoty regresních koeficientů — ty mají v tomto případě velmi nesnadnou interpretaci.

Používání tabulek a grafů v článcích má některá společná pravidla a některá specifická. Tabulky a grafy neuvádíme přímo do textu článku, ale umístíme je buď na samostatné stránky nebo na vyhrazené místo v horní nebo dolní části běžných stránek. LaTeX, na rozdíl od MSWordu, se o umístění plovoucích grafů a tabulek postará automaticky. Každý graf a tabulku očíslujeme a zařadíme k nim legendu (u grafů obvykle pod obrázkem, u tabulek nad). Legenda musí popisovat obsah grafu či tabulky tak podrobně, aby jim čtenář rozuměl bez důkladného studování textu článku. Na každou tabulku a graf musí být v textu odkaz pomocí jejich čísla. Na příslušném místě textu pak shrneme ty nejdůležitější závěry, které lze z tabulky či grafu učinit. Text by

měl být čitelný a srozumitelný i bez prohlížení tabulek a grafů a tabulky a grafy by měly být srozumitelné i bez podrobné četby textu. Na tabulky a grafy odkazujeme pokud možno nepřímo v průběhu běžného toku textu; místo *„Tabulka 1 ukazuje, že muži jsou v průměru vyšší než ženy“* raději napíšeme *„Muži byli významně vyšší než ženy (viz Tabulka 1)“*. Krátký článek (rozsahem do 10 stran A4) by měl obsahovat nejvýše 5–8 tabulek a grafů celkem.

U **tabulek** se doporučuje dodržovat následující pravidla:

- Vyhýbat se svislým linkám. Silnějšími vodorovnými linkami oddělit tabulku od okolního textu včetně legendy, slabšími vodorovnými linkami oddělovat záhlaví sloupců od těla tabulky a jednotlivé části tabulky mezi sebou. V LaTeXu tuto podobu tabulek implementuje balík `booktabs`. Chceme-li výrazněji oddělit některé sloupce od jiných, vložíme mezi ně větší mezeru.
- Neměnit typ, formát a význam obsahu políček v tomtéž sloupci (není dobré do téhož sloupce zapisovat tu průměr onde procenta).
- Neopakovat tentýž obsah políček mnohokrát za sebou. Máme-li sloupec *Rozptyl*, který v prvních deseti řádcích obsahuje hodnotu 0.5 a v druhých deseti řádcích hodnotu 1.5, pak tento sloupec raději zrušíme a vyřešíme to jinak. Například můžeme tabulku rozdělit na dvě nebo do ní vložit popisné řádky, které informují o nějaké proměnné hodnotě opakující se v následujícím oddíle tabulky (např. *„Rozptyl = 0.5“* a níže *„Rozptyl = 1.5“*).
- Čísla v tabulce zarovnávat na desetinnou tečku.
- V tabulce je někdy potřebné používat zkratky, které se jinde v článku nevyskytují. Tyto zkratky můžeme vysvětlit v legendě nebo v poznámkách pod tabulkou. Poznámky pod tabulkou můžeme využít i k podrobnějšímu vysvětlení významu některých sloupců nebo hodnot.

Nakonec zmiňme několik rad týkajících se **grafů**.

- Graf by měl být vytvořen ve velikosti, v níž bude použit ve článku. Zmenšení příliš velkého grafu vede ke špatné čitelnosti popisků.
- Osy grafu musí být řádně popsány ve stejném jazyce, v jakém je psán článek[2]. Kreslíme-li graf hmotnosti proti výšce, nenecháme na nich popisky `ht` a `wt`, ale osy popíšeme *Výška [cm]* a *Hmotnost [kg]*.
- Chceme-li na dvourozměrném grafu vyznačit hodnoty jednotlivých pozorování, dáme pozor, aby se neslily do jednolité černé tmy. Je-li pozorování mnoho, zmenšíme velikost symbolu, kterým je vykreslujeme, anebo raději náhodně vybereme malou část pozorování, kterou do grafu zaneseme. Grafy, které obsahují tisíce pozorování, dělají problémy hlavně v elektronických dokumentech, protože výrazně zvětšují velikost souborů.[3]
- Je-li dokument určen k tisku, vyhýbáme se používání barev. Čáry rozlišujeme typem (plná, tečkovaná, čerchovaná,...), plochy dostatečně rozdílnými intensitami šedé nebo šrafováním. Význam jednotlivých typů čar a ploch vysvětlíme buď v textové legendě ke grafu anebo v grafické legendě, která je přímo součástí obrázku.

[2] Absenci diakritiky však lze jistě tolerovat.

[3] Jeden takový graf může mít klidně několik MB.

4 Jazykové a typografické rady

Článek by měl být napsaný souvislým textem, bez heslovitých vsuvek a bez použití odrážek a číslovaných odstavců. Každá věta by měla být úplná, s podmětem a přísudkem. Autor by se měl vyhýbat jak složitým souvětím tak holým větám. Věty a odstavce by na sebe měly logicky navazovat, systematicky probírat jedno téma po druhém, nepřeskakovat a neodkazovat na fakta, která ještě nebyla zmíněna. Odstavce nesmí být příliš krátké (tři řádky je málo) ani příliš dlouhé; ideální délka odstavce je čtvrt až půl stránky. Krátké články (kolem 10 stránek textu) nemají obsahovat víc než jedno členění, tedy žádné podkapitoly ani mezititulky. Styl by měl být střídmý, bez citově zabarvených a expresívních výrazů. Je třeba psát úsporně a vyhýbat se vycpávkovým slovům a větným konstrukcím, které nenesou žádnou informaci. Mezi takové patří zejména většina výskytů slov *„zejména“*, *„takřka“*, *„velmi“*, *„jak vidíme“* , *„jak bylo řečeno“* a podobně. Autor by si měl po sobě přečíst vše, co napsal, a uvažovat nad každým slovem, vsuvkou, vedlejší větou, jestli se změní smysl sdělení, pokud by tam tento kus nebyl. Jestliže se smysl nezmění, jedná se o vycpávku, kterou je třeba smazat.

Článek by měl být napsán celý v jednotném čase; obvykle je vhodnější zvolit čas minulý než čas přítomný. Odkazuje-li autor na sebe, dělá to často v množném čísle i když článek píše sám (*„udělali jsme. . . “*, *„odhadli jsme. . . “*). Někteří tento postup kritisují jakožto nevhodný *plural majestatis*, ale někomu zas mohou připadat tvary *„udělal jsem. . . “*, *„odhadl jsem. . . “* příliš osobní. To záleží na vkusu.

V českém textu by cizí slova měla být používána s mírou, zvláště pokud mají česká synonyma. Obzvlášť odporné je kopírování anglických slov z výstupu počítačových programů (*„intercept“*, *„P-value“*). U odborných termínů raději používáme jednu variantu, i když se tatáž věc dá nazývat více způsoby: např. nestřídáme pojmy *„interval spolehlivosti“*, *„konfidenční interval“* a *„intervalový odhad“*, ale vybereme si jednu verzi a té se držíme. Zkratky vysvětlujeme tak, že při prvním použití uvedeme celý nezkrácený tvar a zkratku dáme za něj do závorky. Vyhýbáme se kryptickým kódům (rozhodně neoznačíme mladší muže za skupinu I a starší ženy za skupinu II, abychom mohli nadále mluvit jen o rozdílech mezi skupinami I a II). Ve článku zásadně nepoužíváme názvy veličin, které jsme měli zavedeny v programech pro analýzu anebo ve vstupních datech.

Na závěr připojujeme přání hodně štěstí při psací a publikační činnosti.

Doplňující literatura

Šesták, Zdeněk. *Jak psát a přednášet o vědě.* Academia, Praha, 2000.

Higham, Nicolas J. *Handbook of Writing for the Mathematical Sciences.* SIAM, Philadelphia, 1993.

8. 10. 2014