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?

V tomto modulu se připravíme na analýzu kvantitativních dat pomocí statistických metod.

Obsah

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

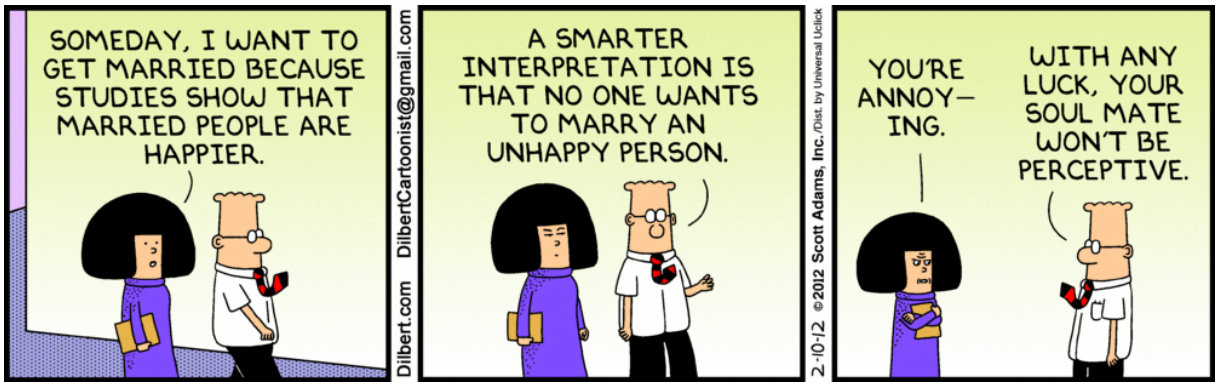
Když se začnou rozvádět nejlepší přátelé, buďte na pozoru. I vašemu vztahu hrozí vysoké riziko rozvodu, zjistili britští experti. Podle nich je rozvod nakažlivý a šíří se jako nějaká nemoc rodinami, pracovním prostředím i skupinou přátel.

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:**

- 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**

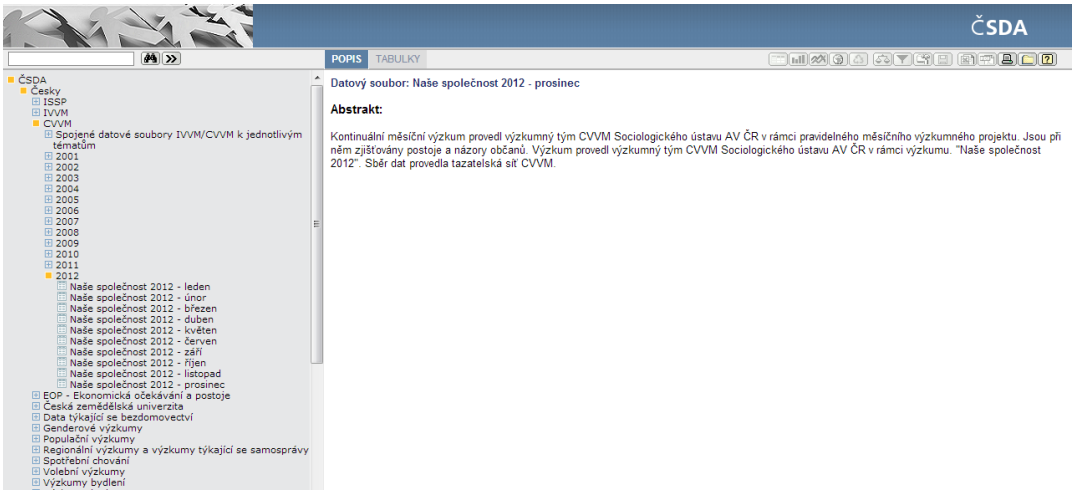
ČSÚ poskytuje informace o státní ekonomice, pohybu osob, srovnání se zahraničím, vědě a výzkumu. Kromě celé řady statistik jsou na stránkách úřadu k dispozici i otevřená data z výsledků voleb.

- **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 přístupk vybraným českým datovým souborům reprezentativních výzkumů. Bez registrace je možné procházet stránky Webu a informace o archivovaných datech.V archivu najdete například datové soubory z realizovaných měsíčních šetření Centra pro výzkum veřejného mínění (CVVM).



- **Repozitáře institucí**

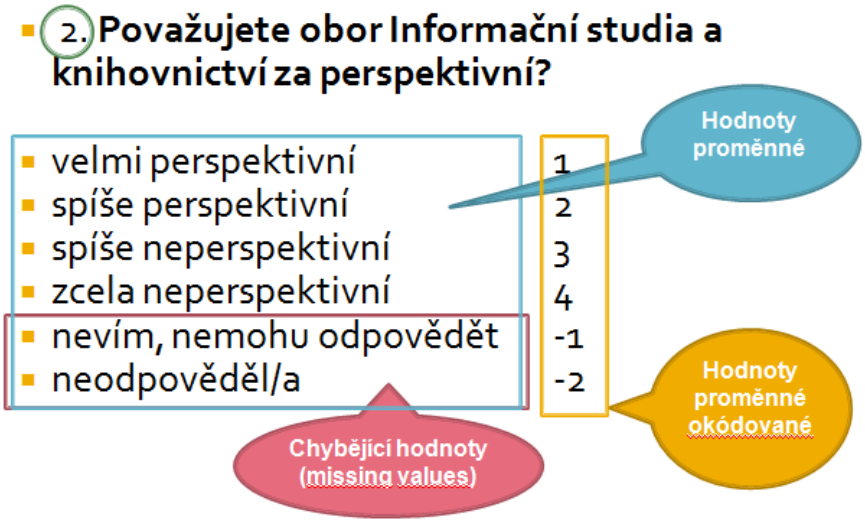
*Některé instituce se mohou rozhodnout poskytnout data ze svých průzkumů k dalším účelům. Příkladem takového rozhodnutí v našem oboru je výzkum **SOAP (Study of Open Access Publishing)** o postojích vědců k open access, který realizovali vědci z CERNu. Mezinárodní data obsahující odpovědi tisíců respondentů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

4 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.