Metodologie pro Informační studia a knihovnictví 2

Modul II:

Práce s daty. GIGO. Datová matice.

Metodologie pro Informační studia a knihovnictví 2

Modul II: Úvod do práce s daty

Co se dozvíte v tomto modulu?

- K čemu vám bude statistická analýza?
- Jaké jsou základní druhy analýzy?
- Kde brát data?
- Jak vypadá datová matice?

1 K čemu je mi statistická analýza?

Každý den se setkáváme zejména v médiích s řadou informací, které pocházejí z kvantitativních výzkumů. Pochopení základů statistické analýzy nám pomůže nejen lépe pochopit, jak tyto informace vznikají, ale také je **lépe a kritičtěji interpretovat**. Často se totiž setkáváme se zjednodušenými a někdy i nesprávnými interpretacemi, které například zaměňují příčinu a následek, opomíjejí vliv dalších proměnných, zjednodušují kauzální vztahy.

Rozvod je nakažlivý, zjistili britští experti

7. července 2010 5:38

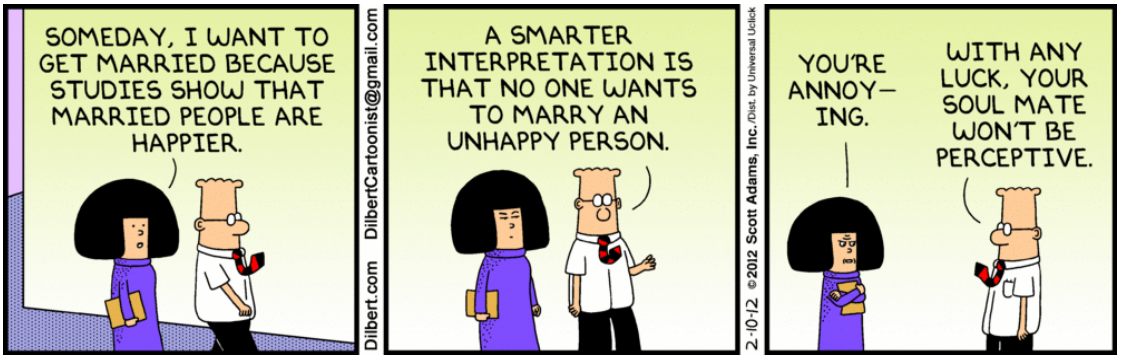
Když se začnou rozvádět nejlepší přátelé, buďte na pozoru. I vašemu vztahu hrozí vysoké riziko rozvodu, zjistili britští experti. Podle nich je rozvod nakažlivý a šíří se jako nějaká nemoc rodinami, pracovním prostředím i skupinou přátel.

Ilustrační foto. | foto: Profimedia.cz

Statistický vtip, který si utahuje z podobných zpráv typu „vědci zjistili, že…“ a který publikoval S. den Hartog ve své dizertační práci odevzdané na Univerzitě v Groningenu, říká:

> *„Je dokázáno, že oslavy narozenin jsou zdravé. Statistici zjistili, že lidé, kteří oslavili více narozenin, se dožívají vyššího věku.“*

A do třetice ukázka podobného vtipu na úkor častých dezinterpretací statistických výzkumů:



Pochopení základů statistiky vám nepomůže ale jen lépe chápat statistické vtipy. **Pomocí statistických metod budete moci například lépe:**

- popsat sledovanou problematiku a výskyt problému v populaci,
- chápat **potřeby své cílové skupiny** (populace),
- rozdělit cílovou skupinu na smysluplné **segmenty** a soustředit na ně cílenou nabídku služeb i marketingovou strategii,
- odhalit **příčiny a následky** jevů,
- spočítat **rizika** spojená se strategickým rozhodováním,
- chápat, jaká čísla a jaké výsledky jsou pro vás skutečně **významné**.

2 Druhy statistické analýzy

Minulý týden jsme se dotkli rozdílů mezi **deskriptivní** a **induktivní** (někdy také inferenční) statistikou.

Deskriptivní statistika se zabývá sběrem, sumarizací a prezentací souborů dat. Je to ta „lehčí" statistika, která je dostupná pomocí běžných nástrojů (kalkulačka, tabulkový procesor).

Pomocí **deskriptivní statistiky** můžeme odpovědět na otázky typu:

- *Jaká je průměrná délka života žen?*
- *Jaká je mediánová hodnota platu knihovníků v ČR?*
- *Jaký je minimální a maximální počet knih, který průměrně za rok přečte student KISKu?*

Induktivní (inferenční) statistika se zabývá zobecňováním výsledků výzkumu na vzorku na populaci. Jinými slovy, pokud vám výsledky ukazují, že z celkového počtu 299 respondentů se 85% inspiruje při výběru knih radou od přátel (to je třeba jeden z výsledků nedávného průzkumu studentek KISKu), induktivní statistika vám pomůže zjistit, s jakou jistotou se toto vaše zjištění dá zobecnit na populaci. Induktivní statistika pracuje s hypotézami a zjišťuje, zda jsou sesbíraná data s těmito hypotézami v souladu.

3 Zdroje dat

V kontextu výzkumů hovoříme o primární a sekundární analýze dat.

Primární analýza

Primární analýza pracuje s originálními daty, která jsme nasbírali přímo pro potřeby výzkumu. Zdrojem kvantitativních dat jsou nejčastěji ***dotazníková šetření*** či výsledky ***experimentálních studií***.

Dotazníkovým šetřením jsme se podrobněji věnovali v předchozím modulu.

Experimentální studie jsou speciální případ výzkumů, kdy se snažíme zjistit vliv jedné proměnné na jiné. Například můžeme srovnávat chyby v bibliografických citacích u studentů, kteří navštěvovali kurz KPM a u studentů, kteří kurz nenavštěvovali. Nemusíme v tomto případě volit jako výzkumnou metodu dotazování, ale podíváme se přímo na citace v závěrečných pracích. V experimentu zkoumáme výzkumnou skupinu, u které se zaměřujeme na to, zda se změna v proměnné (v našem případě absolvování kurzu) promítla i do změny pozorované proměnné (v našem případě správnost citací). Současně si výsledky ověřujeme i na tzv. kontrolní skupině.

Sekundární analýza

Sekundární analýza se soustředí na **analýzu již sesbíraných dat**. Existuje velká množina dat sesbíraných pro účely jiných výzkumů, které se dají využít pro další účely. Zdrojem těchto dat jsou různé výzkumné databáze, ale i webové stránky výzkumných institucí, obrovskou zásobárnou dat jsou instituce veřejné správy.

Hnutí za sdílení výsledků výzkumu se nazývá **open science** či **open data**. Níže naleznete některé příklady zdrojů dat relevantních pro náš obor:

- **Český statistický úřad**

ČSÚ poskytuje informace o státní ekonomice, pohybu osob, srovnání se zahraničím, vědě a výzkumu. Kromě celé řady statistik jsou na stránkách úřadu k dispozici i otevřená data z výsledků voleb.

- **Databáze EUROSTATu**

Databáze EUROSTATu poskytuje informace o regionálních statistikách, ekonomice a financích, průmyslu, obchodu, zemědělství, dopravě, energetice, vědě a technologiích v EU.

- **ČSDA - Český sociálněvědní datový archiv**

ČSDA poskytuje přístup k vybraným českým datovým souborům reprezentativních výzkumů. Bez registrace je možné procházet stránky Webu a informace o archivovaných datech. Varchivu najdete například datové soubory z realizovaných měsíčních šetření Centra pro výzkum veřejného mínění (CVVM).

- **Repozitáře institucí**

*Některé instituce se mohou rozhodnout poskytnout data ze svých průzkumů k dalším účelům. Příkladem takového rozhodnutí v našem oboru je výzkum **SOAP (Study of Open Access Publishing)** o postojích vědců k open access, který realizovali vědci z CERNu. Mezinárodní data obsahující odpovědi tisíců respondentů ve formátech .csv, .xls a .xlsx jsou k dispozici zde. Další repozitáře lze najít např. přes seznam na Datacite nebo přes další služby.*

Poznámka: Právě pro vás připravujeme přehled oborových datových zdrojů s otevřenými daty na http://wiki.knihovna.cz. Až bude zveřejněný, budu informovat! ☺

4 GIGO. Garbage in, garbage out

Úvodní příprava dat na samotnou analýzu není sice nejzajímavější a nejzábavnější částí analytické práce, ale pro kvalitní výsledek je naprosto nezbytná. V této souvislosti se používá

pořekadlo „***garbage in – garage out"*** – pokud jsou na vstupu nekvalitní data, nekvalitní bude i výstup. Proto je v první řadě potřeba věnovat se právě kontrole a čištění dat.

Charles Wheelan (2013) říká, že za každou důležitou výzkumnou studií jsou data, která umožňují analýzu a naopak špatné výzkumy bývají i založené na špatných datech.

Wheelan (2013) identifikuje několik obecných příkladů GIGO:

Zkreslení výsledků kvůli výběru

Možná jste v roce 2008 byli mezi těmi, kdo si byli jistí postupem Strany zelených do krajských zastupitelstev. Stranu zelených totiž volilo hodně lidí z vašeho okolí, tehdejší předseda Bursík předpokládal několikanásobný nárůst počtu zastupitelů, výzkumy veřejného mínění slibovaly Straně zelených zisk 7-9 % voličských hlasů. Po volbách však strana nezískala zastoupení ani v jednom kraji. Podobná situace se opakovala v roce 2010 při volbách do Poslanecké sněmovny. Pokud by řada vysokoškoláků mohla odhadovat výsledky voleb podle statusů a profilových fotografií na Facebooku, Zelení by byli jasnými favority. Přesto ani v těchto volbách nezískali potřebný počet hlasů. Jak je možné, že se tolik lišily odhady (výzkumy) a realita?

Jeden z nejznámějších podobných případů, kde byly špatné výsledky výzkumů ovlivněny špatným výběrem respondentů, a tedy od počátku špatnými daty, byl příklad předvolebního průzkumu, který v roce 1936 realizoval časopis *Literary Digest*. Časopis oslovil před volbami 10 milionů amerických voličů s otázkou, zda budou volit republikána Alfa Landona či demokrata Franklina Roosevelta. 10 milionů je obrovský vzorek, a tak se výsledkům šetření, které přiřkly 57 procent Landonovi, přikládala velká váha. Velký problém byl však ve výběru respondentů. Literary Digest totiž oslovil své předplatitele a také majitele telefonních přístrojů a automobilů (jejich adresy totiž byly veřejně dohledatelné). Pro autory šetření byly potom velkým překvapením výsledky voleb, kde se 60 % zvítězil Franklin Roosevelt. Ukázalo se, že vzorek, byť velký, nebyl v žádném případě reprezentativní vzhledem k celé americké populaci – předplatitelé časopisu Literary Digest, stejně jako majitelé telefonů a vozů, patřili mezi bohatší část společnosti a nebyli rozhodně obrázkem „průměrného Američana". Wheelan dodává: *„Čím větší jsou dobře sestavené vzorky, tím lépe, protože se zmenšuje riziko chyby. Čím větší jsou špatně sestavené vzorky, hromada smetí (garbage) pouze narůstá a čím dál víc zapáchá"* (s. 119).

Speciálním případem zkreslení výsledků kvůli nesprávně vybranému vzorku jsou **ankety** a tzv. **samovýběry**, například dobrovolnické studie. Dobrovolníci, kteří jsou ochotni se přihlásit např. do výzkumu sexuálního chování, nemusí reprezentovat sexuální chování celé populace. Riziko špatného výběru při samosběru se může ještě zvýšit, pokud nabízíme za zapojení do výzkumu odměnu.

Zkreslení výsledků pro publikování

Wheelan (ibid) upozorňuje, že pozitivní výsledky studií mají větší šanci na opublikování, protože jsou tak zajímavější. *„Pokud si vezmete 100 statistických šetření, je pravděpodobné, že jedno z nich bude mít vlastně nesmyslné výsledky – například statistickou asociaci mezi hraním videoher a výskytem rakoviny střev. A tady je ten problém: zatímco 99 studií, které dokázaly nulovou závislost mezi hraním her a rakovinou střev, nebude nikdy publikováno, protože výsledky nejsou dost zajímavé, jediná studie s pozitivními výsledky půjde do tisku a bude se jí věnovat další pozornost"* (s. 121).

Tento efekt byl popsán například u publikování výsledků studií účinnosti léků na depresi – u studií, které dokazovaly účinnost léku, byla publikována velká část, zatímco studie s nepozitivními výsledky vydávány nebyly.

Zkreslení výsledků kvůli paměti

Velká část šetření je založena na zjišťování reálních zážitků a chování respondentů. Ukazuje se ale, že paměť je velmi složitý mechanismus. Wheelan (ibid) zmiňuje harvardskou studii, ve které se vědci dotazovali žen s rakovinou prsu na jejich stravovací návyky. Ukázalo se, že ženy, které onemocněly rakovinou prsu, vykazovaly ve studii větší sklon k předchozí konzumaci tučných jídel oproti zdravým ženám. Ve skutečnosti se však nejednalo o studii závislosti konzumace tuku a výskytu rakoviny prsu, ale o výzkum toho, jaký vliv má onemocnění rakovinou prsu na paměť. Všechny ženy podstoupily dotazování na stravovací návyky léta předtím, než jim byla rakovina diagnostikována. Srovnání výsledků prvního dotazování založeného na měření reálného aktuálního chování a druhého šetření zjišťujícího stejné chování v minulosti, ukázalo, že fakt onemocnění má vliv na to, jak si ženy „převyprávěly" svou minulost vlivem hledání příčin onemocnění.

Tento druh zkreslení je tedy velkým rizikem studií, které zjišťují minulé chování.

Survivorship bias – „klam přeživších"

Tzv. klam přeživších je chybou, která je založena na vyšší viditelnosti těch, kteří „přežili" určitý proces. Je sofistikovanější obdobou „mazáckých pouček typu „maturita je hračka". Například pokud bychom zjišťovali spokojenost se studiem na KISKu na absolventech našeho oboru, dobrali bychom se pravděpodobně jiných čísel, než kdybychom zjišťovali spokojenost se studiem mezi všemi studenty, tedy i těmi, kteří z nějakého důvodu studium nedokončili. Klam přeživších tedy může často vést k optimističtějším závěrům.

Klam zdravého uživatele

Tzv. klam zdravého uživatele byl popsán v epidemiologii.

- Do výzkumných studií o zdraví se například hlásí obecně zdravější lidé – prostě proto, že se více zajímají o zdraví.
- Lidé, kteří berou vitamíny, jsou zdravější. Prostě proto, že je to *ten druh lidí*, kteří berou pravidelně vitamíny (tito lidé ale také pravděpodobnější pravidelně sportují, sledují své zdraví a věnují se prevenci).

Do vztahů, které mezi proměnnými sledujeme, zkrátka vstupují ještě další proměnné, a ty je potřeba hlídat. Jinak se nemůžeme vyvarovat omylů, které se dají shrnout pod heslo ***„garbage in – garbage out“***.

Vliv nepozorovaných proměnných

Disman (2002) ukazuje, že do analýzy mohou vstupovat další proměnné s rizikem ovlivnění výsledků. Tato rizika je potřeba hlídat:

1. **Nepravá korelace**. Ačkoliv se může zdát, že proměnná A ovlivňuje proměnnou B, může existovat ještě třetí nepozorovaná či neanalyzovaná proměnná C, která ovlivňuje A i B.
   (C → A ^ C → B)
2. **Vývojová sekvence.** V tomto případě se nám opět zdá, že proměnná A ovlivňuje proměnnou B a může tomu skutečně tak být. Co však nepozorujeme, je proměnná 0, která ovlivňuje proměnnou A.
   (0 → A → B)
3. **Chybějící střední člen.** Tato situace nastává, pokud jsme do analýzy nezařadili proměnnou, která je ovlivňována proměnnou A a dále ovlivňuje proměnnou B.
   (A → X → B)
4. **Dvojí příčina.** Závislá proměnná B může mít více příčin, ale ne všechny jsou zahrnuty do výzkumu.
   (A+X+Y→ B)

Zdroje chybných dat při zápisu

Chyby v datech mohou vznikat i při zápisu do datového souboru. Obvykle se jedná o posuny desetinných čárek, záměnu znaků či další chyby při přepisování (například záměna „O“ a „0“).

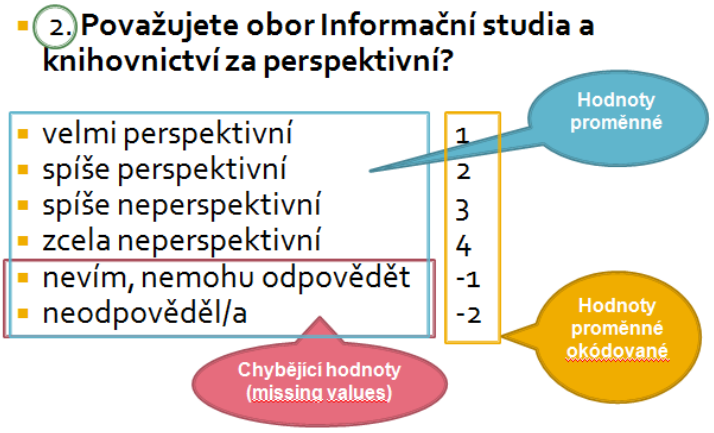
Pokud vás téma chyb v analýze zaujalo, přečtěte si třeba článek ***Why Most Published Research Findings Are False?***

5 Práce s datovým souborem

Datové soubory (matice) mají specifickou podobu. V tabulce se zapisují respondenti do jednotlivých řádků, kde každý sloupec představuje jednu proměnnou.

Pro práci s velkým množstvím dat a pro práci ve specializovaných softwarech se využívá kódování hodnot proměnných. Okódovaná otázka může vypadat například takto:

Příklad kódování jednoduché otázky



Příklad kódování baterie otázek

Spokojenost s nabídkou kurzů	Velmi souhlasím	Spíše souhlasím	Ani souhlasím, ani nesouhlasím	Spíše nesouhlasím	Vůbec nesouhlasím	Nevím / nemohu odpovědět
Povinné (A) kurzy mají logickou časovou posloupnost.	1	2	3	4	5	-1
Obsahy jednotlivých povinných (A) kurzů se nepřekrývají.	1	2	3	4	5	-1
Jsem spokojen/a s tematickou šíří nabídky povinně volitelných (B) kurzů.	1	2	3	4	5	-1
Jsem spokojen/a s počtem nabízených povinně volitelných (B) kurzů.	1	2	3	4	5	-1

Hodnoty proměnné se dělí na tzv. **validní hodnoty** a **chybějící hodnoty** (missing values):

- **Validní hodnoty** jsou ty hodnoty, které započítáváme do analýzy. Jsou to všechny varianty odpovědí, které pro nás mají vysokou informační hodnotu.
- **Chybějící hodnoty** jsou ty hodnoty, kdy respondent zvolí odpověď typu „nevím / nemohu se rozhodnout / nemohu odpovědět" nebo otázku přeskočí a odpověď vůbec neposkytne. I tyto druhy odpovědí pro nás mohou mít informační hodnotu (např. pokud existuje na některou otázku vysoký počet odpovědí „nevím" nebo neodpovědí, měli bychom se zamyslet nad tím, zda respondenti otázce rozumí).

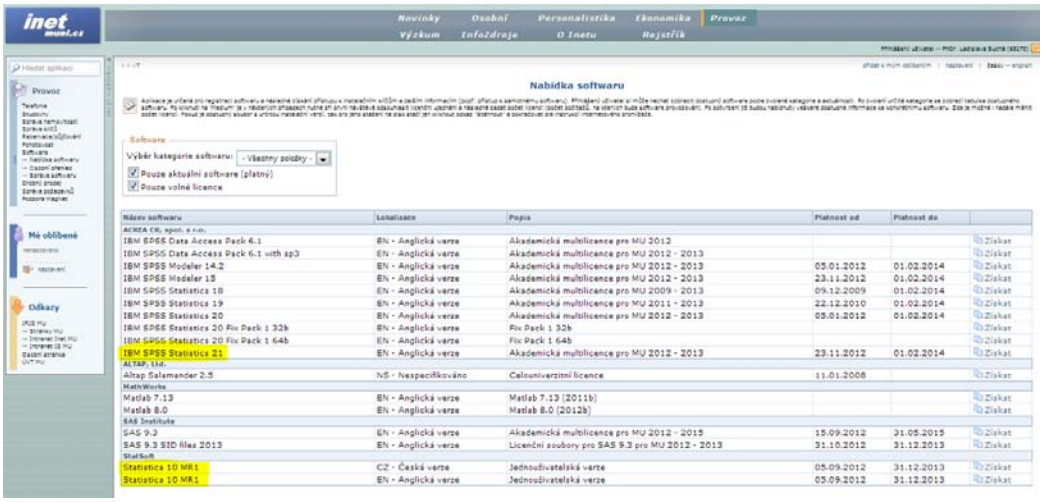
V kódování se validní hodnoty označují čísly od jedné výše, chybějícím hodnotám se dává číslice, která je na první pohled odliší (např. 99 nebo záporná číslice, např. -1).

5 Nástroje pro sběr a analýzu dat

Datové soubory lze vytvářet v různých programech.

- **Online nástroje**. Při využití online nástrojů lze data editovat často přímo v online datasetu. Téměř všechny online aplikace ale poskytují i možnost expertu dat do formátů .xls, .csv nebo .sav (formát pro SPSS).
- **Běžné tabulkové procesory**. Nejdostupnější variantou pro práci s daty jsou běžně dostupné tabulkové procesory – například MS Excel, Open Office Calc nebo Google Spreadsheets.
- **Speciální desktopové nástroje pro statistickou analýzu**. Pro statistickou analýzu existují i specializované nástroje, od free nástrojů (nejrozšířenější je pravděpodobně prostředí R) až po profesionální placené nástroje. Pro studenty FF MU jsou k dispozici zdarma programy SPSS a Statistica.

Programy SPSS a Statistica najdete v INETu. Po přihlášení se se svým UČO a sekundárním heslem najdete programy v sekci Provozní služby – Software – Nabídka softwaru.


inet
Novinky
Osobní
Personalistika
Ekonomika
Provoz
Výzkum
Infozdroje
O Inetu
Rejstřík
Nabídka softwaru
Software
Výběr kategorie softwaru:
- Všechny položky -
Pouze aktuální software (platný)
Pouze volné licence
Název softwaru
Lokalizace
Popis
Platnost od
Platnost do
ACREA CR, spol. s r.o.
IBM SPSS Data Access Pack 6.1
EN - Anglická verze
Akademická multilicence pro MU 2012
IBM SPSS Data Access Pack 6.1 with sp3
EN - Anglická verze
Akademická multilicence pro MU 2012 - 2013
IBM SPSS Modeler 14.2
EN - Anglická verze
Akademická multilicence pro MU 2012 - 2013
05.01.2012
01.02.2014
IBM SPSS Modeler 15
EN - Anglická verze
Akademická multilicence pro MU 2012 - 2013
23.11.2012
01.02.2014
IBM SPSS Statistics 18
EN - Anglická verze
Akademická multilicence pro MU 2009 - 2013
09.12.2009
01.02.2014
IBM SPSS Statistics 19
EN - Anglická verze
Akademická multilicence pro MU 2011 - 2013
22.12.2010
01.02.2014
IBM SPSS Statistics 20
EN - Anglická verze
Akademická multilicence pro MU 2012 - 2013
05.01.2012
01.02.2014
IBM SPSS Statistics 20 Fix Pack 1 32b
EN - Anglická verze
Fix Pack 1 32b
IBM SPSS Statistics 20 Fix Pack 1 64b
EN - Anglická verze
Fix Pack 1 64b
IBM SPSS Statistics 21
EN - Anglická verze
Akademická multilicence pro MU 2012 - 2013
23.11.2012
01.02.2014
ALTAP, Ltd.
Altap Salamander 2.5
NS - Nespecifikováno
Celouniverzitní licence
11.01.2008
MathWorks
Matlab 7.13
EN - Anglická verze
Matlab 7.13 (2011b)
Matlab 8.0
EN - Anglická verze
Matlab 8.0 (2012b)
SAS Institute
SAS 9.3
EN - Anglická verze
Akademická multilicence pro MU 2012 - 2015
15.09.2012
31.05.2015
SAS 9.3 SID files 2013
EN - Anglická verze
Licenční soubory pro SAS 9.3 pro MU 2012 - 2013
31.10.2012
31.12.2013
StatSoft
Statistica 10 MR1
CZ - Česká verze
Jednouživatelská verze
05.09.2012
31.12.2013
Statistica 10 MR1
EN - Anglická verze
Jednouživatelská verze
05.09.2012
31.12.2013
Získat