Metodologie pro Informační studia a knihovnictví 2

Modul I: Úvod do kurzu

Co je cílem?

Na konci kurzu byste měli:

- umět navrhnout a zpracovat kvantitativní výzkum,
- umět provést základní analýzu dat (především deskriptivní analýzu) a přehledně a atraktivně tato data prezentovat,
- bez problémů zrealizovat výzkum do své bakalářské nebo magisterské práce.

Jak to bude probíhat?

Úvodní hodina bude realizována formou setkání v D21. Zbytek semestru každý pracuje individuálně. Budeme pracovat s daty, které si sami sesbíráme.

1. Na začátku bude vyhlášený výzkumný problém: výzkumné otázky a podotázky.
2. Sestrojíme si měřící nástroj – dotazník, který budeme šířit online nebo klasickou formou. Navrhneme také kvóty, které by měl splňovat náš vzorek.
3. Každý sesbírá 10 odpovědí – celkem budeme mít nakonec odpovědi od cca 700 respondentů.
4. Vyčistíme data a připravíme se na práci s nimi.
5. Budeme se postupně učit analyzovat odpovědi – od nejjednodušší frekvenční analýzy až po vytváření kontingenčních tabulek a posuzování vztahu mezi proměnnými.
6. Velkou většinu kurzu budeme pracovat s kvantitativními daty. Poslední hodiny tedy věnujeme základům práce s daty kvalitativními.
7. Výstupem bude závěrečná zpráva z výzkumu.

Software, nástroje

Data lze analyzovat s pomocí celé řady nástrojů. My budeme primárně pracovat s Excelem. Návody v kurzu budou pracovat s verzí MS Office 2010. Alternativně lze k analýze dat ale využít i jiné nástroje, například SPSS nebo free software R. Důležitý je výsledek.

Hodnocení

Hodnocení se bude skládat ze tří složek:

- průběžné práce z hodiny na hodinu,

- závěrečné zprávy z výzkumu a
- závěrečného testu (ten proběhne v PC učebně a bude mít formu praktického řešení zadaného úkolu).

Pro splnění povinností pro zápočet je nutné odevzdat **8 z 10 úkolů** a získat alespoň **30 z možných 50 bodů** (30 bodů za závěrečnou zprávu, 20 bodů za test).

Osnova kurzu

1	Úvod do kurzu • Úvod do analýzy dat • Kvantitativní a kvalitativní analýza
2	Analýza v kvantitativním výzkumu I • Datový soubor – zápis dat, kódování, úvod do práce s Excelem, s SPSS.
3	Analýza v kvantitativním výzkumu II • Importy, exporty souborů • Práce s vybranými daty • Čištění dat • Odstranění chyb při zpracování • Co s chybějícími hodnotami? • Druhy proměnných a jejich kontrola (kategorizovaná, nekategorizovaná data) • Co s extrémními hodnotami?
4	Deskriptivní analýza I • Základní popis nominálních dat (četnosti, modus) • Základní popis ordinálních dat (minimum, maximum, medián) • Tabulka četností (správná podoba: absolutní, relativní četnosti, validní četnosti, chybějící hodnoty)
5	Deskriptivní analýza II • Základní popis kardinálních dat (aritmetický průměr) • Pokročilý popis kardinálních dat (šikmost, špičatost, percentily, rozptyl, směrodatná odchylka)
6	Vytváření nových proměnných • Rekódování • Počítání • Seřazení hodnot
7	Kontingenční tabulky (třídění druhého stupně) • Správná podoba kontingenční tabulky
8	Vyhodnocování otevřených odpovědí v dotazníku
9	Induktivní statistika • Zobecňování na populaci (úskalí) • Ověřování hypotéz
10	Vizualizace dat a nástroje. • Jaké grafy používat pro jaké příležitosti? • Jaké nástroje jsou k dispozici?
11	Analýza v kvalitativním výzkumu • Kódování kvalitativních dat
12	Shrnutí obsahu kurzu, časté chyby v analýze dat.

Metodologie pro Informační studia a knihovnictví 2

Modul I: Opakování. Základní termíny. Tvorba dotazníku. Online nástroje pro sběr a administraci dat.

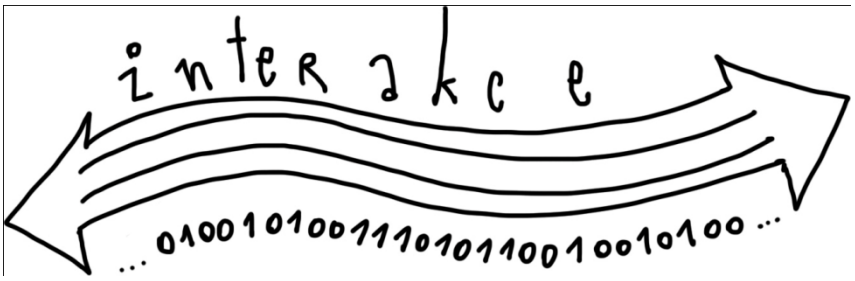
Minulý semestr jsme se věnovali návrhu dobrého výzkumu. Víme tedy, jak zrealizovat výzkum tak, aby byl validní a reprezentativní. V tomto semestru se budeme zabývat prací s daty – ukážeme si, jak co nejlépe vytěžit a zhodnotit výsledky výzkumu.

Co se dozvíte v tomto modulu?

- Co je to proměnná, hodnota proměnné, jaké jsou druhy proměnných?
- Co jsou to hypotézy a kdy se používají?
- Jaká jsou základní pravidla pro tvorbu (online) dotazníků?
- Jak vybrat vzorek pro výzkum?

Základní termíny

Při vyhodnocování kvantitativních výzkumů pracujeme s velkým množstvím **standardizovaných dat**. Proto je velmi důležité vědět, jak tato data organizovat a dále s nimi pracovat.



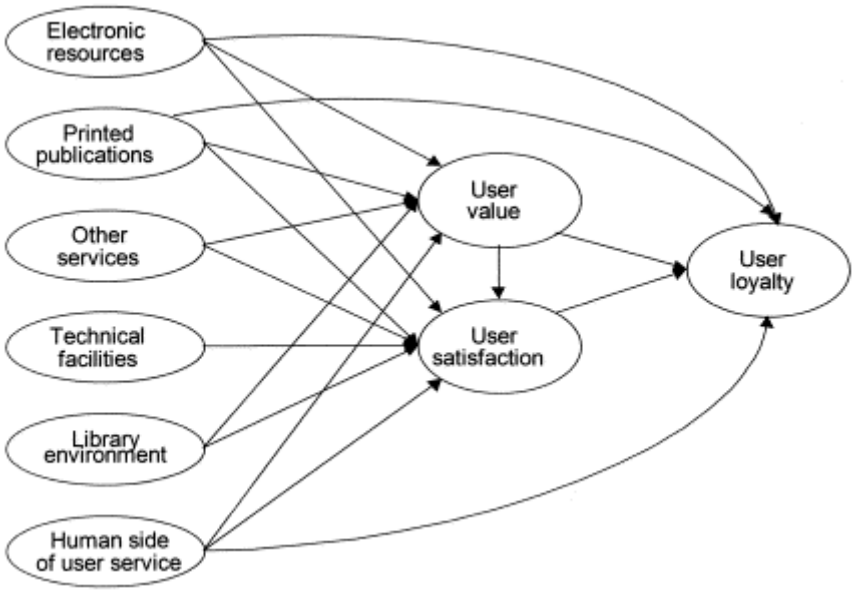
Nejprve si pojďme zopakovat základní termíny z minulého semestru:

Proměnná (znak) a její hodnoty

Při měření kvantitativních jevů nejprve vytváříme operační definici pojmů, které měříme. Tzv. **operacionalizace** znamená převedení těchto pojmů na indikátory neboli měřitelné proměnné (znaky). Např. *vzdělání* respondentů můžeme měřit jako proměnnou „dokončené vzdělání", která bude nabývat hodnot „základní", „středoškolské bez maturity", středoškolské s maturitou" atd. (Reichel 2009, s. 55).

Operacionalizace pojmu *vzdělání* respondentů je určitě jednoduchým příkladem. Existují ale pojmy, které je složitější operacionalizovat – příkladem může být např. *spokojenost uživatele s knihovnou*. Při operacionalizaci podobně složitých pojmů je dobré postupovat ve třech krocích:

a. **Vyhledat množství definic pojmu.** Například *spokojenost uživatele s knihovnou* můžeme definovat pomocí normy ISO jako „vnímání zákazníka týkající se stupně splnění jeho požadavku" (ČSN ISO 9000). Jinou definici můžeme nalézt v modelu spokojenosti a loajality uživatelů vyvinutý ve spolupráci dánských knihoven pod odbornou garancí Dánského knihovního úřadu (Styrelsen for Bibliotek og Medier).



Obr. 1: The user satisfaction and loyalty model (Martensen & Gronholdt, 2003)

b. Zvolit si nejrelevantnější definici. Hned na začátku výzkumu se musíte rozhodnout, jaká definice je pro váš výzkum nejrelevantnější. Můžete si také stanovit definici vlastní, musíte ji však vždy zdůvodnit.

c. Proveďte **dekompozici pojmu** / výzkumného problému. Výše uvedená definice Martensena a Gronholdta například říká, že model spokojenosti zahrnuje spokojenost s elektronickými zdroji, spokojenost s tištěnými publikacemi, s ostatními službami, s technickým zázemím, s prostředím v knihovně a s personálem. Každá z těchto položek jde dále ještě rozložit na nižší měřitelné jednotky. Například spokojenost s personálem můžeme ještě rozložit na tyto indikátory:
   - spokojenost s odborností personálu,
   - spokojenost s ochotou personálu,
   - spokojenost s dostupností personálu.

Podobným způsobem bychom si mohli rozložit všechny součásti pojmu *spokojenost uživatele s knihovnou*., Na konci bychom dostali určité množství indikátorů, které by tuto spokojenost měřily.

Při konstrukci výzkumného nástroje (v případě kvantitativních výzkumů typicky dotazníku, ale může se jednat např. i o tabulku pro zapisování výsledků standardizovaného pozorování) je třeba dbát na to, aby proměnné byly ***rozlišitelné*** (znak nabývá různých hodnot), ***úplné*** (jsou vyčerpány všechny hodnoty) a ***jednoznačné*** (hodnoty znaku se nepřekrývají, jsou výlučné).

Reliabilita a validita indikátorů

Indikátory jsou **realiabilní** neboli **spolehlivé**, získáváme-li opakovaně stejné výsledky. Znamená to, že všichni respondenti porozuměli otázce stejně, respondenty nemohla ovlivnit různým způsobem osoba tazatele, kódování bylo provedené stejně pro všechny případy. Spolehlivost indikátorů ale sama o sobě nezaručuje, že měření bude validní.

Indikátory jsou **validní** neboli **platné**, měří-li skutečně pojem, který jsme měřit chtěli. Pokud respondenti rozumí pojmu stejně jako my a chápou stejně měřítka, která používáme.

Druhy proměnných

Rozlišujeme tři druhy proměnných:

- **Nominální proměnné** nabývají nečíselných hodnot a nelze je uspořádat hierarchicky či podle velikosti (nemůžeme určit, která hodnota proměnné je vyšší než jiná. Speciálním případem nominálních hodnot jsou **dichotomické proměnné** (muž/žena, ano/ne). Nominální proměnnou může být např. stav, bydliště, oblíbená barva apod.
- **Ordinální proměnné** nabývají hodnot, u kterých můžeme s jistotou tvrdit, že jedna je vyšší než druhá, nemůžeme však s jistotou tvrdit, o kolik je vyšší. Ordinální proměnnou je například vzdělání, volně formulované frekvence činností.
- **Kardinální proměnné** nabývají skutečných měřitelných číselných hodnot – kardinální proměnnou je například věk, počet dětí, výše platu. Speciálním případem kardinálních proměnných jsou intervalové proměnné (např. výše platu měřená intervaly 0-10000, 10001-20000, 20001-30000…).

Ovlivňující proměnná se nazývá **nezávisle proměnná,** ovlivňovaná proměnná se nazývá **proměnná závislá**.

Hypotézy

Hypotézy jsou výroky **o dosud neprokázaných vztazích mezi dvěma nebo více proměnnými**. Jsou formulovány stručně a jasně, ideálně ve formě jedné oznamovací věty. Hypotézy by měly být vždy formulované tak, aby byly ověřitelné, a všechny proměnné musí být definovány operacionálně (jinými slovy tak, že je jasné, jak budou měřeny).

Při formulaci hypotéz postupujeme tak, že si nejprve stanovíme tzv. **vstupní hypotézu**, kterou si zpřesníme na několik **pracovních hypotéz** tak, že provedeme dekompozici výzkumného problému (rozložíme si tedy zkoumané koncepty na co nejmenší zkoumatelné dílky). Pracovní hypotézy lze postupně zpřesnit a formulovat i jako hypotézy statistické.

Příklad vstupní a pracovní hypotézy:

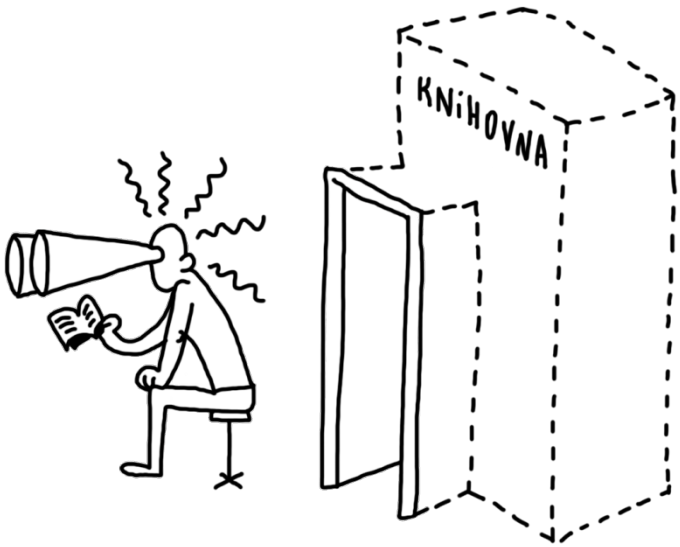
- H (vstupní hypotéza): Lidé chodí do knihovny z mnoha důvodů, které se liší podle jejich sociodemografických, vzdělanostních a dalších charakteristik.
- H1 (pracovní hypotéza 1): Důvody chození do knihovny se liší podle pohlaví.
- H2 (pracovní hypotéza 2): Důvody chození do knihovny se liší dle věku.
- H3 (pracovní hypotéza 3): Důvody chození do knihovny se liší dle dosaženého vzdělání.
- …

V tomto příkladu jsme provedli dekompozici výzkumného problému a máme nyní čtyři proměnné (závislou proměnnou „důvody chození do knihovny" a tři nezávislé proměnné: pohlaví, věk a dosažené vzdělání).

Deskriptivní a induktivní statistika

Formulace hypotéz je důležitá zejména ve výzkumech, kdy používáme ***induktivní statistiku***. Ve výzkumech, které počítají jen se ***statistikou deskriptivní***, si vystačíme s výzkumnými otázkami.

Deskriptivní statistika	Induktivní statistika
• Jednoduchý popis a sumarizace souborů • Dává odpověď na otázku po prostých četnostech hodnot jednotlivých proměnných (třídění prvního stupně), případně po četnostech rozlišených dle kategorií (třídění druhého a dalších stupňů) • Na základě prosté deskriptivní analýzy nelze zobecňovat na populaci	• Způsob přenášení závěrů analýzy z výzkumného vzorku na celou populaci. • Je založená na matematických statistických modelech a teoriích pravděpodobnosti. • Pomocí induktivní statistiky lze zobecnit na populaci s určitým stupněm spolehlivosti.



Populace a vzorek

Začneme opět opakováním z předchozího semestru. V **kvantitativním výzkumu** nepracujeme s celou populací, o které chceme vypovídat. Pracujeme se vzorkem, který by měl populaci co nejlépe reprezentovat. Připomeňme si základní pojmy:

Populace

Populace je množina jednotek, které chceme koumat. Předpokládáme, že závěry našich výzkumů jsou platné pro celou populaci.

Vzorek

Vzorek je množina jednotek reprezentujících populaci. Jsou to jednotky, které skutečně pozorujeme či měříme.

V kvantitativním výzkumu je **plán výzkumu** včetně výběru vzorku **známý předem**. Klíčová je přitom znalost populace – některé typy reprezentativních výběrů pracují se seznamem jednotek v populaci, jiné vycházejí ze znalosti charakteristik populace.

Pro určení **velikosti reprezentativního vzorku** se často používá tato tabulka:

Velikost populace	Velikost vzorku (pravděpodobnostní výběry)
Do 100 jednotek	80 %
Do 1000 jednotek	40 %
Do 10 000 jednotek	7,5 %
Do 100 000 jednotek	1,5 %
Do 1 000 000 jednotek	0,25 %
Do 10 000 000 jednotek	0,045 %

Jedná se však o velmi zjednodušený přehled. Velikost a složení výběru vzorku závisí na řadě faktorů, především na:

- míře **homogenity** populace,
- **členitosti** zkoumaných znaků,
- používání dalších **stupňů třídění**,
- zamýšlené míře **statistické pravděpodobnosti** výpovědí.

Jednoduše řečeno, čím složitější bude náš výzkumný záměr, čím více znaků budeme sledovat, čím podrobnější třídění budeme používat (například nebudou nás zajímat rozdíly ve sledované vlastnosti jen podle pohlaví, ale i podle pohlaví a věku a dosaženého vzdělání), čím heterogennější bude naše populace ve sledovaných znacích, tím větší potřebujeme výzkumný vzorek.

Typy výběrů

Existuje celá řada typů výběrů, které se uplatňují v kvantitativním výzkumu.

Reprezentativní výběry		Nereprezentativní výběry
Pravděpodobnostní výběry	**Nepravděpodobnostní výběry**	
Prostý náhodný výběr	Kvótní výběr	Snowball technika
Systematický výběr		Teoretický výběr

Náhodný stratifikovaný výběr	Výběr typických případů
Náhodný skupinkový výběr	Výběr kritických případů
	Účelový výběr

Podrobně jsou druhy výběrů popsány na **wiki.knihovna.cz**.

Tvorba dotazníku

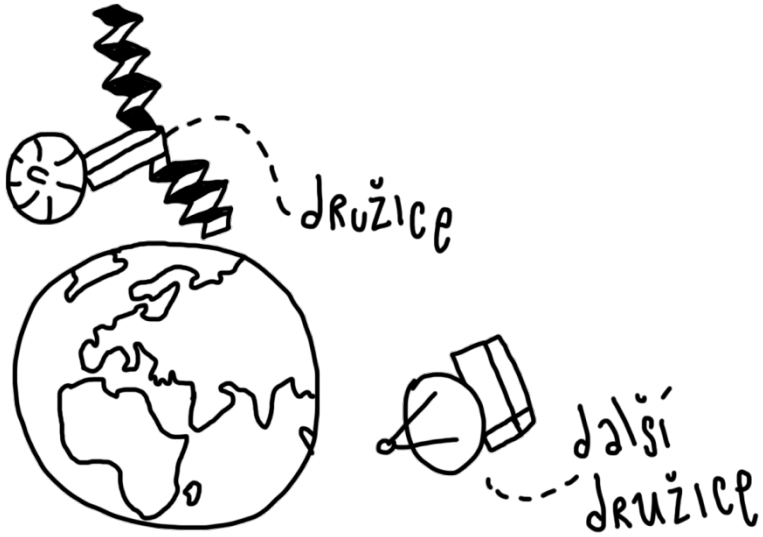
Velkou většinu semestru budeme trávit analýzou dat, která získáme standardizovaným dotazníkem. O tvorbě dotazníků existuje celá řada zdrojů, proto si pojďme zopakovat jen to nejdůležitější.

Než začneme vytvářet dotazník…

Před samotným vytvořením dotazníku bychom si měli být jistí odpovědí na následující otázky:

- Máme dobře **rozpracovaný výzkumný problém**? Provedli jsme dobře dekompozici a máme tedy pojmenované a definované všechny proměnné, které potřebujeme změřit?
- Budeme přebírat **existující dotazník** nebo vytvoříme **dotazník nový**? Neexistují už ověřené měřící nástroje, které měří to, co chceme?
- **Jak budeme dotazník distribuovat**? Stačí nám papírová forma nebo budeme vytvářet online dotazníkový formulář? Jaké požadavky na dotazník máme?

O rozpracování výzkumného problému už byla řeč, pojďme se podívat na vytváření a distribuci dotazníku…



Existující nebo nový dotazník?

Vytvářet nový dotazník není jednoduchá činnost – je zapotřebí hlídat reliabilitu i validitu výzkumného nástroje, a to vyžaduje celou řadu testů. Existuje celá řada **standardizovaných** dotazníků i sad otázek, které můžeme využít.

Využití standardizovaného dotazníku má navíc velkou výhodu – můžeme naše data porovnat snadno se závěry jiných výzkumníků.

Výhody využití existujících a vytváření nových dotazníků

Využití existujících nástrojů	Vytvoření nového dotazníku
• Používáme již otestovaný nástroj (validita, reliabilita). • Máme možnost porovnat naše výsledky s výsledky předchozích výzkumů. • Pokud nejsme zkušení výzkumníci, je převzetí dotazníku sázkou na jistotu.	• Nadesignování dotazníku přímo pro náš výzkumný problém. • Možnost přizpůsobení lokálním podmínkám.

V praxi lze samozřejmě oba přístupy kombinovat – například využít sady otázek z existujících dotazníků a další otázky k nim přidat. Na tomto principu fungují například dotazníky zjišťující spokojenost se službami knihoven (http://vyzkumy.knihovna.cz/dotazniky).

V našem oboru je příkladem standardizovaného dotazníku například **dotazník na měření fenoménu úzkosti z knihoven** od Sharon Bostick. Dotazník byl přeložen do českého jazyka v rámci bakalářské práce Moniky Machovské.

Online nástroje pro sběr a administraci dat

Existuje celá řada online nástrojů pro sběr a administraci dat. Liší se cenou (ta se pohybuje od licencí zdarma až po tisícové položky za měsíc), funkcemi, dostupnými variantami otázek, ale i možnostmi distribuce a designu dotazníků. Než se rozhodnete, v jaké aplikaci budete dotazník vytvářet, je tedy potřeba rozmyslet všechny požadavky, které by měl nástroj splňovat (některé nástroje neumí speciální typy otázek, některé neumožňují grafickou customizaci šablon).

Příklady služeb pro tvorbu dotazníků:

- Google Spreadsheets
- Survey Monkey
- SurveyGizmo
- Polldaddy
- Survs
- Survio (české)
- Easyresearch (český)

Služby Survio a Easyresearch mohou studenti KISKu využívat v rámci svého studia zdarma.

Voucher na GOLD licenci služby Survio pro studenty KISKu je
survio-kisk-gold-2012-JsnxUh3s.

Specifika online výzkumů[1]

Většina studentských výzkumů probíhá dnes formou online dotazníků. Pro online dotazování se často používá zkratka CAWI (Computer Assisted Web Interviewing). Online dotazníky umožňují získat velké množství dat rychleji, efektivněji a často i pro respondenty zajímavou formou. Je-li k online dotazování využito správného nástroje, poskytuje tato forma dotazování **řadu výhod**, především:

- výrazně snižuje časový interval sběru dat,
- eliminuje chyby při sběru dat (filtrování otázek a implementace logických posloupností se odehrává bez toho, aby o tom respondent věděl),
- poskytuje možnost okamžité a průběžné kontroly výsledků a jejich základní analýzy,
- obsahuje možnost zapojení multimédií (videa, obrazu, zvuku…),
- eliminuje finanční náklady na tisk dotazníků a tazatele atd.,
- umožňuje i vyplnění poměrně dlouhého dotazníku (např. delšího než při telefonickém dotazování), aniž by respondent formulář opustil,
- získaná data lze snadno exportovat do různých formátů (.csv, .xls, .sav atd.),
- distribuce dotazníku může probíhat současně několika kanály (zasláním html odkazu, zveřejněním na stránkách atd.),
- snadnou integraci dotazníku do webových stránek knihovny formou iframe nebo pop-up oken - pro umístění stačí jen jednoduché vložení kódu (embed).

Metoda CAWI má však i řadu **slabých stránek**, se kterými je potřeba počítat:

- výzkumná skupina, kterou jsme schopni oslovit, nemusí být reprezentativní vzhledem k celkové populaci registrovaných čtenářů knihovny – podle údajů Eurostatu (2009) má pouze 54 % českých domácností přístup k internetu, přičemž obrovský rozdíl v používání internetu existuje mezi různými věkovými skupinami (zatímco ve věkové skupině 12-19 let využíván internet 90 % populace, mezi seniory nad 60 let pouze 14 %),
- obtížná je možnost kontroly identity respondenta,
- míra návratnosti závisí na formě distribuce dotazníku,
- snižuje možnost pokládat komplexní a náročné otázky.

Design a implementace online dotazníků tedy vyžaduje velký důraz na kontrolu rizika nízké návratnosti, reprezentativnosti a výběrové chyby (Vicente, Reis, 2010). Na řadu úskalí online dotazování upozorňuje také Ryšavý (Ryšavý 2011), který připomíná, že online dotazování bývá většinou **cenzovní šetření s nízkou návratností** a s výsledky nelze zacházet jako s reprezentativními daty získanými pomocí náhodného výběru. Nemůžeme v nich pracovat se statistickou významností tak, jako v reprezentativních šetřeních. Ryšavý však tyto druhy výzkumů nezavrhuje – nízká návratnost nemusí automaticky znamenat chybné výsledky: „ Jde spíše o to, jak zhodnotit informace o nerespondentech, které mají výzkumníci k dispozici, a jak tuto zprávu zprostředkovat nejen odbornému publiku.“ (Ryšavý 2011, s. 100).

[1] Vychází z článku Suchá, L. Měření spokojenosti online: Sběr a administrace online dotazníků. Knihovny současnosti 2012 - sborník. Ostrava: Sdružení knihoven ČR, 2010.

6 tipů na zvýšení návratnosti online dotazníků

Faktorům ovlivňujícím návratnost online dotazníků byla v posledních letech věnována řada výzkumů. Vicente a Reis (2010) stanovili šest oblastí, které míru návratnosti online dotazníků ovlivňují:

1. ***Obecná struktura dotazníku***. Předchozí výzkumy ukázaly, že existují rozdíly mezi návratností online dotazníků, které jsou rozděleny na několik stran a delších dotazníků, kde musí respondent rolovat stránku. Rozdělení dotazníku do několika stran bez nutnosti rolování zvyšuje míru vyplněných a dokončených dotazníků (Lozar Manfreda, Batagejl a Vehovar, 2002).
2. ***Délka dotazníku***. Délka dotazníku má vliv především na míru opuštění online formuláře v průběhu jeho vyplňování, dále také na míru výběru odpovědí typu „nevím, nemohu odpovědět", která stoupá spolu s délkou dotazníku (Deutskens et al, 2004).
3. ***Sledování pokroku ve vyplňování***. Součástí online dotazníků by mělo být i sledování pokroku ve vyplňování (formou grafického znázornění nebo zprávy Zbývá vyplnit XX % otázek / stran dotazníku atd.). Důležitější než samotný pokrok je ale vnímání tohoto pokroku samotným respondentem (nevidí-li hned zpočátku jasný pokrok ve vyplňování, míra opuštění dotazníku se zvyšuje).
4. ***Vizuální prezentace***. Množství výzkumů sledovalo i roli vizuální prezentace. Vyšší míra vyplnění online dotazníků je zaznamenána u designově střídmých dotazníků. Oproti tomu vizualizace určitých prvků (např. log, ukázek zboží) může zvyšovat míru odpovědí na konkrétní otázku (Deutskens, et al. 2004).
5. ***Interaktivita***. Interaktivní dotazník reaguje na respondenta – využívá filtrovacích otázek, změny řazení odpovědí atd. Interaktivita dotazníku ovlivňuje především kvalitu získaných dat.
6. ***Formát otázek a odpovědí***. Složité otázky a otevřené otázky zvyšují dle výzkumů míru opuštění dotazníku před jeho kompletním vyplněním (Lozar Manfreda, Vehovar, 2002). Významný rozdíl oproti tomu nebyl nalezen např. mezi odpověďmi formou tlačítek (radio buttons) a výběrem odpovědí (drop down boxes).

Formulace otázek v dotazníku

Ještě důležitější než zvýšení návratnosti je však mít **metodologicky správně postavený dotazník**, s **dobře formulovanými otázkami**. K formulaci otázek můžeme přistoupit tehdy, máme-li dobře provedenou operacionalizaci a jsou nám tedy jasné všechny proměnné, na které se chceme dotazovat.

Existuje celá řada požadavků pro formulaci otázek v dotazníku (např Jarrett & Gaffney 2009). Mezi nejdůležitější patří tyto zásady:

- ✓ Používejte **jednoduchý jazyk**, kterému respondenti rozumějí. Potřebujete se zeptat, zda respondentům vyhovuje v knihovně mezinárodní desetinné třídění? Zda využívají hledání podle signatur? Zda by využili QR kódy? Zeptejte se tak, aby tomu rozuměli!
- ✓ Ptejte se tak **jednoznačně**, aby všichni rozuměli otázce stejně. Zeptáte-li se například *„Jak často čtete?"* (Někteří respondenti mohou odpovídat na to, jak často čtou knihy, jiní mohou do odpovědi zařadit i noviny a časopisy atd.).

- ✓ Ptejte se na **reálné zážitky či chování**. Vyvarujte se hypotetických otázek, na které nemohou respondenti odpovědět na základě vlastní zkušenosti (předvídat vlastní budoucí chování je vždy problematické).
- ✓ Ptejte se vždy **pouze na jednu věc**. Otázky, které ve skutečnosti obsahují dvě podotázky, se nazývají odborně ***dvouhlavňové***. Jedná se o otázky typu „Jak často čtete knihy a časopisy? (respondent může číst každý den časopisy, ale knihy jen výjimečně, přesto bude mít na tuto otázku stejné skóre jako vášnivý čtenář).
- ✓ Nepokládejte **návodné otázky**. Příkladem návodné otázky může být například sugestivně položený dotaz *„Všeobecně se ví, že čtení pomáhá dětem zlepšovat vztah ke světu i k sobě samým. Čtete se svými dětmi?"*. Speciálním případem návodných otázek může být používání autorit (např. *„Souhlasíte s názorem emeritního profesora XY, že…?"*
- ✓ Dejte si pozor na **používání záporů**. Například otázka *„Souhlasíte s tím, že by se erotická beletrie neměla půjčovat náctiletým?"* nemusí být pro všechny jednoduše pochopitelná.
- ✓ Používejte pojmy, které mají pro každého **stejný význam**. Například časová určení „často", „málokdy" atd. může každý vnímat jinak.
- ✓ **Seskupujte otázky** podle smyslu tak, aby jim respondent velmi snadno rozuměl a neztratil se v kontextu.
- ✓ Používejte **jasné časové a místní rámce**. Chcete-li se zeptat například na frekvenci čtení, připojte upřesnění časového období (například „Jak často jste v uplynulém roce četl beletrii?"
- ✓ Dejte si pozor na **příčinnou souvislost**. Příkladem otázky, která podsouvá příčinnou souvislost je třeba *„Zařídil/a jste si kvůli rostoucí ceně knih průkazku do knihovny?"*.
- ✓ Podobně si dejte pozor i na **neodůvodněné předpoklady**. Špatně položená otázka je i *„S ohledem na stav našeho školství, myslíte si, že je dobré poslat své dítě do státní školy?"*
- ✓ Vyčerpejte **všechny možnosti**. Nabízíte respondentům všechny možnosti odpovědi? Pokud si nejste jistí, připojte odpověď *„nevím"* a *„jiné*, prosím upřesněte".
- ✓ **Zbavte se zbytečných otázek!** Projděte si dotazník ještě jednou a dejte si pozor, zda nemáte zbytečně v dotazníku otázky, které se opakují, či zbytečně zjišťují stejnou věc (pokud neslouží ke kontrole kvality odpovědí).

Dramaturgie dotazníku

Stejně důležitá jako dobrá formulace otázek je i **dramaturgie dotazníku**. Na začátek dotazníku je vhodné umístit jednodušší otázky, které uvádějí do tématu. Složitější otázky a obsáhlejší baterie je vhodné umístit doprostřed dotazníku, naopak otázky citlivější a otázky na sociodemografické údaje je vhodné umístit na konec dotazníku.

U složitějších dotazníků je vhodné využívat **filtrační otázky**, které filtrují nerelevantní otázky. Například na názor na práci knihovního personálu na výpůjčním pultu se budeme ptát pouze respondentů, kteří dříve uvedli, že využívají výpůjční knihovní služby.

U složitějších dotazníků je také vhodné využívat přechodů, vysvětlení – **provést respondenta dotazníkem**. Pokud je to možné, dělte otázky tematicky do oddílů a sdělte respondentům, co je v jednotlivých oddílech čeká.

Průvodní dopis

Při tvorbě dotazníku nezapomínejte na komunikaci s respondentem, především na oslovení. To by mělo obsahovat:

- **oslovení** (personalizované oslovení znamená mnohem vyšší míru návratnosti),
- **zdůvodnění výzkumu** (pokud se vám podaří přesvědčit respondenta, že je výzkum důležitý a potřebný, zafunguje to dokonce lépe než dárky a odměny za vyplnění),
- **prohlášení o anonymitě**,
- informaci o tom, **kdo výzkum dělá** a **jak bude naloženo s daty**.

Instrukce pro vyplňování

Především v případě složitých dotazníků dbejte na pečlivé instrukce. Ne všechny online aplikace umožňují vkládat text mezi otázky (umí to například Survs.com). Pokud potřebujete respondenta provést dotazníkem, ověřte si, zda to aplikace, kterou chcete použít, umí.

Jarrett a Gaffney (2009) o psaní instrukcí ve webových formulářích a dotaznících říkají:

a. **Pište instrukce jednoduchým jazykem**: využívejte známé termíny, aktivní věty (oslovujte přímo respondenta, nepožívejte trpný rod), nepoužívejte dlouhá souvětí a nekonečné odstavce, místo nich volte raději odrážky – rozbijte graficky text, používejte smysluplné nadpisy sekcí dotazníku.
b. **Seškrtejte instrukce, které jsou zbytečné**: zbavte se textu, který není nutný.
c. Vložte instrukce opravdu **tam, kde jsou potřebné**.

A závěrečná rada…

Svůj dotazník vždy nejprve otestujte! Testování vám pomůže odhalit chyby, kterých jste si nemuseli při tvorbě otázek všimnout. Pro testování vám stačí několik málo respondentů, ideálně přímo z vaší výzkumné populace. Pokud to z nějakého důvodu možné, otestujte dotazník alespoň na svých spolužácích, rodině, přátelích…

Literatura

Connaway, L. S., & Powell, R. R. (2004). Basic research methods for librarians. Library and information science text series. Westport, Conn: Libraries Unlimited.

Denscombe, M. (2006). Web-based questionnaires and the mode effect: An evaluation based on completion rates and data contents of near-identical questionnaires delivered in different modes.*Social Science Computer Review*, 24(2) 246-254.

Deutskens E., de Ruyter K., Wetzels M., Oosterveld P. (2004). Response rate and response quality of internet-based surveys: an experimental study. *Mark. Lett*. 15(1), 21–36.

Eurostat. (2009) Internet usage in 2009 - Households and Individuals. WWW: http://epp.eurostat.ec.europa.eu/cache/ITY_OFFPUB/KS-QA-09-046/EN/KS-QA-09-046-EN.PDF.

Jarrett, C., & Gaffney, G. (2009). *Forms that work: Designing Web forms for usability.* Amsterdam: Morgan Kaufmann.

Lozar Manfreda, K., Batagelj, Z. & Vehovar, V. (2002): Design of Web Survey Questionnaires: Three Basic Experiments. *Journal of Computer-Mediated Communication*, 7 (3).

Martensen, A., Gronholdt, L. (2003). Improving library users' perceived quality, satisfaction and loyalty: an integrated measurement and management system, *Journal of Academic librarianship*, 2003, 29(3), pp.140-147.

Reichel, J. Kapitoly metodologie sociálních výzkumů. Praha: Grada, 2009.

Ryšavý, D. (2011). Úskalí on-line dotazování při měření postojů vysokoškoláků a pracovníků vysokých škol. Data a výzkum - SDA Info, 2011, roč. 5, č. 1, s. 85-103.

Vicente, P., Reis E., (2010). Using Questionnaire Design to Fight Nonresponse Bias in Web Surveys. *Social Science Computer Review*, May 1, 2010; 28(2): 251 - 267.

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

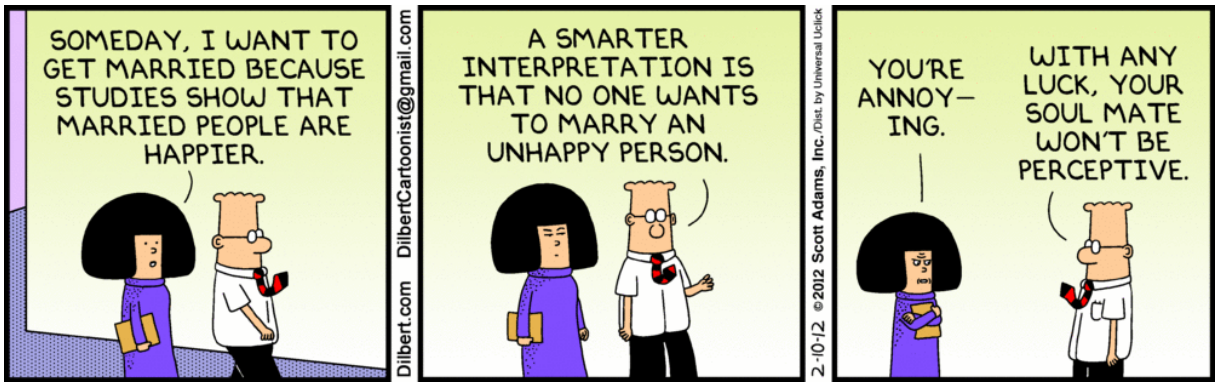
Když se začnou rozvádět nejlepší přátelé, buďte na pozoru. I vašemu vztahu hrozí vysoké riziko rozvodu, zjistili britští experti. Podle nich je rozvod nakažlivý a šíří se jako nějaká nemoc rodinami, pracovním prostředím i skupinou přátel.

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**

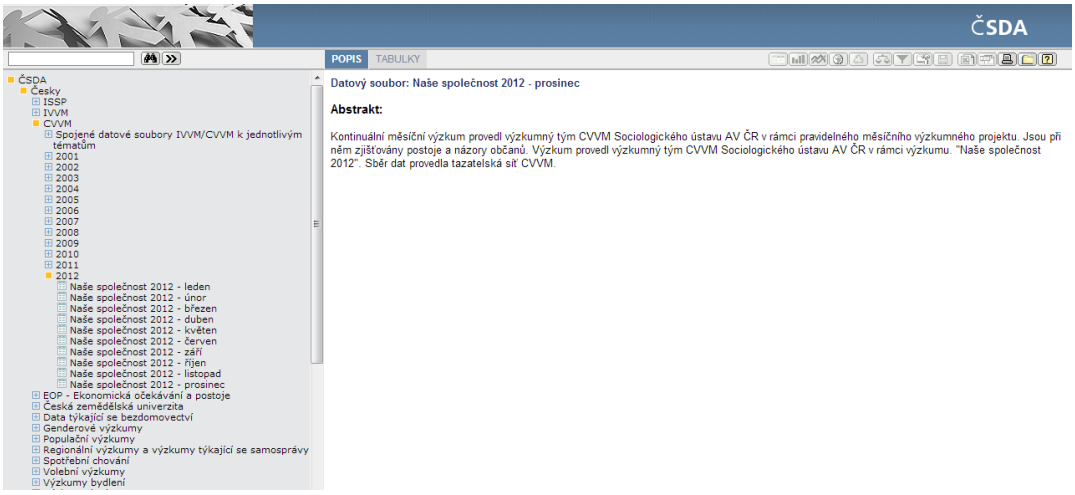
ČSÚ poskytuje informace o státní ekonomice, pohybu osob, srovnání se zahraničím, vědě a výzkumu. Kromě celé řady statistik jsou na stránkách úřadu k dispozici i otevřená data z výsledků voleb.

- **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**

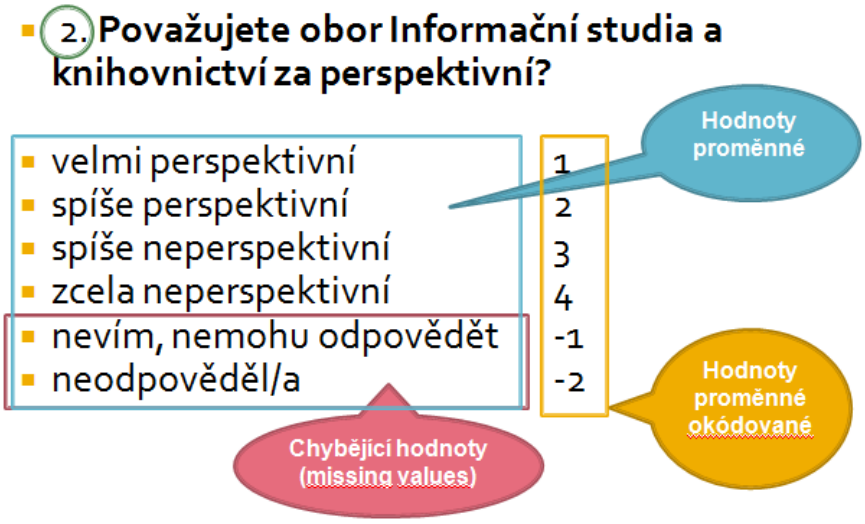
*Některé instituce se mohou rozhodnout poskytnout data ze svých průzkumů k dalším účelům. Příkladem takového rozhodnutí v našem oboru je výzkum **SOAP (Study of Open Access Publishing)** o postojích vědců k open access, který realizovali vědci z CERNu. Mezinárodní data obsahující odpovědi tisíců respondentů 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).

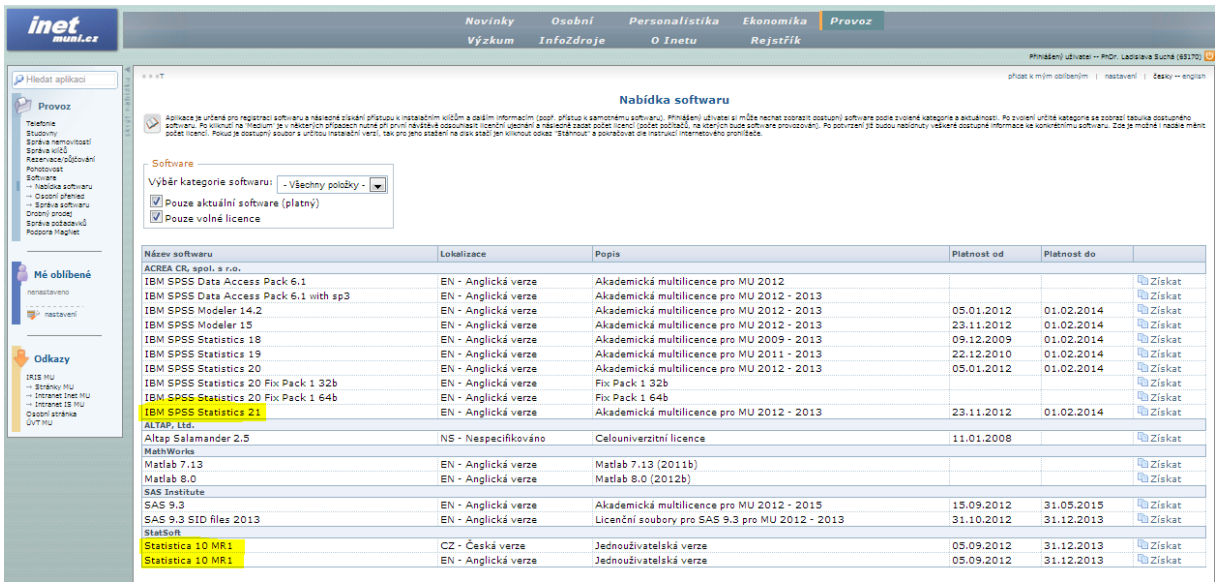
V kódování se validní hodnoty označují čísly od jedné výše, chybějícím hodnotám se dává číslice, která je na první pohled odliší (např. 99 nebo záporná číslice, např. -1).

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.



Metodologie pro Informační studia a knihovnictví 2

Modul 3: GIGO. Popis dat. Kontrola a čištění dat.

Co se dozvíte v tomto modulu?

- Proč je potřeba dbát na kvalitu dat na vstupu
- Jak popsat výběrový soubor a na jaké hodnoty proměnných dávat pozor při kontrole?
- Jak vybrat jen určité případy (nový dataset)
- Jak postupovat v Excelu a v Google Spreadsheets?

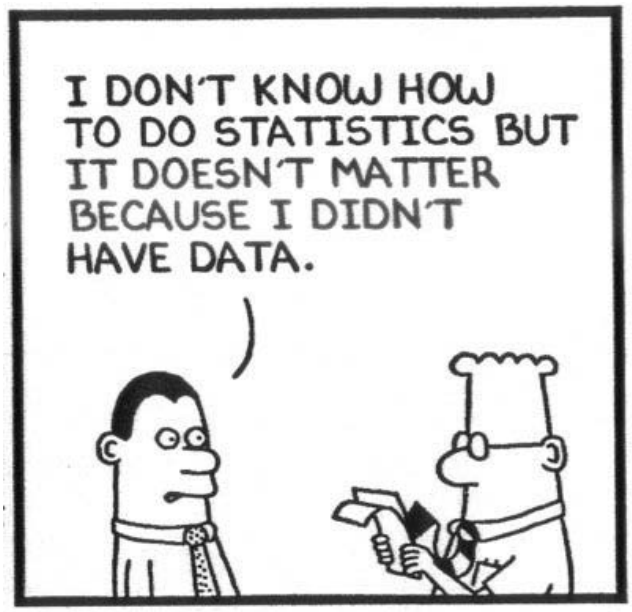
V tomto modulu si připravíme dataset k samotné analýze. To, zda budete mít na konci analýzy smysluplné výsledky, do značné míry záleží právě na tom, jakou míru pozornosti budete věnovat počáteční kontrole dat.

Obsah

1 Pravidlo 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 pouze narůstá a čím dál více 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 nejsou tak zajímavé. *„Pokud si vezmete 100 statistických šetření, je pravděpodobné, že jedno z nich bude mít naprosto 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. 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 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?***

2 Popis a kontrola dat

Prvním úkolem výzkumníka je popis výběrového souboru. Charakteristikou vzorku by měla začít každá analýza i analytická kapitola v bakalářské či diplomové práci. Zajímá nás například:

- Kolik je ve výběrovém souboru jednotek?
- Kolik je v souboru mužů a žen?
- Kolik je v souboru lidí se ZŠ/SŠ/VŠ vzděláním?
- Jak je v souboru distribuován věk?

Toto rozložení může být vyjádřeno v **absolutních**, **relativních**, či **kumulativních relativních četnostech**.

- **Absolutní četnost** udává absolutní číslo – hodnotu četnosti varianty proměnné v souboru.
  Například: V souboru je 1456 mužů a 1201 žen.
- **Relativní četnost** udává **podíl** četnosti varianty proměnné v souboru.
  Například: V souboru je 24 % osob se základním vzděláním.
- **Kumulativní relativní četnost** udává kumulativní podíly variant proměnné v souboru (nejsou použitelné pro nominální proměnné).
  Například: V souboru je 36 % respondentů, kteří mají alespoň maturitu (tedy nejen úspěšní středoškoláci s maturitou, ale také vysokoškoláci se všemi variantami diplomů).

Popis a kontrola kategorizovaných dat

Tabulky četností

Pro zobrazení základních hodnot popisu rozložení hodnot kategorizovaných proměnných (tedy proměnných nominálních a ordinálních s menším počtem variant odpovědí) se používá tzv. **tabulka četností**. Ta obsahuje jak absolutní, tak relativní četnosti hodnot proměnných. Takto vypadá správná a kompletní tabulka četností:

Jaké je Vaše vzdělání?				
		Četnost odpovědí	**Relativní četnost**	**Validní relativní četnost**
Validní hodnoty	Základní	46	7,5 %	7,6 %
	Základní vyučen /střední bez maturity	62	10,1 %	10,2 %
	Střední s maturitou	307	50,1 %	50,5 %
	Pomaturitní nástavba, VOŠ	40	6,5 %	6,6 %
	Vysokoškolské	153	25,0 %	25,2 %
	Celkem validní hodnoty	608	99,2 %	100,0 %
Chybějící hodnoty (neví, neodpověděl/a)	Chybějící hodnoty	5	0,8 %	
Celkem		**613**	**100,0 %**	

V praxi se často používá jen zkrácená verze tabulky obsahující pouze validní četnosti:

Jaké je Vaše vzdělání?	**Četnost odpovědí**	**Validní relativní četnost**
Základní	46	7,6 %
Základní vyučen /střední bez maturity	62	10,2 %
Střední s maturitou	307	50,5 %
Pomaturitní nástavba, VOŠ	40	6,6 %
Vysokoškolské	153	25,2 %
Celkem	**608**	**100,0 %**

Před počítáním četností je ale potřeba zkontrolovat data. Kontrolujeme, zda se nachází v platném intervalu (například proměnná pohlaví nabývá v našem souboru pouze hodnot 1 a 2, všechny ostatní varianty by měly být omyly).

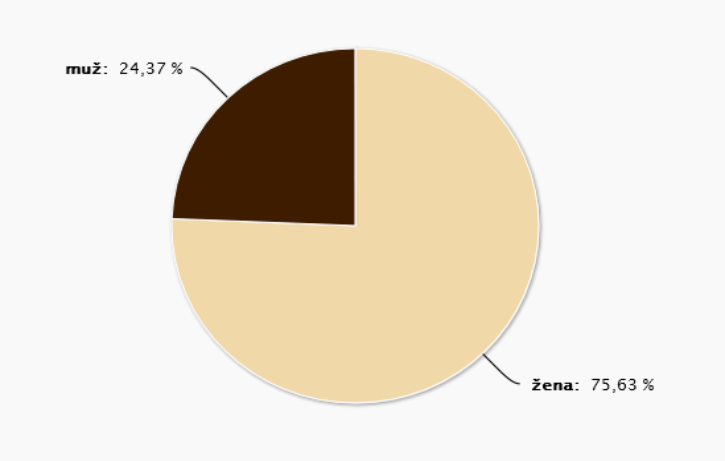
Grafy četností

Pro znázornění rozložení četností se využívají i grafy znázorňující četnosti hodnot proměnných. Nejznámějšími variantami jsou koláčový a sloupcový graf.

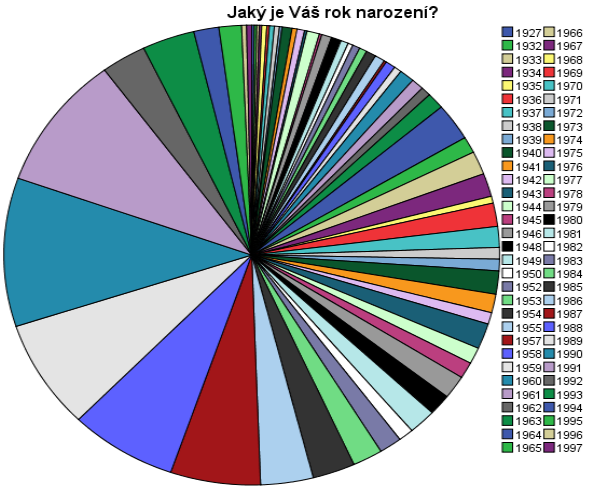
Koláčový graf je vhodný:

- pro třídění prvního stupně (jedna datová řada),
- pro porovnání četností u nominálních proměnných, které nemají příliš mnoho hodnot (méně než 7),
- pokud hodnoty, které chcete vykreslit, nejsou nulové,
- pokud hodnoty představují část celku.

Příklad proměnné, kde je vhodné využít koláčový graf:



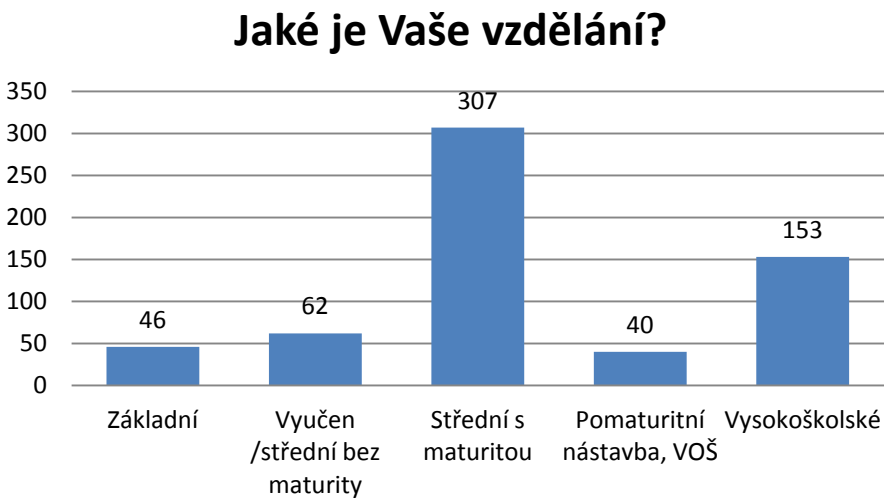
Příklad proměnné, kde NENÍ vhodné využít koláčový graf:



Sloupcový graf je vhodný pro:

- porovnání položek,
- ordinální proměnné a kardinální proměnné s menším počtem kategorií,
- znázornění změn za časové období (třídění druhého stupně).

Příklad sloupcového grafu:



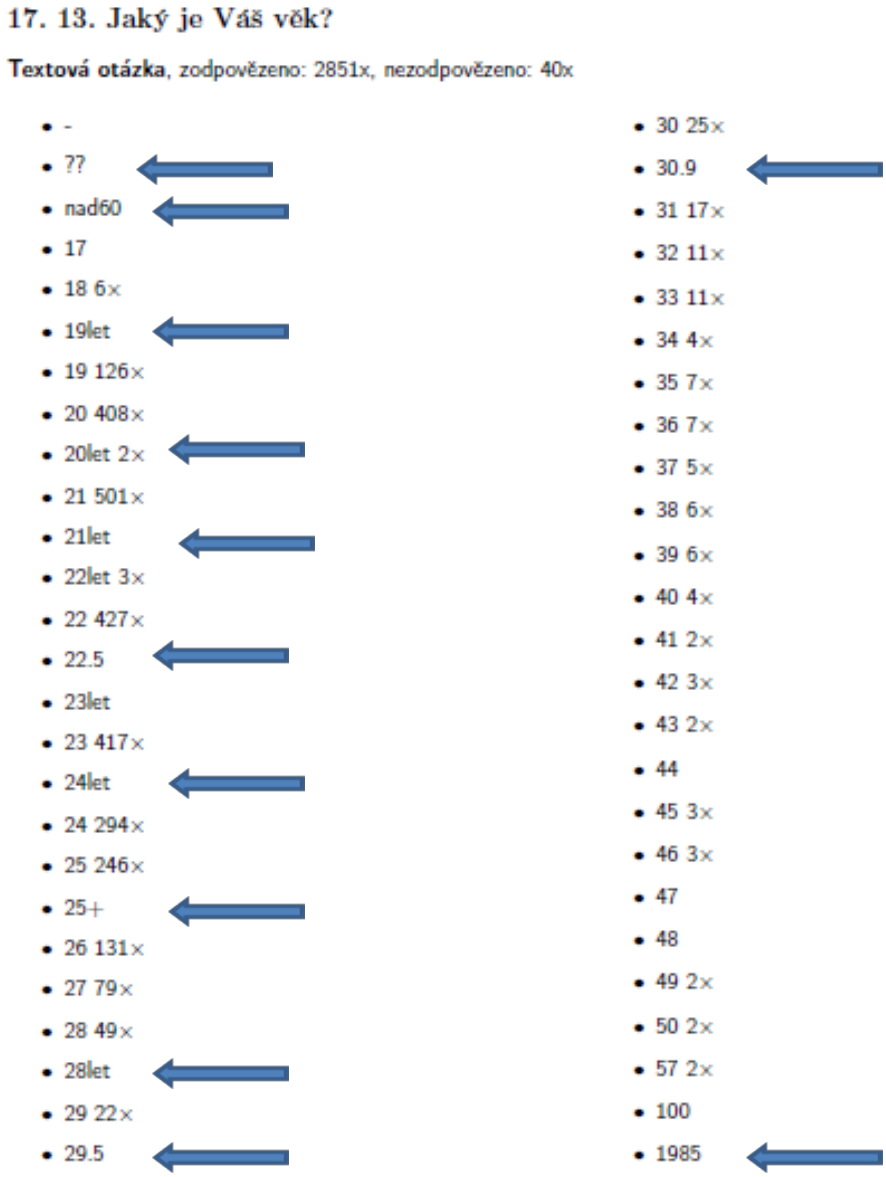
Grafy se v Excelu vkládají pomocí funkce „**Grafy**“ na listu „**Vložení**“.

Popis a kontrola nekategorizovaných dat

Pro první kontrolu nekategorizovaných dat nám bude stačit podívat se na **minimální** a **maximální** hodnoty dat. Například u proměnné „rok narození“ by naši respondenti neměli být narozeni později než v roce 1995 (máme rok 2013 a respondenti měli být starší 18 let). Dřívější datum narození není jasné, ale nejstarší občance ČR je momentálně 109 let, držme se tedy limitu 1904 jako

nejmenšího možného roku narození. U hodnot 1904–1995 tedy máme důvod domnívat se, že jsou v pořádku. Často se však mohou vyskytnout chyby vzniklé při zápisu (např. rok 11982 či naopak vynechání číslice – rok 198). Tato data je potřeba opravit.

Někdy se může stát, že respondenti nevědí, jak odpovědět. Potom můžete na jednoduchou otázku („*Kolik je vám let*“) získat velmi různé formáty odpovědí:



17. 13. Jaký je Váš věk?

Textová otázka, zodpovězeno: 2851x, nezodpovězeno: 40x

- -
- ??
- nad60
- 17
- 18 6×
- 19let
- 19 126×
- 20 408×
- 20let 2×
- 21 501×
- 21let
- 22let 3×
- 22 427×
- 22.5
- 23let
- 23 417×
- 24let
- 24 294×
- 25 246×
- 25+
- 26 131×
- 27 79×
- 28 49×
- 28let
- 29 22×
- 29.5
- 30 25×
- 30.9
- 31 17×
- 32 11×
- 33 11×
- 34 4×
- 35 7×
- 36 7×
- 37 5×
- 38 6×
- 39 6×
- 40 4×
- 41 2×
- 42 3×
- 43 2×
- 44
- 45 3×
- 46 3×
- 47
- 48
- 49 2×
- 50 2×
- 57 2×
- 100
- 1985

Co s chybnými daty?

Narazíme-li na chybnou hodnotu, máme v zásadě několik možností:

- **Zjistit chybu a nahradit chybný zápis správnou hodnotou.** Například pokud chyba vznikla při přepisu papírového dotazníku do elektronické tabulky, je možné dotazník dohledat a chybu opravit. Stejně postupujeme i v případě, že respondenti nevyplnili pole tak, jak jsme chtěli (např. hodnotu „23let“ si překódujeme jen na „23“).

- Pokud není možné zjistit chybu, můžeme **prohlásit odpověď za chybějící** a nakládat s ní, jako by nebyla otázka vůbec zodpovězena. Variantně můžeme respondenta úplně vyřadit ze souboru.

Co s chybějícími daty?

Kromě chybných dat je potřeba zkoumat i **chybějící hodnoty**. Vyplatí se před samotnou analýzou zkontrolovat, kolikrát se vyskytly v odpovědích varianty „nevím / nemohu odpovědět".

Jsou odpovědi rozděleny náhodně? Nemá výskyt nevím souvislost s nějakou jinou proměnnou?

Pro kontrolu můžeme rozdělit soubor na skupiny záznamů s chybějícími hodnotami a bez nich, porovnat charakteristiky obou souborů, nebo nechat korelovat vyplnění/nevyplnění s jinou proměnnou (o korelacích bude řeč v dalších modulech).

3 Práce s datovým souborem

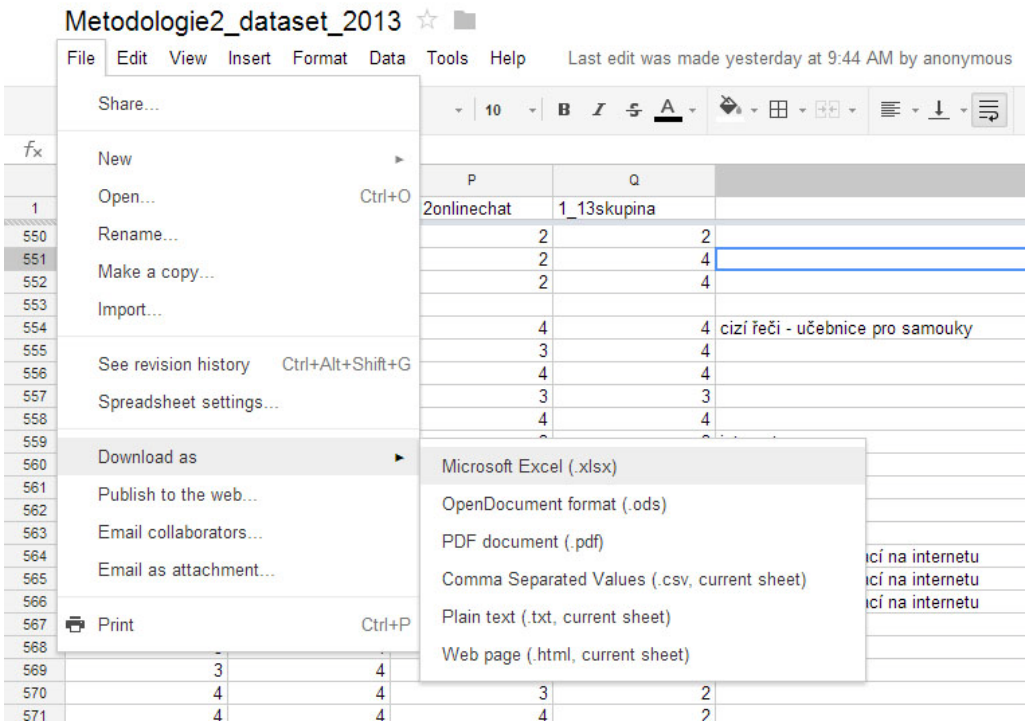
Dřív než začneme pracovat s datovým souborem, je potřeba zmínit několik zásad.

1. Ať už pracujeme v jakémkoliv programu, je vždy důležité pravidelně **zálohovat data**. Ponechte si zálohovaný původní datový soubor, ať se k němu v případě nejistot můžete vrátit. Zálohujte si také průběžnou práci – při analýze často vytváříte nové proměnné, o které byste mohli bez zálohování přijít. Při nepozornosti si také můžete přemazat některá data, proto je vhodné mít zazálohovaných několik posledních verzí souborů s daty.
2. Pokud pracujete ve **sdíleném souboru**, dbejte na to, aby byly kroky jednotlivých výzkumníků odlišitelné a zpětně dohledatelné. Pokud to prostředí neumožňuje, zvažte jinou variantu způsobu práce s daty.
3. Než začnete analyzovat, data **zkontrolujte** a **pečlivě popište**.

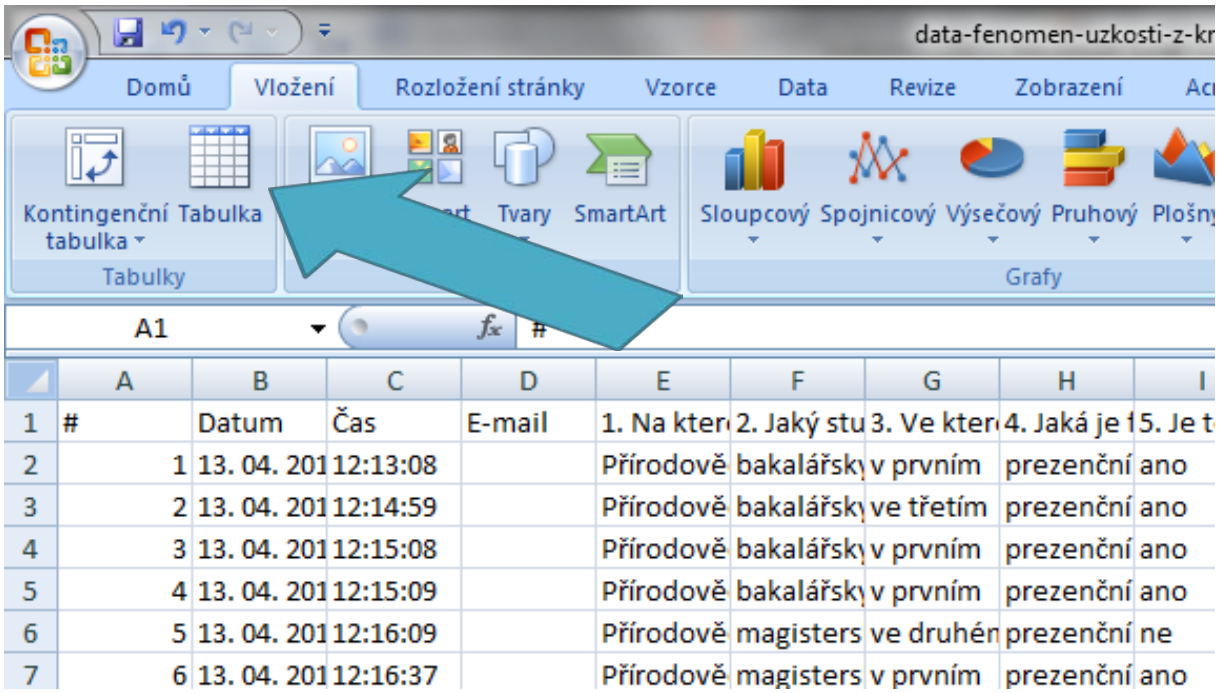
Stažení tabulky

V tomto semestru budeme pracovat se souborem, který jsme si společně vytvořili v Google dokumentech. Většinu operací, které budeme používat, lze provádět přímo v Google Spreadsheets. Pro práci v Excelu je možné si stáhnout tabulku z Google dokumentů pomocí funkce „**Download as**".

Stažení souboru ve formátu .xls:

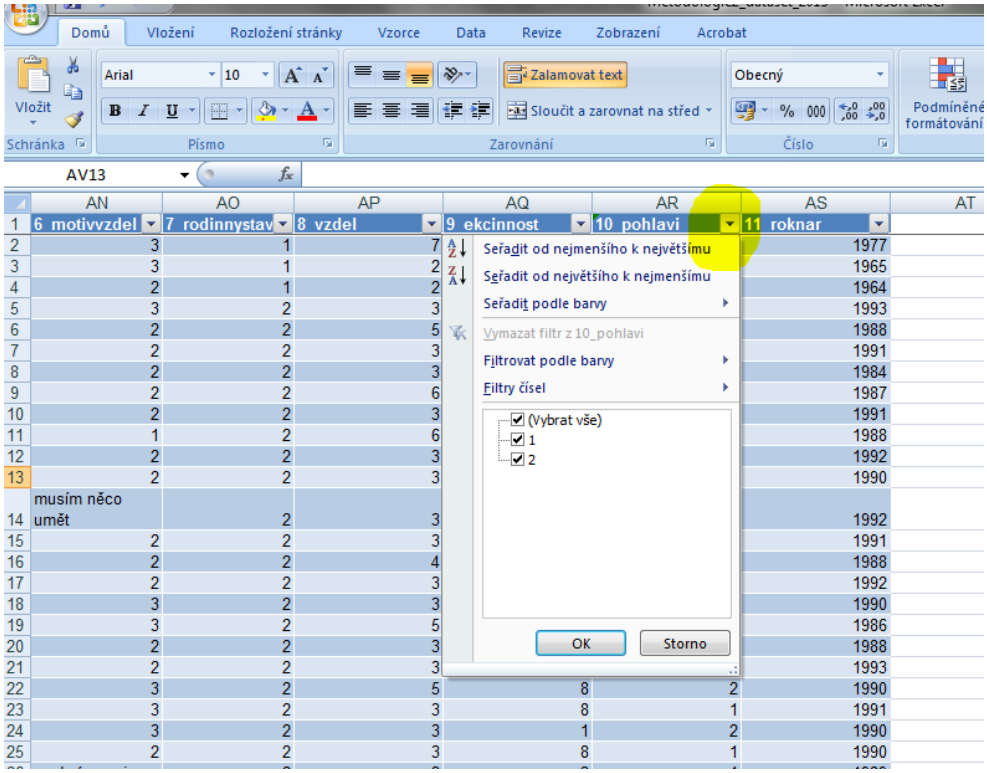


V Excelu je poté pro práci s daty vhodné data převést na inteligentní tabulku pomocí funkce „**Tabulka**“ v listu “**Vložení**”:



Excel rozpozná záhlaví a převede data na přehlednější tabulku.

Někdy nechceme pracovat s celým datovým souborem, ale zajímají nás například pouze ženy. V Excelu si můžeme jednoduše vyfiltrovat rozkliknutím položky v záhlaví:



	AN	AO	AP	AQ	AR	AS	AT
1	6 motivvzdel	7 rodinnystav	8 vzdel	9 ekcinnost	10 pohlavi	11 roknar	
2	3	1	7			1977	
3	3	1	2			1965	
4	2	1	2			1964	
5	3	2	3			1993	
6	2	2	5			1988	
7	2	2	3			1991	
8	2	2	3			1984	
9	2	2	6			1987	
10	2	2	3			1991	
11	1	2	6			1988	
12	2	2	3			1992	
13	2	2	3			1990	
14	musím něco umět	2	3			1992	
15	2	2	3			1991	
16	2	2	4			1988	
17	2	2	3			1992	
18	3	2	3			1990	
19	3	2	5			1986	
20	2	2	3			1988	
21	2	2	3			1993	
22	3	2	5	8	2	1990	
23	3	2	3	8	1	1991	
24	3	2	3	1	2	1990	
25	2	2	3	8	1	1990	

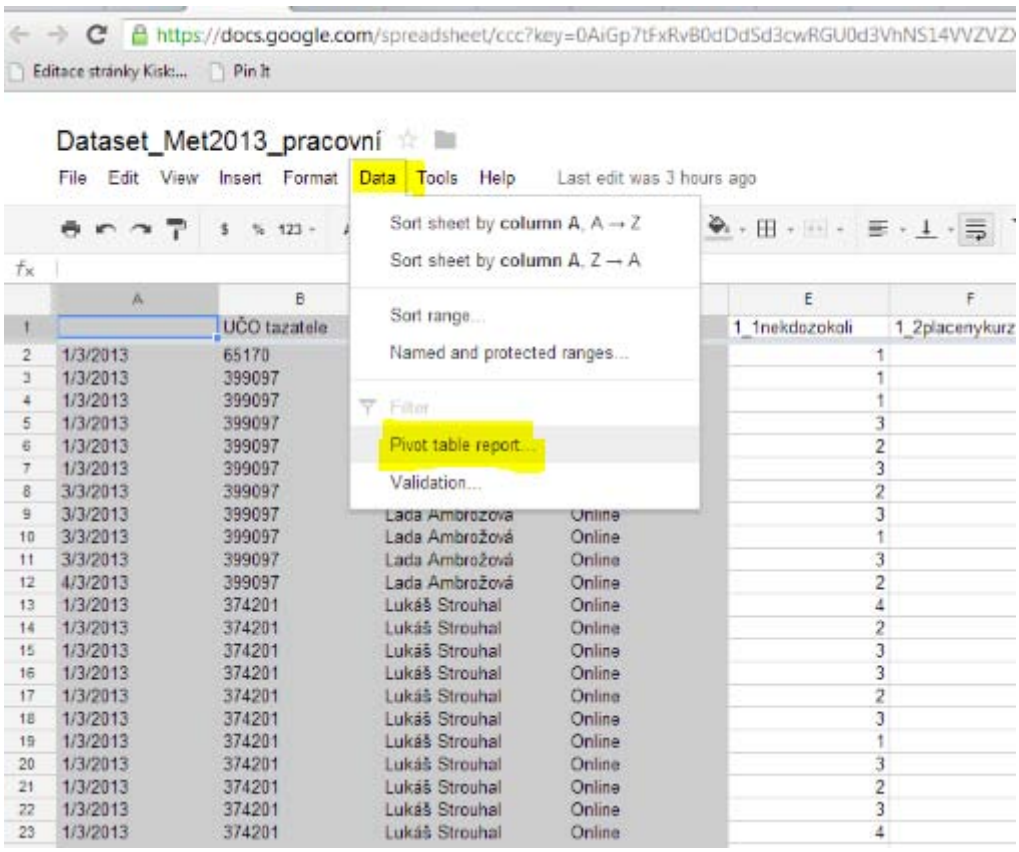
Popis rozložení hodnot proměnných

Pro počítání absolutních četností v Excelu slouží příkaz **COUNTIF**.

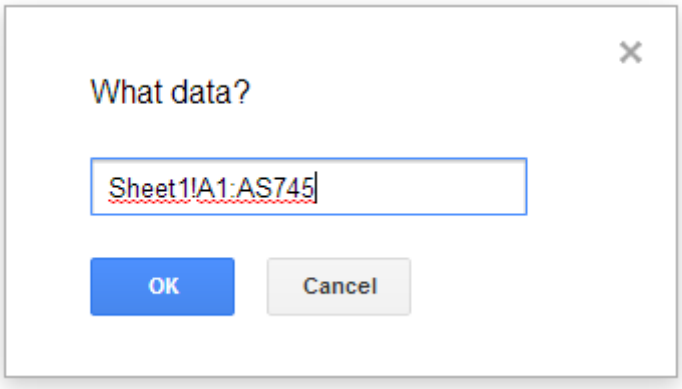
A	B
Prodejce	**Faktura**
Novák	15 000
Novák	9 000
Horák	8 000
Horák	20 000
Novák	5 000
Veselý	22 500
Vzorec	**Popis (výsledek)**
=COUNTIF(A2:A7;"Novák")	Počet faktur od Nováka (3)
=COUNTIF(A2:A7;A4)	Počet faktur od Horáka (2)
=COUNTIF(B2:B7,"< 20000")	Počet faktur s hodnotou nižší než 20 000 (4)
=COUNTIF(B2:B7,">="&B5)	Počet faktur s hodnotou vyšší nebo rovnou 20 000 (2)

Zdroj: http://office.microsoft.com

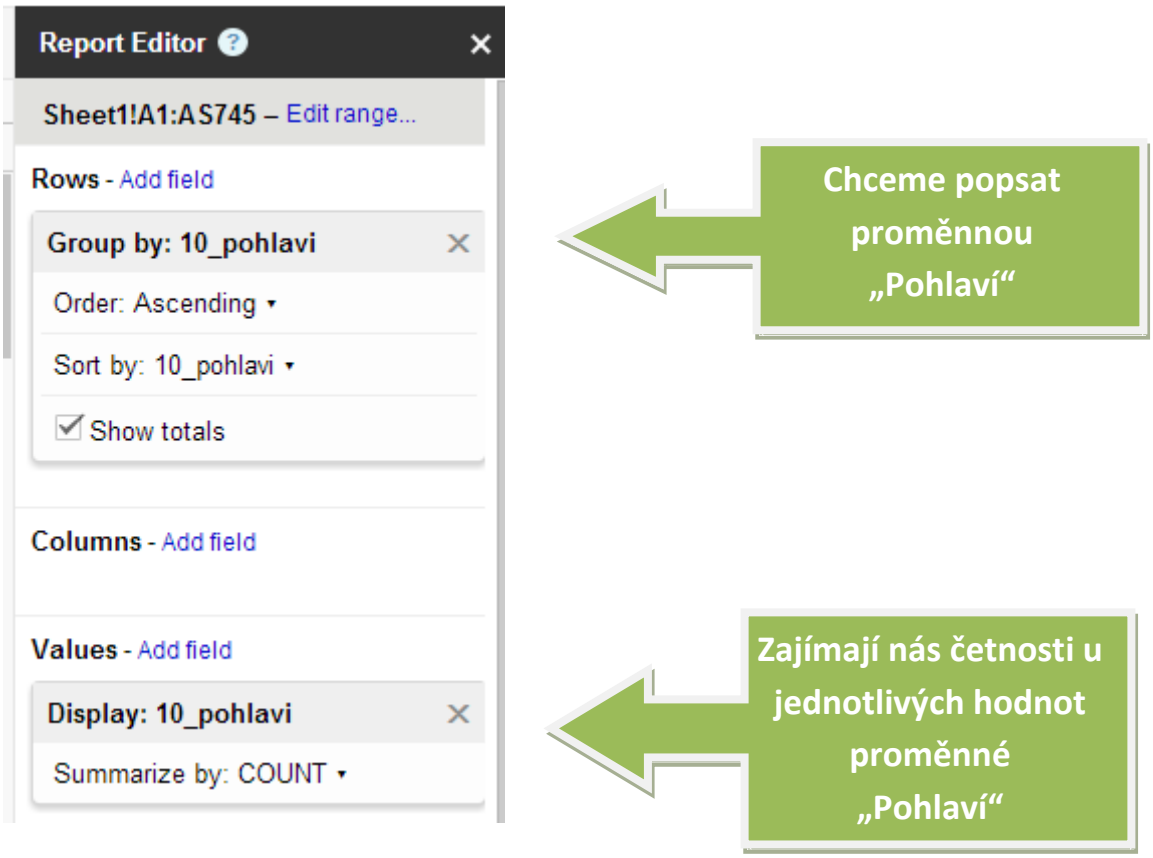
Příkaz COUNTIF nám spočítá výskyt konkrétní varianty hodnoty proměnné. Pro vytvoření tabulky četností je však užitečnější funkce „pivot tables“. Najdete ji v sekci „Data“.



Aplikace se vás nejprve zeptá na rozsah dat. Dávejte si pozor, abyste zahrnuli celou tabulku.



Nová tabulka se vám objeví na novém listu. Tabulku četností vytvoříte tak, že v položce „Řádky“ / „Rows“ specifikujete proměnnou, kterou chcete popsat a proces výpočtu hodnot. Pro tabulku četností budeme nejčastěji používat příkaz „COUNT“.

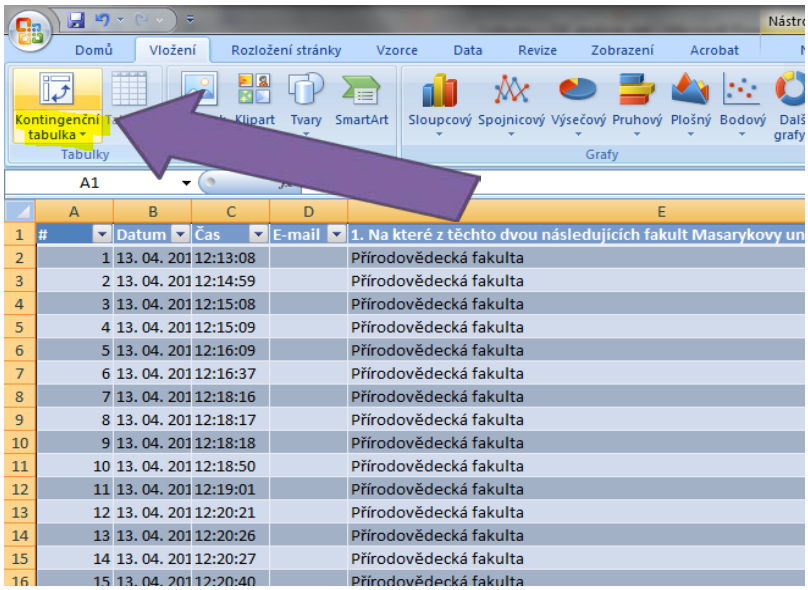


Zpracování v Google Spreadsheets může chvilku trvat, proto buďte trpěliví, pokud tabulka nebude hned reagovat na zadané změny.

Pokud jste si nepřekódovali odpovědi předem, výsledná tabulka bude obsahovat naše kódy, před publikováním je tedy třeba ji ještě upravit – místo kódů (např. „1“) by výsledná tabulka měla obsahovat reálné hodnoty proměnných (např. „muž“).

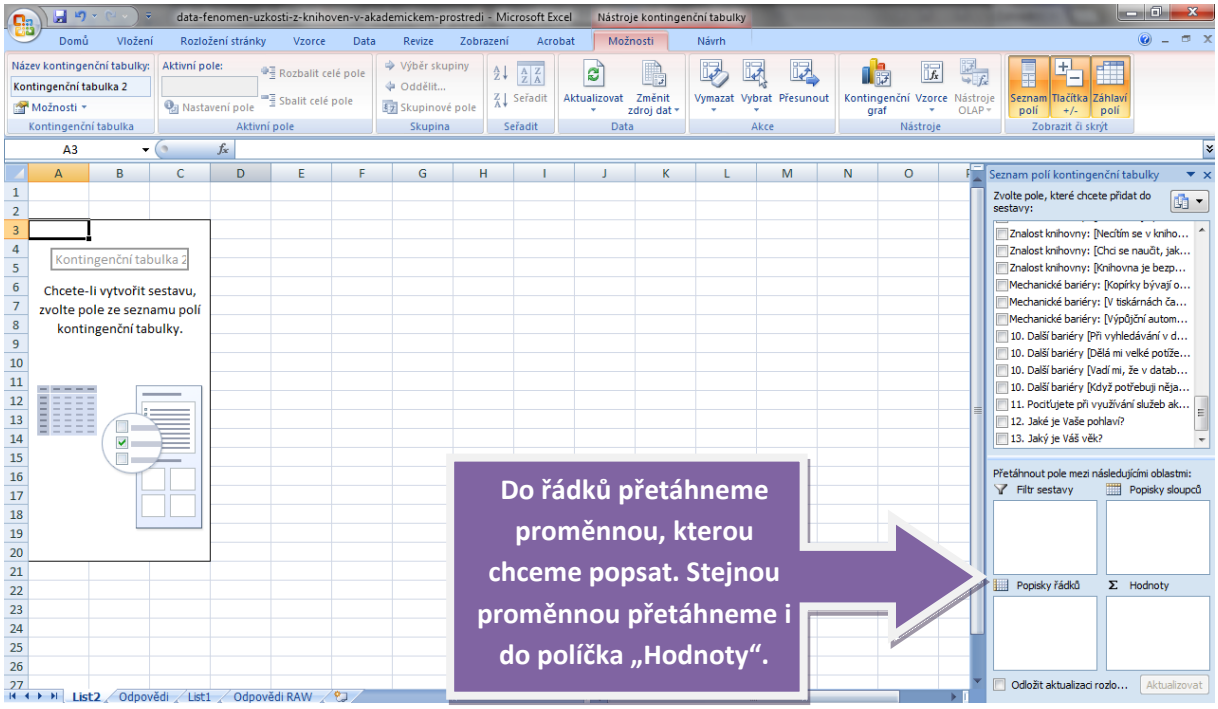
Jste:	**Četnost odpovědí**	**Validní relativní četnost**
Muž	80	40 %
Žena	120	60 %
Celkem	**200**	**100 %**

Pokud jste se rozhodli pracovat v Excelu, je postup velmi podobný. Tabulku vytvoříte tak, že označíte data, se kterými chcete pracovat, a zvolíte možnost „**Kontingenční tabulka**“ na kartě „**Vložení**“.



#	Datum	Čas	E-mail	1. Na které z těchto dvou následujících fakult Masarykovy un
1	13. 04. 201	12:13:08		Přírodovědecká fakulta
2	13. 04. 201	12:14:59		Přírodovědecká fakulta
3	13. 04. 201	12:15:08		Přírodovědecká fakulta
4	13. 04. 201	12:15:09		Přírodovědecká fakulta
5	13. 04. 201	12:16:09		Přírodovědecká fakulta
6	13. 04. 201	12:16:37		Přírodovědecká fakulta
7	13. 04. 201	12:18:16		Přírodovědecká fakulta
8	13. 04. 201	12:18:17		Přírodovědecká fakulta
9	13. 04. 201	12:18:18		Přírodovědecká fakulta
10	13. 04. 201	12:18:50		Přírodovědecká fakulta
11	13. 04. 201	12:19:01		Přírodovědecká fakulta
12	13. 04. 201	12:20:21		Přírodovědecká fakulta
13	13. 04. 201	12:20:26		Přírodovědecká fakulta
14	13. 04. 201	12:20:27		Přírodovědecká fakulta
15	13. 04. 201	12:20:40		Přírodovědecká fakulta

Na novém listu se objeví prostředí pro tvorbu kontingenčních tabulek. Pro tvorbu tabulek četností budeme využívat zatím jen možnosti popisů řádků:

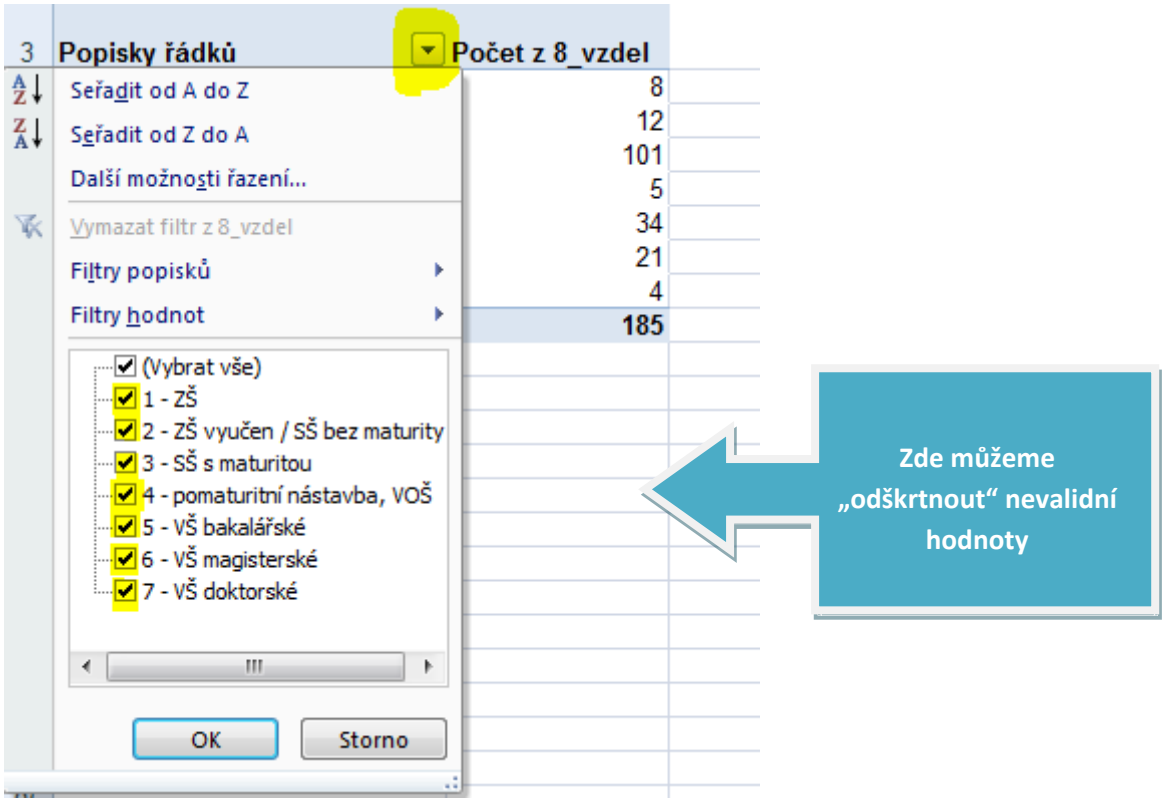


Pro ukázku si vytvořme tabulku se vzděláním:

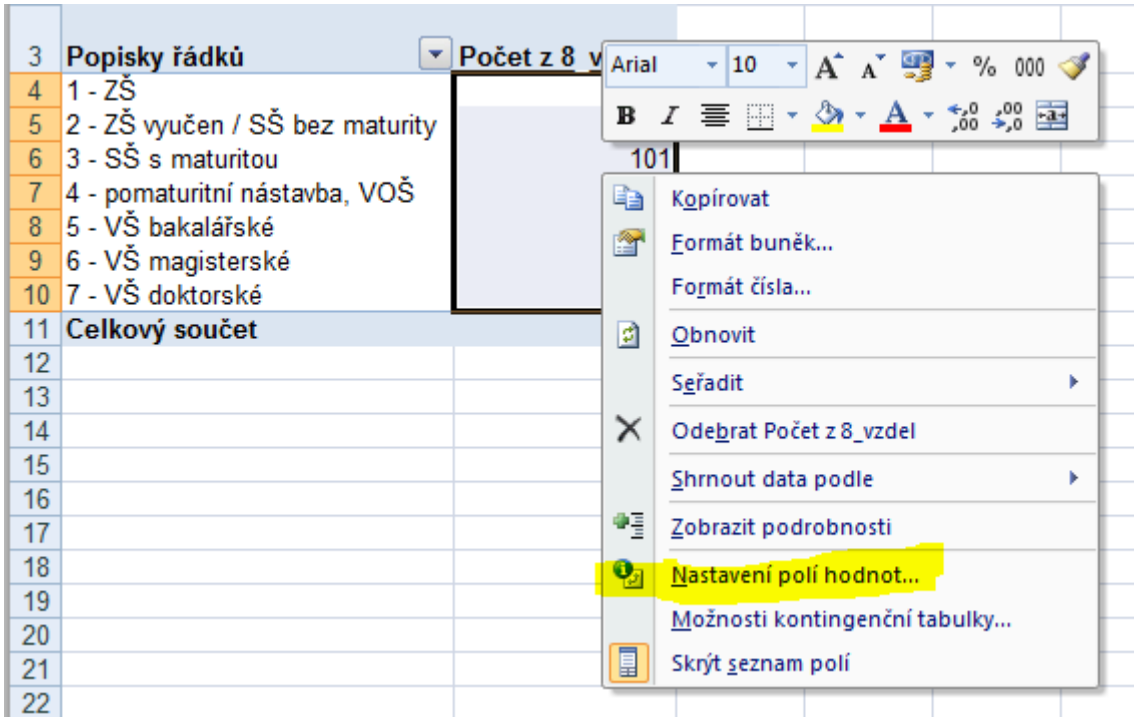
Popisky řádků	Počet z 8_vzdel
1 - ZŠ	8
2 - ZŠ vyučen / SŠ bez maturity	12
3 - SŠ s maturitou	101
4 - pomaturitní nástavba, VOŠ	5
5 - VŠ bakalářské	34
6 - VŠ magisterské	21
7 - VŠ doktorské	4
Celkový součet	**185**

Tato tabulka ukazuje hodnoty výskytu jednotlivých variant odpovědí u proměnné „Dokončené vzdělání“

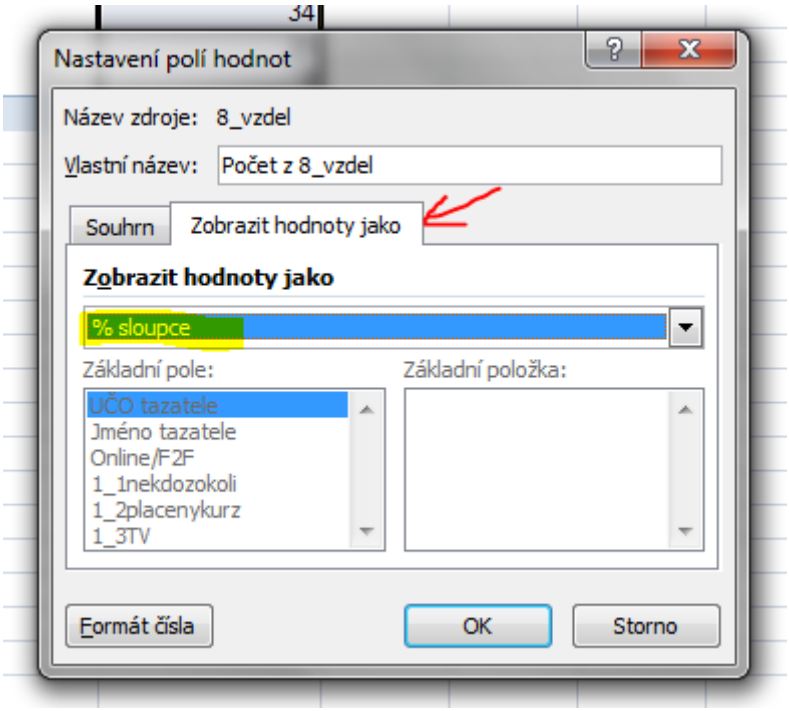
Pokud máme v otázce varianty odpovědí, které nechceme zahrnovat do analýzy (tzv. nevalidní odpovědi – tedy odpovědi typu „nevím“, „neodpověděl“), můžeme je odškrtnout v rozbalovacím menu:



Chceme-li přepočítat absolutní četnosti na relativní četnosti, klikneme na datovou oblast pravým tlačítkem myši a zvolíme možnost „**Nastavení polí hodnot**“:



Vybereme záložku „**Zobrazit hodnoty jako**“ a zvolíme „% sloupce“. Absolutní hodnoty se přepočítají na procenta:



Získáme tak **relativní četnosti**:

Popisky řádků	Počet z 8_vzdel
1 - ZŠ	4,32%
2 - ZŠ vyučen / SŠ bez maturity	6,49%
3 - SŠ s maturitou	54,59%
4 - pomaturitní nástavba, VOŠ	2,70%
5 - VŠ bakalářské	18,38%
6 - VŠ magisterské	11,35%
7 - VŠ doktorské	2,16%
Celkový součet	**100,00%**

Minimální a maximální hodnoty

Minimální a maximální hodnoty lze rozpoznat už z popisu rozloženíproměnných. U spojitých nekategorizovaných dat ale popis rozložení četností nepoužíváme, proto je výhodnější znát příkaz na rychlé zjištění minimálních a maximálních hodnot. V Excelu i v Google Spreadsheet se tyto hodnoty zjišťují pomocí funkce MIN a MAX. Zapisují se do políčka jako příkaz ve tvaru

„=MIN(datová oblast)“ či **„=MAX(datová oblast)“**

	A
1	Data
2	10
3	7
4	9
5	27
6	2

Vzorec	Popis (výsledek)
=MIN(A2:A6)	Nejmenší z výše uvedených čísel (2)
=MIN(A2:A6;0)	Nejmenší z výše uvedených čísel a čísla 0 (0)

	A
1	Data
2	10
3	7
4	9
5	27
6	2

Vzorec	Popis (výsledek)
=MAX(A2:A6)	Největší z výše uvedených čísel (27)
=MAX(A2:A6;30)	Největší z výše uvedených čísel a čísla 30 (30)

Využívejte podpory a nápovědy!

Pokud si nejste jistí provedením příkazu, využívejte podpory Microsoft Office i Google Spreadsheets. Na internetu lze najít také spoustu videotutorialů a návodů. V nejhorším případě pište na sucha@phil.muni.cz ☺.

Literatura

Disman, M. (2002) Jak se vyrábí sociologická znalost. Praha: Karolinum.

Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124.

Wheelan, Ch. (2013) Naked Statistics. New York: W. W. Norton & Company Ltd.

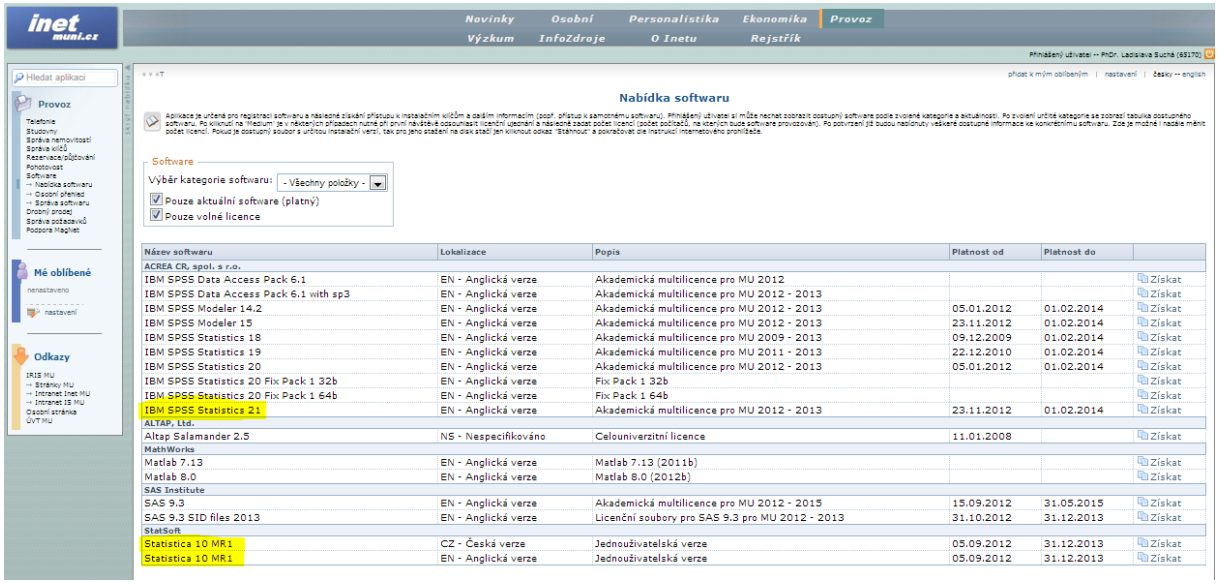
Návod pro práci s SPSS

Návody pro práci s programem SPSS pro kurz Metodologie pro Informační studia a knihovnictví 2 (jaro 2013)

Ladislava Zbiejczuk Suchá

Instalace programu

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



Program si můžete stáhnout ve formátu ISO. Pro spuštění je tedy nutné jej vypálit na DVD nebo vytvořit virtuální disk. Při registraci nezapomeňte uvést registrační kód dostupný v INETu.

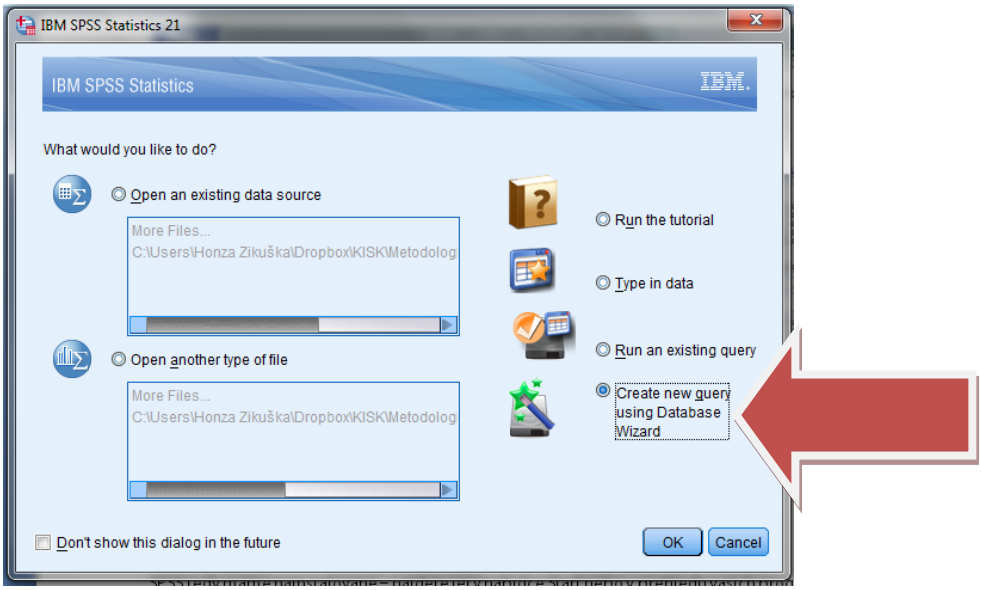
Dostupných je hned několik druhů licencí – doporučuji vybrat licenci **IBM SPSS Statistics 21** (nejnovější verze programu).

Otevření souborů s daty

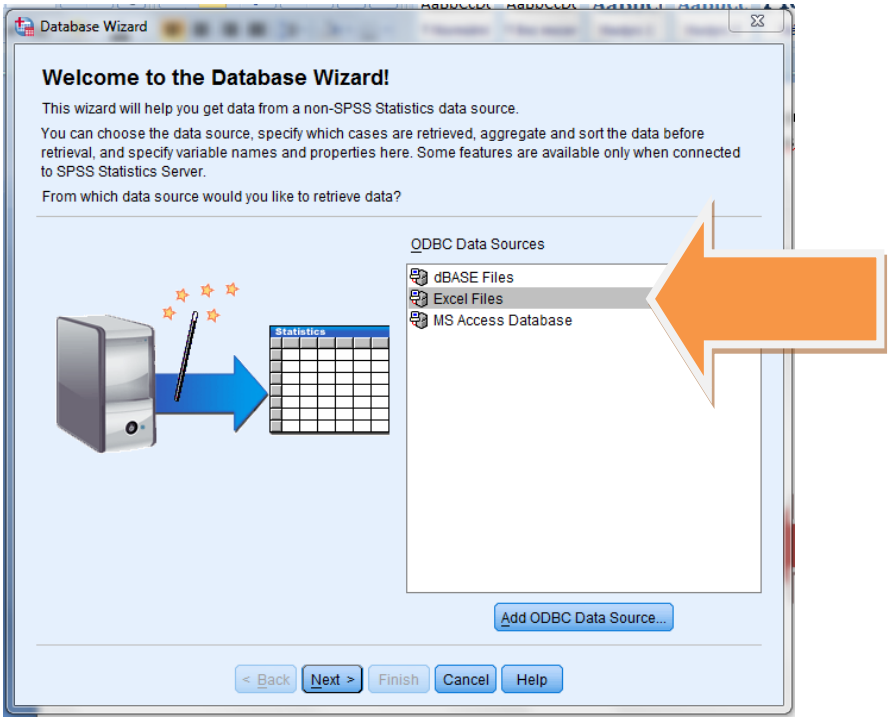
SPSS tedy máme nainstalované – najdete jej v nabídce Start nebo v přehledu vašich programů. Do SPSS můžete data dostat několika způsoby – ten nejzákladnější je přímé tvoření datasetu v SPSS. My ale budeme potřebovat pracovat s daty, která již máme ve formátu .xls.

Postupovat budeme následovně:

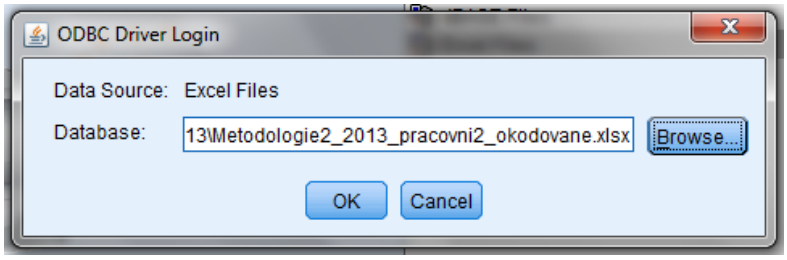
1. Uložíme si na své PC datový soubor ve formátu pro Excel (najdeme jej v ISu).
2. Pro převedení excelového souboru do souboru typu .sav spustíme „Database Wizzard“:



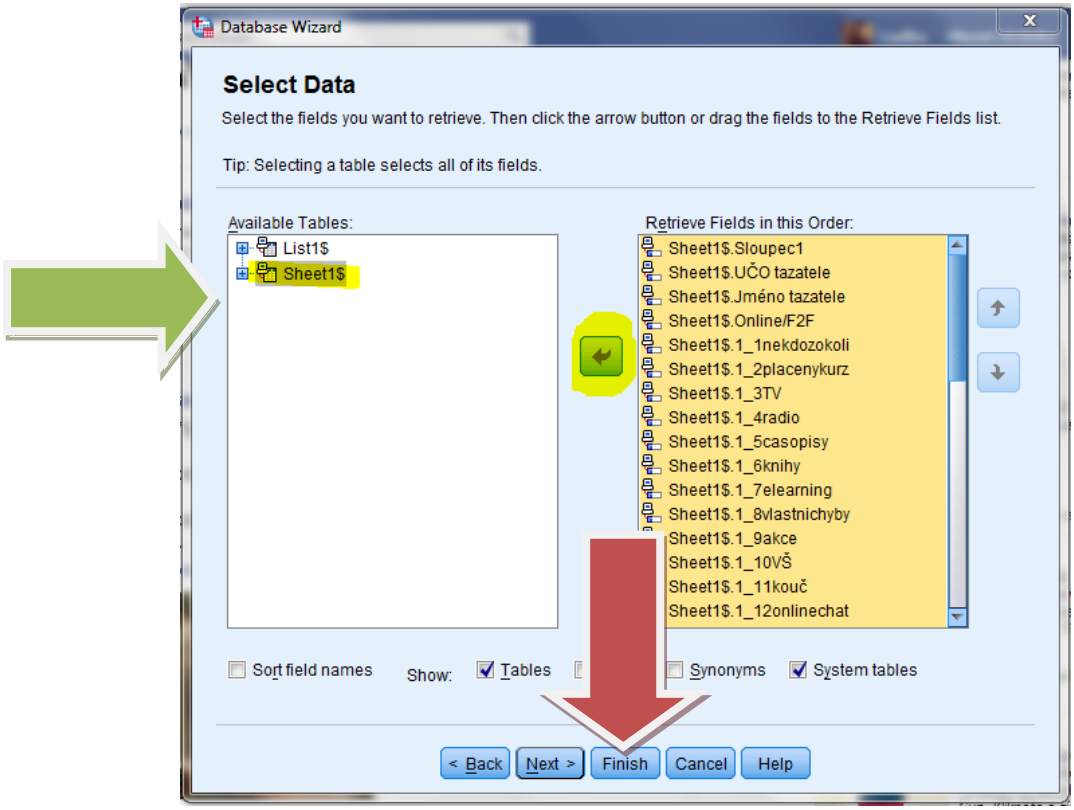
3. Z nabízených možností v dalším okně si vyberte „Excel files“:



4. Vyberte soubor ze svého PC:



5. Vyberte si oblast, kterou chcete převést a poté potvrďte stistem „**Finish**“

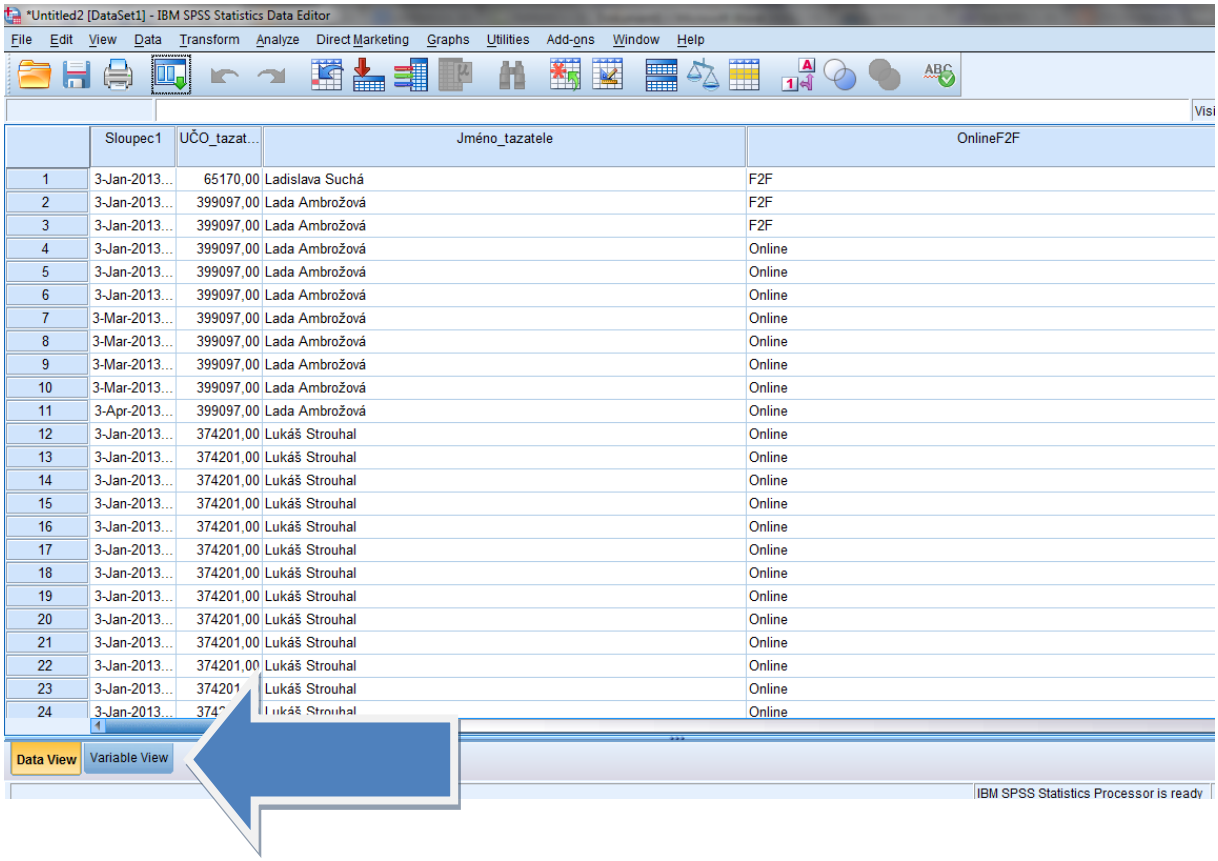


6. V počítači se vám otevřou dvě nová okna. Jedno přímo s datasetem a druhé je tzv. „Output“ – okno, kam se zapisují procesy a výsledky operací SPSS.

Práce s datasetem

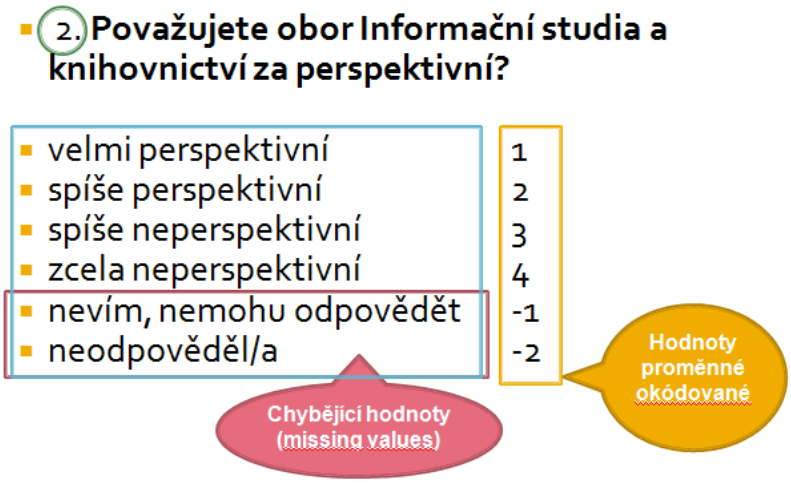
Dataset je neprve potřeba upravit a popsat. Všimněte si, že v SPSS lze přepínat mezi dvěma druhy zobrazení:

- pohled na data,
- pohled na proměnné.

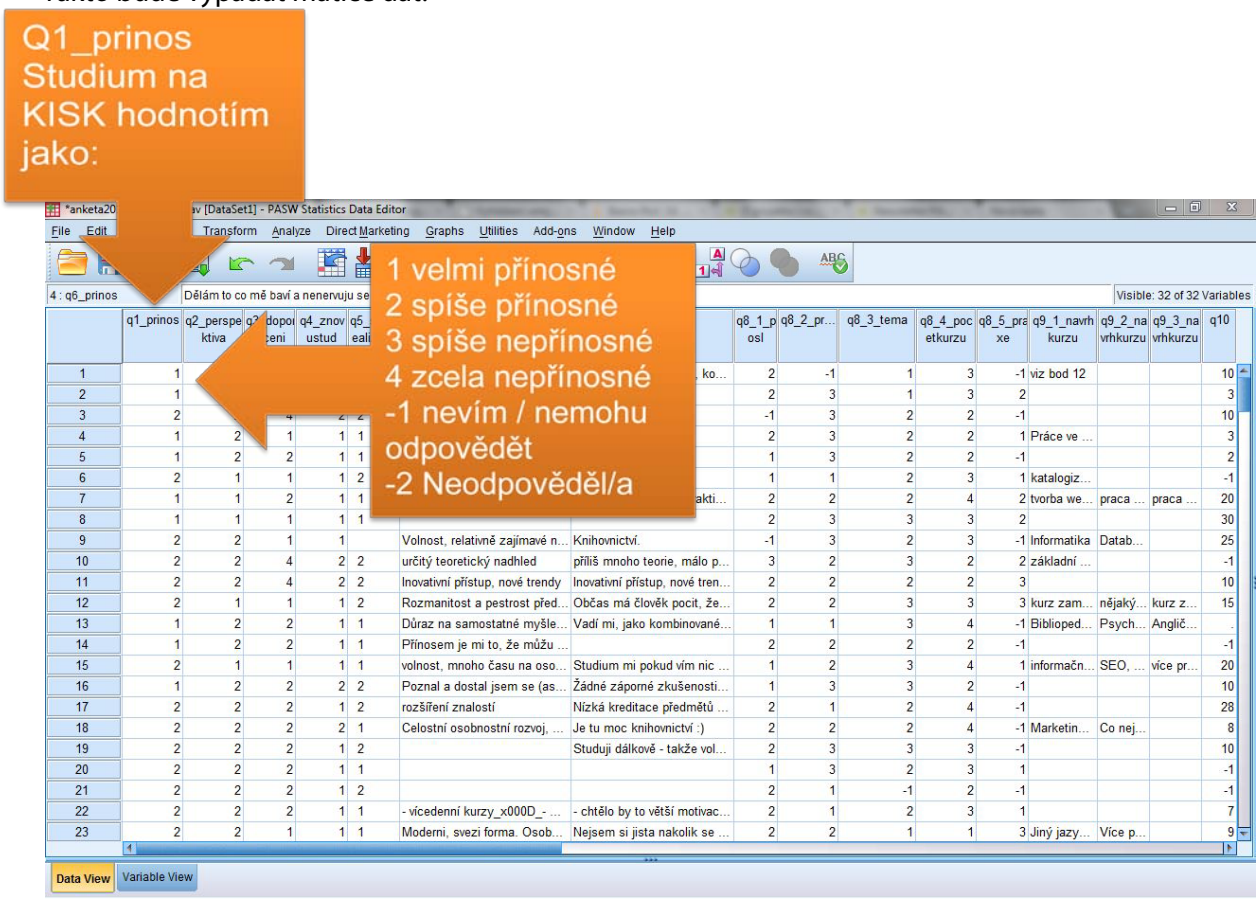


Pohled na data je velmi podobný tomu, co znáte z Excelu – co řádek, to respondent, co sloupec, to proměnná. **Pohled na proměnné** upřesňuje parametry jednotlivých proměnných.

Ukažme si to na příkladu této otázky:

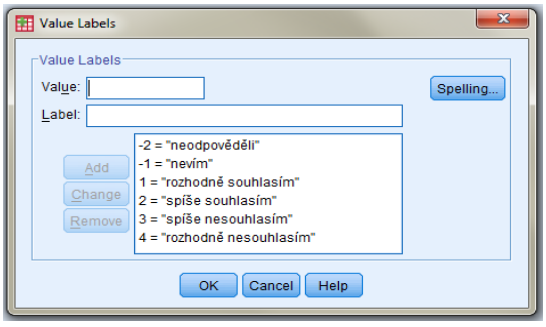


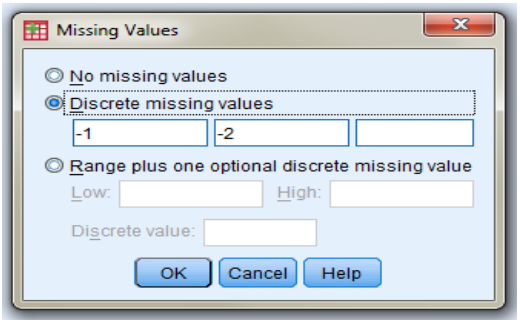
Takto bude vypadat matice dat:



Zároveň je potřeba popsat jednotlivé proměnné na kartě **Variable view**:

- *Name: zkrácené označení proměnné.*
- *Typ: číselné/slovní (SPSS potřebuje vědět, jaké operace může provádět s jednotlivými proměnnými)*
- *Decimal: desetinná místa (pouze kardinální proměnné) – automaticky jsou nastavena dvě desetinná místa, snižte si jejich počet na 0.*
- *Label: většinou kopírujeme znění otázky.*
- *Value labels: hodnoty proměnné – popíšeme všechny hodnoty proměnné včetně „missing values“*
- *Missing values: které hodnoty nezahrnujeme do dané analýzy – SPSS s nimi v konkrétních operacích nebude počítat.*
- *Measure: typ proměnné (nominální/ordinální/kardinální)*



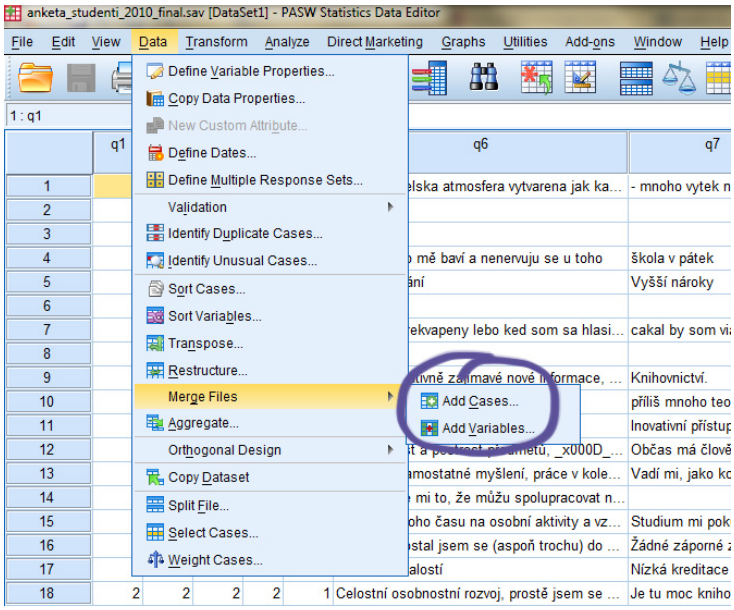


Ve studijních materiálech v ISu máte již datasety s popsanými proměnnými.

Slučování datových souborů

Někdy potřebujeme sloučit více datových souborů. Máme na výběr dvě varianty:

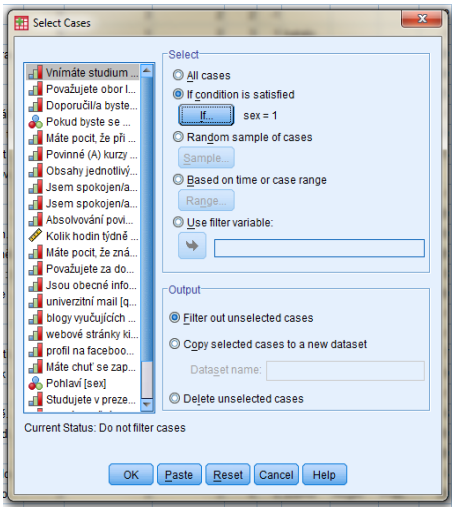
- Chceme sloučit více dat o stejných případech: Merge Files → Add variables
- Chceme sloučit soubory s různými jednotkami a stejnými proměnnými Merge Files → Add Cases



Výběr případů

Někdy naopak potřebujeme pracovat jen s některými případy (například se ženami):

- **Data → Select Cases**
- Lze vybírat náhodně nebo dle kritéria – pokud např. chceme pracovat jen s muži, pak musíme použít proceduru IF



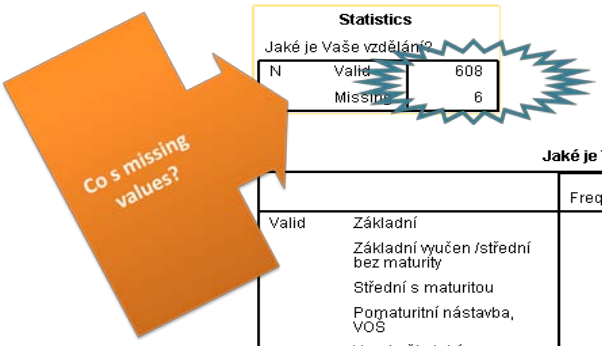
Kontrola dat

V SPSS probíhá kontrola dat se stejnou logikou jako v jakénkoliv jiném programu. Její provedení je jen jednodušší, protože SPSS je přizpůsobeno na provádění statistických operací. SPSS má také tu výhodu, že nám v Outputu dává tabulky již v té podobě, v jaké by se měly objevit v odborné práci – tedy kompletní tabulky četností s nevalidními validními absolutními i relativními hodnotami.

Pro použití v odborné práci je pouze třeba **přeložit popisky tabulek**.

Kontrola kategorizovaných dat

SPSS nám prostřednictvím jednoduchého příkazu **Analyze → Descriptive Statistics → Frequencies** (zde si vyberete konkrétní proměnnou) vrátí počet validních a nevalidních hodnot proměnných. Výsledky najdeme v okně Output:



Statistics

Jaké je Vaše vzdělání?

N	Valid	608
	Missing	6

Jaké je Vaše vzdělání?

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Základní	46	7,5	7,6	7,6
	Základní vyučen /střední bez maturity	62	10,1	10,2	17,8
	Střední s maturitou	307	50,0	50,5	68,3
	Pomaturitní nástavba, VOŠ	40	6,5	6,6	74,8
	Vysokoškolské	153	24,9	25,2	100,0
	Total	608	99,0	100,0	
Missing	System	6	1,0		
Total		614	100,0		

Stejně jako v případě SPSS nás bude zajímat výpis četností jednotlivých výskytů hodnot proměnné. Zde máme příklad chybného zápisu jména studentky či chybného zápisu v proměnné „pohlaví“:

Jméno výzkumníka

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Andrea Szászová	10	1,6	1,6	1,6
	Babora Ondrušová	1	,2	,2	1,8
	Barbora Ondrušová	9	1,5	1,5	3,3
	Barbora Pitašová	10	1,6	1,6	4,9
	Blanka Justová	10	1,6	1,6	6,5
	Blanka Svobodová	10	1,6	1,6	8,1
	Dagmar Chládková	11	1,8	1,8	9,9
	Dagmar Šiková	10	1,6	1,6	11,6
	Dalibor Bláha	10	1,6	1,6	13,2
	Daniela Králová	10	1,6	1,6	14,8

Statistics

Pohlaví

N	Valid	613
	Missing	1

Chyba: proměnná „Pohlaví“ by neměla nabývat hodnoty 7

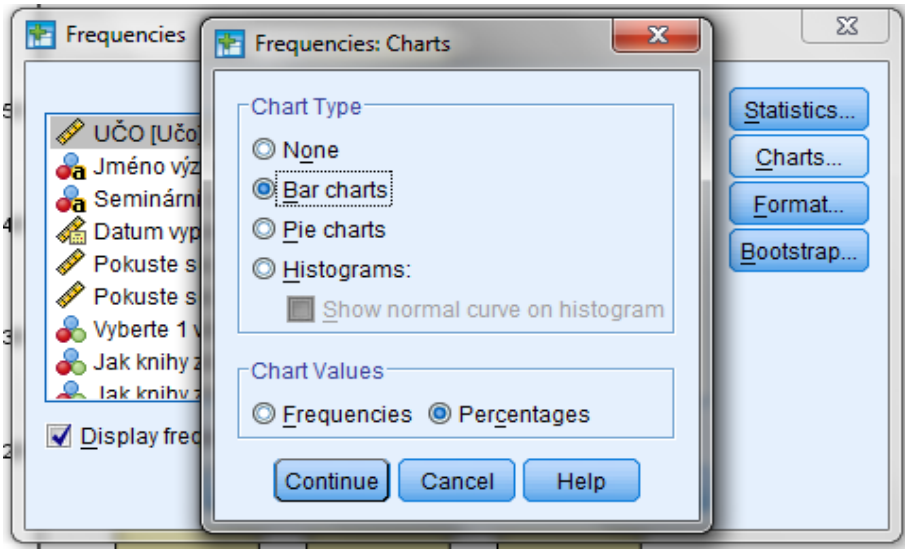
Pohlaví

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Muž	279	45,4	45,5	45,5
	Žena	333	54,2	54,3	99,8
	7	1	,2	,2	100,0
	Total	613	99,8	100,0	
Missing	System	1	,2		
Total		614	100,0		

Poté co naleznete chybná data, můžete je v datasetu vyhledat pomocí příkazu CTRL+F stejně jako v Excelu.

Tabulky četností a grafy v SPSS

Tabulky četností v SPSS získáme příkazem Analyze → Descriptive Statistics → Frequencies . Grafy vytvoříme cestou Analyze → Descriptive Statistics → Frequencies → **Charts**.



Modus a medián v SPSS

Modus, medián a aritmetický průměr jednoduše získáte v SPSS touto cestou:

Analyze → Descriptive Statistics → Frequencies → **Statistics → Mean, Median, Mode**

Metodologie pro Informační studia a knihovnictví 2

Modul 4: Kódování a rekódování. Deskriptivní statistika – popis dat I

Co se dozvíte v tomto modulu?

- Co zjišťujeme u nominálních proměnných?
- Co zjišťujeme u ordinálních proměnných?
- Jak zjistit modus a medián?
- Jak popsat grafy?

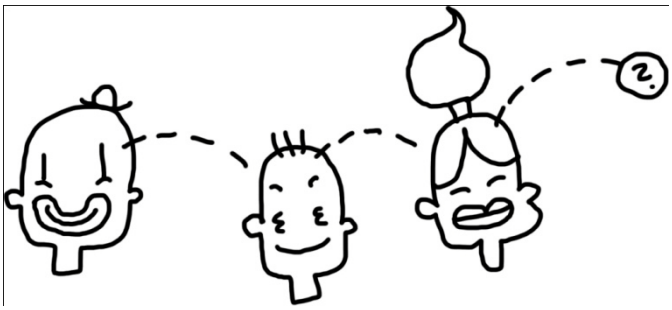
V tomto modulu si ukážeme, jak popsat kategorizovaná data.

Obsah

1 Kódování a rekódování

Náš dataset obsahuje kódy odpovědí. Tedy například místo odpovědi „velmi spokojen/a“ je uveden kód odpovědi „1“. Kódy jsou užitečné z nejen pro statistické softwary a aplikace, ale i pro snížení chybovosti při zápisu dat. Například pokud kombinujete sběr dat online a offline, je výhodné si nechat z online aplikace vygenerovat pouze číselné kódy a odpovědi sesbírané v terénu „doťukat“ do tabulky ve formě kódů. Výrazně se tím šetří čas, i pokud budete dále zpracovávat data pouze v Excelu.

Další práce s datasetem se liší, pokud pracujete v Excelu a ve statistickém software – například v SPSS.



Práce v Excelu

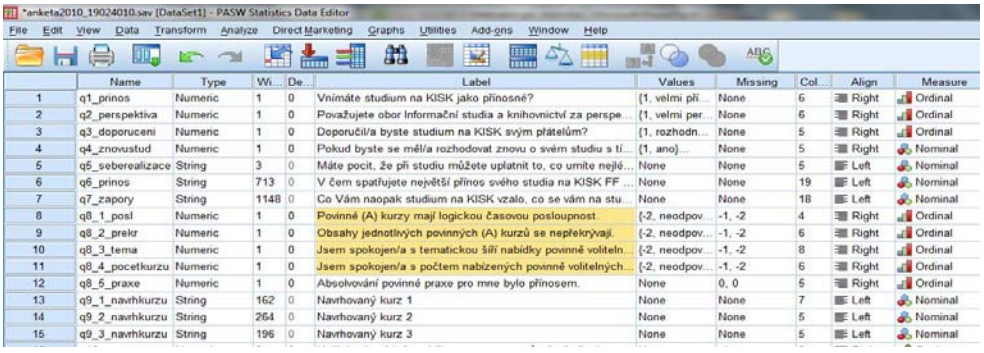
Pokud pracujete v Excelu, je výhodné si data opět překódovat tak, aby se vám ve výsledných tabulkách opět objevovaly celé odpovědi, nikoliv jen kódy. Jednoduše to uděláte pomocí funkce „Najít a nahradit“ (CTRL+H).

> *Protože takovou práci s dokumenty určitě umíte, dataset nemusíte celý překódovat, ale budete jej mít nahraný v ISu již překódovaný.*

Pokud se rozhodnete pracovat s okódovaným souborem, nezapomeňte ve výsledných tabulkách přepsat kódy odpovědí na skutečné odpovědi.

Práce v SPSS

Pro práci v SPSS má smysl kódy ponechat. SPSS pracuje přímo s kódy, kterým přiřazujete popisky („labels“). Jednoduše přepínáte mezi zobrazením responsí a zobrazením proměnných a jejich popisu.


*anketa2010_19024010.sav [DataSet1] - PASW Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Wi...	De...	Label	Values	Missing	Col...	Align	Measure
1	q1_prinos	Numeric	1	0	Vnímáte studium na KISK jako přínosné?	{1, velmi při...	None	6	Right	Ordinal
2	q2_perspektiva	Numeric	1	0	Považujete obor Informační studia a knihovnictví za perspe...	{1, velmi per...	None	6	Right	Ordinal
3	q3_doporuceni	Numeric	1	0	Doporučil/a byste studium na KISK svým přátelům?	{1, rozhodn...	None	5	Right	Ordinal
4	q4_znovustud	Numeric	1	0	Pokud byste se měl/a rozhodovat znovu o svém studiu s tí...	{1, ano}...	None	5	Right	Nominal
5	q5_seberealizace	String	3	0	Máte pocit, že při studiu můžete uplatnit to, co umíte nejlé...	None	None	5	Left	Nominal
6	q6_prinos	String	713	0	V čem spatřujete největší přínos svého studia na KISK FF ...	None	None	19	Left	Nominal
7	q7_zapory	String	1148	0	Co Vám naopak studium na KISK vzalo, co se vám na stu...	None	None	18	Left	Nominal
8	q8_1_posl	Numeric	1	0	Povinné (A) kurzy mají logickou časovou posloupnost.	{-2, neodpov...	-1, -2	4	Right	Ordinal
9	q8_2_prekr	Numeric	1	0	Obsahy jednotlivých povinných (A) kurzů se nepřekrývají.	{-2, neodpov...	-1, -2	6	Right	Ordinal
10	q8_3_tema	Numeric	1	0	Jsem spokojen/a s tematickou šíří nabídky povinně voliteln...	{-2, neodpov...	-1, -2	8	Right	Ordinal
11	q8_4_pocetkurzu	Numeric	1	0	Jsem spokojen/a s počtem nabízených povinně volitelných...	{-2, neodpov...	-1, -2	6	Right	Ordinal
12	q8_5_praxe	Numeric	1	0	Absolvování povinné praxe pro mne bylo přínosem.	None	0, 0	5	Right	Ordinal
13	q9_1_navrhkurzu	String	162	0	Navrhovaný kurz 1	None	None	7	Left	Nominal
14	q9_2_navrhkurzu	String	264	0	Navrhovaný kurz 2	None	None	5	Left	Nominal
15	q9_3_navrhkurzu	String	196	0	Navrhovaný kurz 3	None	None	5	Left	Nominal

Odlišná je i práce s tzv. „missing values“ (chybějícími hodnotami. Zatímco při práci v SPSS nebo statistických softwarech je vhodné je okódovat odlišným způsobem (např. -1 nebo 99) a programu „říci“, že se jedná o chybějící hodnoty, se kterými nemá počítat, při práci v Excelu můžeme ponechat políčka volná, případně i ponechat popisy typu „Neví/neodpověděl“. To, že se tyto hodnoty nezahrnují do analýzy, označujeme až při samotné tvorbě tabulky četností (viz předchozí týden).

2 Deskriptivní statistika (kategorizované proměnné)

Nejprve malé opakování:

- **Deskriptivní statistika** se zabývá popisem dat, jejich sumarizaci a prezentací.
- **Kategorizované proměnné** jsou všechny proměnné, jejichž hodnoty se nacházejí v určitých kategoriích. Jedná se tedy o nominální, ordinální i kardinální proměnné (pouze ale kardinální poměrové).

Různé druhy proměnných umožňují různé druhy popisu.

Popis nominálních proměnných

U nominálních proměnných zjišťujeme:

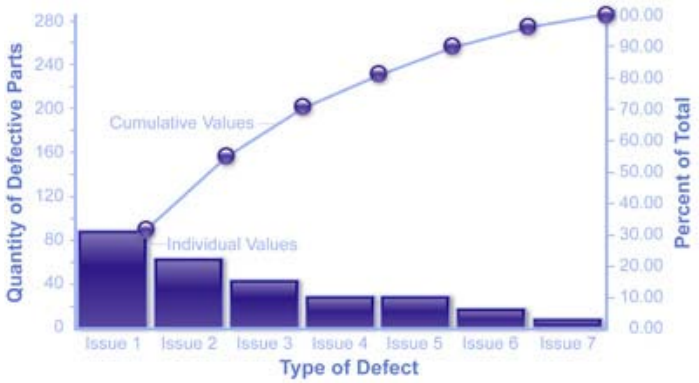
- **rozložení četností** variant znaku (pomocí tabulek četností),
- nejčastěji zastoupenou kategorii – **modus** (modálních kategorií někdy může být více než 1),
- **variační poměr**, který se vypočítá tak, že od jedné odečteme podíl četnosti modální kategorie a velikosti souboru.

Rozložení nominální proměnné můžeme – je-li to vhodné – znázornit i tzv. **Paretovým diagramem**. Paretův diagram (nebo také Paretův graf) kombinuje sloupcový a čárový graf. Sloupci jsou vyznačené četnosti jednotlivých kategorií seřazené podle velikosti, čarou je vyznačená kumulativní četnost. Paretův graf se využívá ve strategickém rozhodování a jako nástroj zlepšování kvality – dokáže velmi účinně zvýraznit důležité kategorie od nedůležitých – tzv. „vital few“ vs. „trivial many“ (Levine & Stephan 2010)

Paretův graf získáme v Excelu z této tabulky:

	Četnost	Kumulativní relativní četnost
Položka A		
Položka B		
Položka C		
Položka D		

Příklad Paretova diagramu:

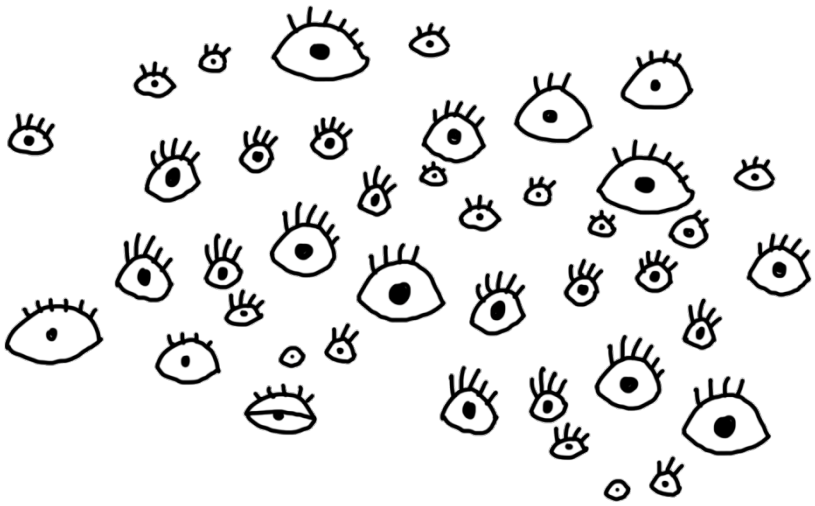


(zdroj: http://www.billiondollargraphics.com/paretochart.html)

Popis ordinálních proměnných

U ordinálních proměnných zjišťujeme:

- rozložení četností variant znaku (pomocí tabulek četností),
- nejčastěji zastoupenou kategorii – **modus** (modálních kategorií někdy může být více než 1),
- **medián** (mediánovou kategorii),
- variační poměr,
- další vlastnosti, jako je ordinální variance či normalizovaná ordinální variance (dorvar – těmi se ale nebudeme dopodrobna zabývat).



Rozložení četností

Zjištění rozložení četností je základní operací popisné statistiky. Ukázali jsme si jej už v minulém modulu. Při popisu rozložení četností vytvoříme vždy:

- tabulku četností,
- graf četností (koláčový či sloupcový).

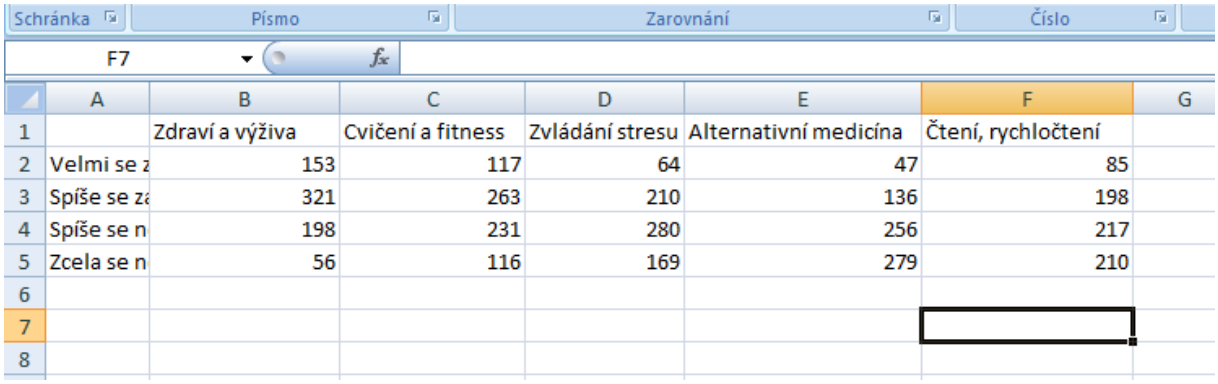
V grafu i v tabulce četností pracujeme vždy s validními četnostmi (tedy nezahrnujeme odpovědi typu „nevím“ nebo „*neodpověděl/a*“.

V případě nominálních proměnných je pro přehlednost vhodné kategorie ve sloupcovém diagramu seřadit dle výskytu od největší po nejmenší.

Porovnání rozložení četností

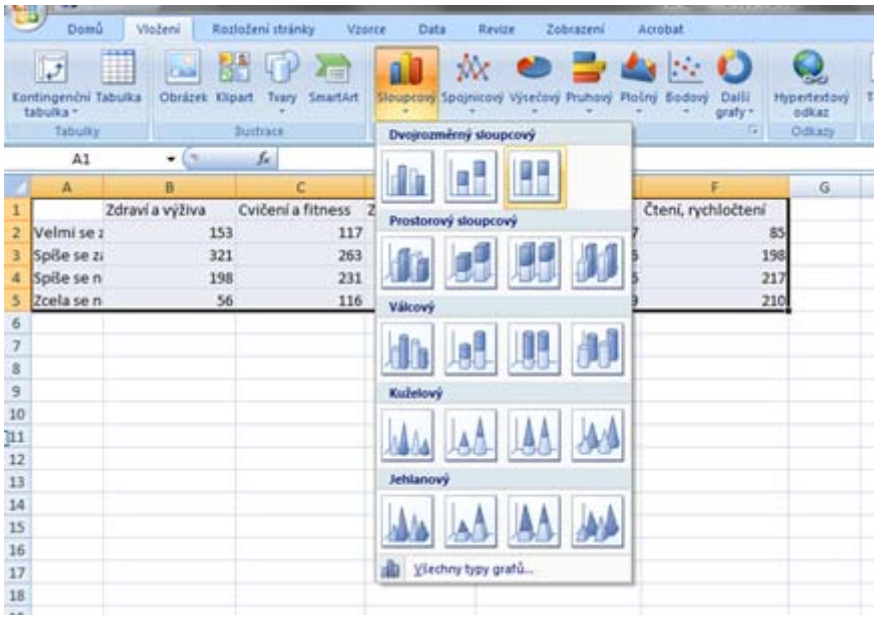
Pro zobrazení porovnání rozložení četností u baterií otázek se používají **skládané sloupcové grafy**.

Skládaný sloupcový graf můžete vytvořit tak, že si připravíte tabulku s absolutními validními četnostmi u jednotlivých kategorií:

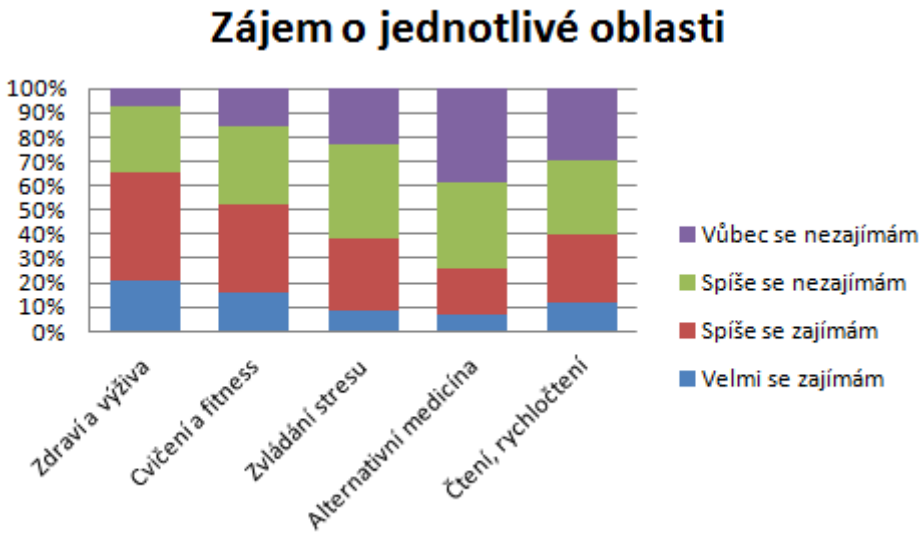


	A	B	C	D	E	F	G
1		Zdraví a výživa	Cvičení a fitness	Zvládání stresu	Alternativní medicína	Čtení, rychločtení	
2	Velmi se z	153	117	64	47	85	
3	Spíše se za	321	263	210	136	198	
4	Spíše se n	198	231	280	256	217	
5	Zcela se n	56	116	169	279	210	
6							
7							
8							

Tabulku si označíte a zvolíte možnost „Vložení“ – „Grafy“ – „Sloupcový“.



Výsledkem je skládaný sloupcový graf, který přehledně ukazuje rozdíly v rozložení jednotlivých proměnných.



Modus a medián

Pro připomenutí z minulého semestru si uveďme, v čem se liší MODUS a MEDIÁN (obě udávají tzv. míry centrální tendence a často se pletou):

> ***MODUS*** *je hodnota, která se v datech vyskytuje nejčastěji.*
> ***MODÁLNÍ KATEGORIE*** *je tedy nejpočetněji zastoupená kategorie.*
>
> ***MEDIÁN*** *dělí řadu výsledků seřazených podle velikosti na dvě stejně početné poloviny.*
> ***MEDIÁNOVÁ KATEGORIE*** *je ta, ve které je dosaženo 50% všech údajů, postupujeme-li od první kategorie výše.*

Jestliže je počet položek ve výzkumném souboru lichý, pak platí:

$$\textbf{Medián} = x_{(n+1)/2}$$

Jestliže je počet položek ve výzkumném souboru sudý, pak platí:

$$\textbf{Medián} = 0{,}5(x_{n/2}+x_{n/2+1})$$

Představte si otázku na počet dětí. Odpovědi respondentů jsou {0,1,1,2,2,3,5}.

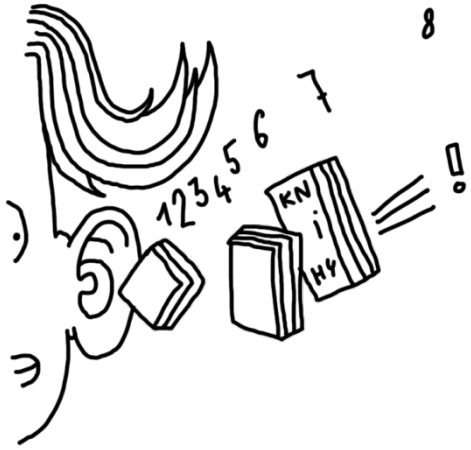
- *V souboru jsou dvě modální kategorie (tedy kategorie s nejvyšším počtem výskytů) – jsou to hodnoty 1 a 2.*
- *Mediánová kategorie je 2. Medián je na rozdíl od aritmetického průměru málo citlivý k odlehlým (extrémním) hodnotám. Pokud by byly odpovědi respondentů {0,1,1,2,2,3,5,10}, medián stale zůstává roven 2.*

Modus a medián v Excelu

V Excelu existují na výpočet mediánu a modu jednoduché příkazy MEDIAN a MODE. Syntaxe zápisu je snadná:

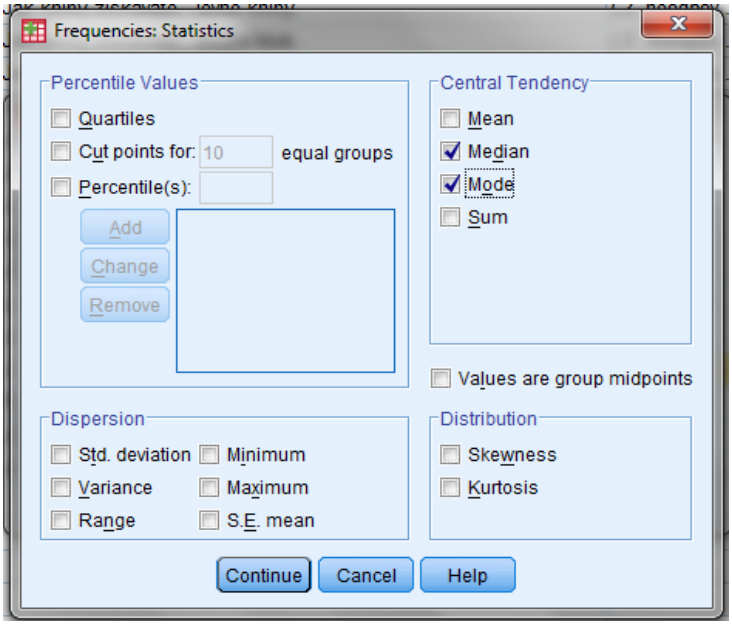
- =MEDIAN(datová oblast) – např. =MEDIAN(A1:A730)
- =MODE(datová oblast) – např. =MODE(A1:A730)

(Příkazy vypočítají medián a modus ze sloupce A, řádků 1-730.)



Modus a medián v SPSS

V SPSS vyberete v nabídce položky Analyze > Descriptive Statistics > Frequencies (zde zvolíte proměnnou) > Statistics > Median, Mode.



3 Tipy pro vytváření grafů

Levine a Stephan (2010) shrnují několik tipů pro prezentaci dat prostřednictvím grafů v akademickém prostředí:

- vždy si vyberte ten nejjednodušší graf,
- vždy používejte popisek grafu,
- popište obě osy,
- vyvarujte se ilustrací a zbytečného používání grafiky na pozadí nebo okrajích grafu,
- vyvarujte se používání módních piktogramů, které by mohly ztížit čitelnost dat,
- vertikální osa by měla začínat nulou (pokud nezačíná negativními hodnotami).

V neakademickém prostředí (např. pro účely marketingu) je využití grafiky vhodné, v prostředí akademickém je na prvním místě čitelnost dat. 3D efekty a vkládání obrázků mohou znemožnit čtení hodnot dat. Další tipy pro vytváření grafů najdete třeba zde.

Literatura

Hendl, J. *Přehled statistických metod analýzy dat*. Praha : Portál 2009

Levine, D. M., & Stephan, D. (2010). *Even you can learn statistics: A guide for everyone who has ever been afraid of statistics*. Upper Saddle River, N.J: FT Press.

Metodologie pro Informační studia a knihovnictví 2

Modul 5: Popis nekategorizovaných dat

Co se dozvíte v tomto modulu?

- Kdy používat modus, průměr a medián.
- Co je to směrodatná odchylka.
- Jak popsat distribuci dat.
- Jak zobrazovat spojité proměnné.

Obsah

Nekategorizované proměnné

Nekategorizované proměnné jsou ty proměnné, které mohou nabývat všech hodnot z daného intervalu. Může jednat o plat, věk, počet obyvatel města, délku pracovní zkušenosti v měsících…

Aritmetický průměr

Aritmetický průměr je třetí mírou centrální tendence. U kardinálních dat lze jako míry centrální tendence využívat všechny tři:

- modus,
- medián,
- aritmetický průměr.

Aritmetický průměr je ukazatelem „průměrné" hodnoty, nemusí být ale vždy ukazatelem nejvhodnějším – vhodné je jej kombinovat s mediánem. Aritmetický průměr je totiž velmi citlivý na extrémní hodnoty. I jedna extrémní hodnota může výrazně posunout aritmetický průměr.

Příklad: V roce 2010 byl podle serveru Platy.cz průměrný měsíční plat 23 300 Kč. Medián byl však na hodnotě 21 000 Kč. Znamená to, že průměr vychýlil menší počet jedinců s výrazně vyšším platem.

Průměrný měsíční plat (v Kč)	Medián (Kč)	Rozdíl (v %)
23 300	21 000	11%

Zdroj: Platy.cz

Pro připomenutí:

Modus se používá, pokud:

- rozdělení má více vrcholů,
- chceme zjistit nejčastější hodnoty.

Medián používáme, pokud:

- jsou data ordinální nebo kardinální,
- chceme znát střed rozložení dat,
- (v kombinaci s průměrem) pokud soubor obsahuje extrémní hodnoty,
- jestliže je rozložení dat zešikmené.

Aritmetický průměr je vhodné používat, pokud

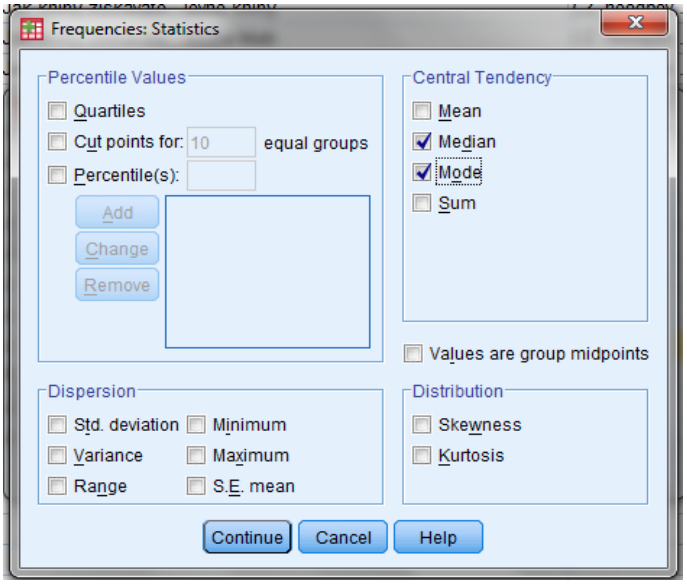
- jsou data kardinální,
- rozložení je symetrické,
- chceme použít statistické testy. (Hendl 2009)

Aritmetický průměr v Excelu

- Příkaz **PRŮMĚR**

Aritmetický průměr v SPSS

Pro zjištění hodnot měr centrální tendence v SPSS zadáte Analyze → Descriptive Statistics → Frequencies → **Statistics → Mean, Median, Mode**



Minimum, maximum a rozpětí

První charakteristiky nekategorizovaných dat, na které se díváme už při fázi čištění dat, jsou **minimální** a **maximální hodnoty**. Z nich také snadno spočítáme **rozpětí**.

Rozpětí je nejjednodušší míra variability a snadno se vypočítá jako rozdíl mezi nejvyšší a nejnižší hodnotou.

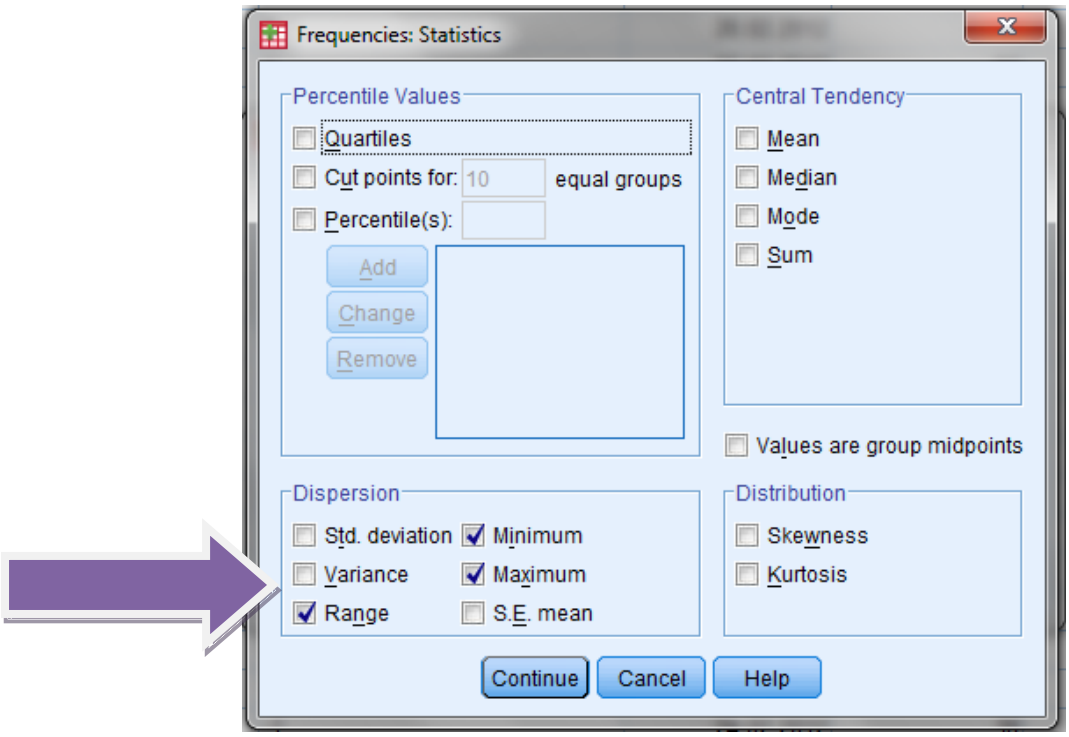
> *Např. Je-li minimální hodnota 18 a maximální 1024, rozpětí hodnot proměnné v souboru je 106.*

Minimum, maximum a rozpětí v Excelu

- Příkaz **MIN(oblast hodnot)**
- Příkaz **MAX(oblast hodnot)**
- Rozpětí jako rozdíl hodnot MAX a MIN

Minimum, maximum a rozpětí v SPSS

Vypočítání rozpětí můžete v SPSS zadat tímto řetězcem: **Analyze – Frequencies – Statistics:**



Rozptyl a směrodatná odchylka

Rozptyl je definován jako střední hodnota kvadrátů odchylek od střední hodnoty (průměru). Vyjadřuje variabilitu rozdělení souboru náhodných hodnot kolem její střední hodnoty. Při průměrování odchylek dělíme číslem n-1.

S rozptylem úzce souvisí **směrodatná odchylka**. Ta se vypočítá jako odmocnina z rozptylu. Vrací tedy míru rozptýlenosti do měřítka původních dat. V podstatě nám říká, uvnitř jakého intervalu okolo průměru leží zvolené procento případů – tedy čím je směrodatná odchylka menší, tím lépe pro aritmetický průměr.

Hendl (2009) srozumitelně vysvětluje, jak dochází k výpočtu směrodatné odchylky:

1. Nejprve si vypočítáme všechny odchylky od průměru (např. při hodu kostkou vždy spočítáme odchylku konkrétní hozené hodnoty od celkového průměru).
2. Umocnění na druhou převede záporné odchylky na kladná čísla. Zároveň zvýrazní váhu extrémnějších odchylek.
3. Sečteme kvadratických odchylek.
4. Dělením číslem n-1 získáme průměrnou kvadratickou odchylku.
5. Odmocnina (v případě směrodatné odchylky) převede výsledek do původního měřítka dat.

Pro názornost si pojďme ukázat příklad, který dobře znáte – hodnocení vyučujících na KISKu a směrodatnou odchylku tohoto hodnocení.

```
Zajímavost předmětu              není vůbec zajímavý .***X(*)**... je velmi zajímavý
Přínosnost předmětu              není vůbec přínosné ***X*(*)*.... je velmi přínosné
Obtížnost obsahu                       velmi snadný .....(.)**X** velmi obtížný
Náročnost na přípravu                  velmi snadný .....(*)X*... velmi obtížný
Dostupnost studijních zdrojů  velmi špatně dostupné .....(.).**X* velmi dobře dostupné
Jak učitel učí                         velmi špatný .***X(*)**... vynikající
Učitel jako odborník                 není odborníkem .....(.)***X* je odborníkem
```

```
Zajímavost předmětu              není vůbec zajímavý .....(.)...*X je velmi zajímavý
Přínosnost předmětu              není vůbec přínosné .....(.)..*X* je velmi přínosné
Obtížnost obsahu                       velmi snadný **X**(.)..... velmi obtížný
Náročnost na přípravu                  velmi snadný *X**.(.)..... velmi obtížný
Dostupnost studijních zdrojů  velmi špatně dostupné .....(.)**X** velmi dobře dostupné
Jak učitel učí                         velmi špatný .....(.)...*X vynikající
Učitel jako odborník                 není odborníkem .....(.)....X je odborníkem
```

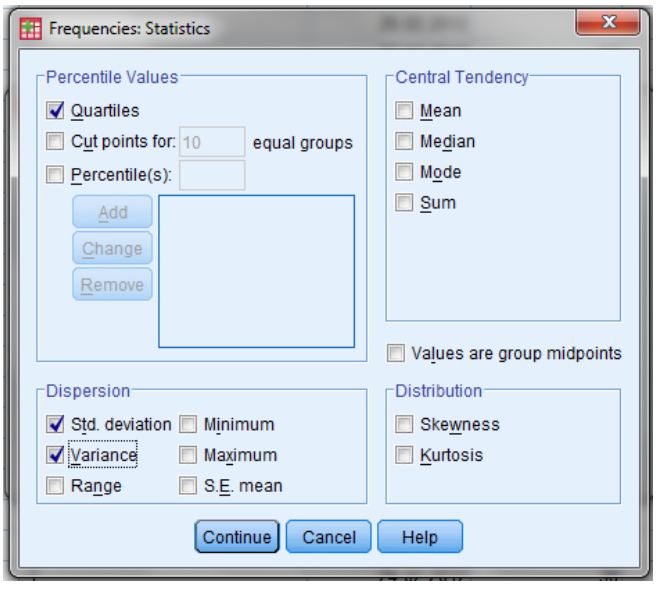
Průměrné hodnocení proměnné „Učitel jako odborník" je u obou vyučujících podobné – jeden vyučující má průměrné hodnocení 9, druhý má průměrné hodnocení 10. Směrodatná odchylka (zvýrazněná hvězdičkami) nám ale poskytne rychlou další informaci – říká nám, jak moc se hodnocení všech respondentů pohybovalo kolem průměru. Vidíme, že zatímco v druhém případě se hodnocení výjimečně shodovalo a studující se shodli na tom, že učitel je skutečný odborník, v prvním případě nebyla shoda zdaleka tak veliká.

Rozptyl a směrodatná odchylka v Excelu

- rozptyl – příkaz **VAR**
- směrodatná odchylka – příkaz **SMODCH.VÝBĚR**

Rozptyl a směrodatná odchylka v SPSS

Vypočítání rozptylu a směrodatné odchylky můžete v SPSS zadat tímto řetězcem: **Analyze – Frequencies – Statistics:**



Percentily

Percentil x je hodnota, pro kterou platí, že x procent případů má hodnotu menší nebo rovnu percentilu x.

Nejčastěji se využívají:

- **MEDIÁN** (x50)
- **KVARTILY** (x25, x50, x75)
- **DECILY** (x10, x20, x30, x40, x50, x60, x70, x80, x90)

Například vás může zajímat, jak jsou rozloženy příjmy obyvatel v horním a spodním percentilu. Tato informace spolu s mediánem ukazuje, jak moc jsou rozevřené pomyslné nůžky mezi „horní" a „spodní" vrstvou společnosti.

Jak vysoký je medián proti průměrné mzdě? (ve vybraných zemích OECD)

Země	spodních 10 %	medián	horních 10 %
Švédsko	56 %	89,8 %	150,9 %
Finsko	62,3 %	89,5 %	147,9 %
Kanada	44,6 %	89,1 %	166,9 %
Dánsko	60,9 %	89 %	150,4 %
Norsko	63,2 %	88,9 %	149 %
Japonsko	52,4 %	87,6 %	162,7 %
Nový Zéland	51,2 %	87,2 %	160,6 %
Německo	43,4 %	87 %	165,7 %
Česko	**49,3 %**	**85,2 %**	**153,1 %**
Itálie	56,1 %	85,1 %	156,6 %
Švýcarsko	56,6 %	84,9 %	153,4 %
Belgie	60,4 %	84,5 %	153,4 %
Nizozemí	51,7 %	84 %	158,8 %

Zdroj: http://finexpert.e15.cz/jak-se-lisi-prumerna-mzda-a-median

Percentil v Excelu

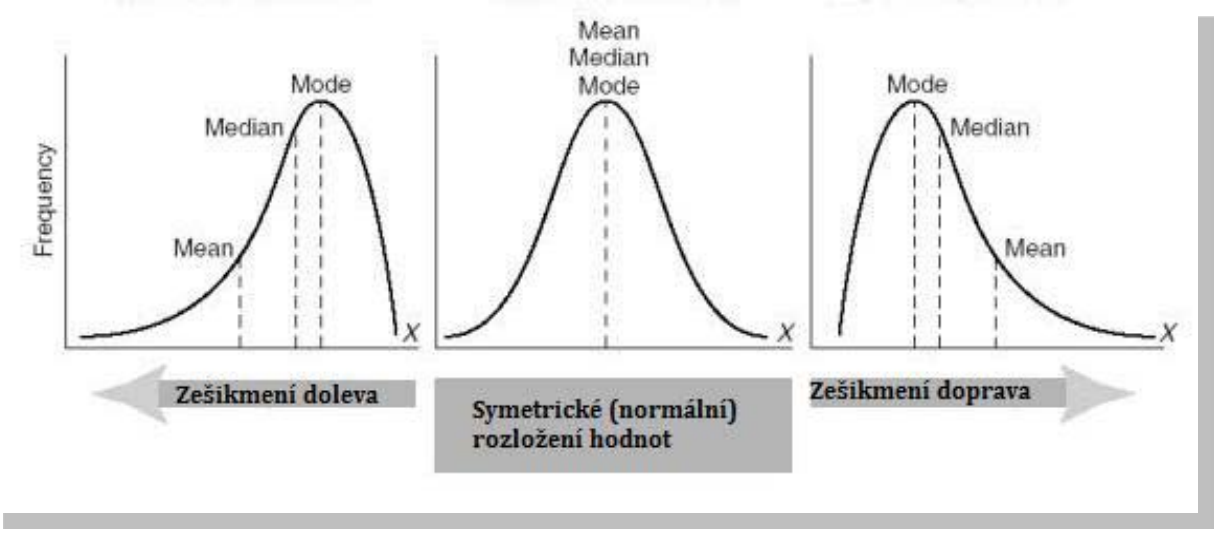
- Příkaz **PERCENTIL** (rozpětí dat;hodnota percentilu z intervalu 0-1)
  Tedy např. percentil 50 můžeme zapsat jako =PERCENTIL(A1:A30;0,5)

Percentil v SPSS

Vypočítání rozptylu a směrodatné odchylky můžete v SPSS opět zadat tímto řetězcem: **Analyze – Frequencies – Statistics (políčko Percentile Values)**.

Šikmost a špičatost

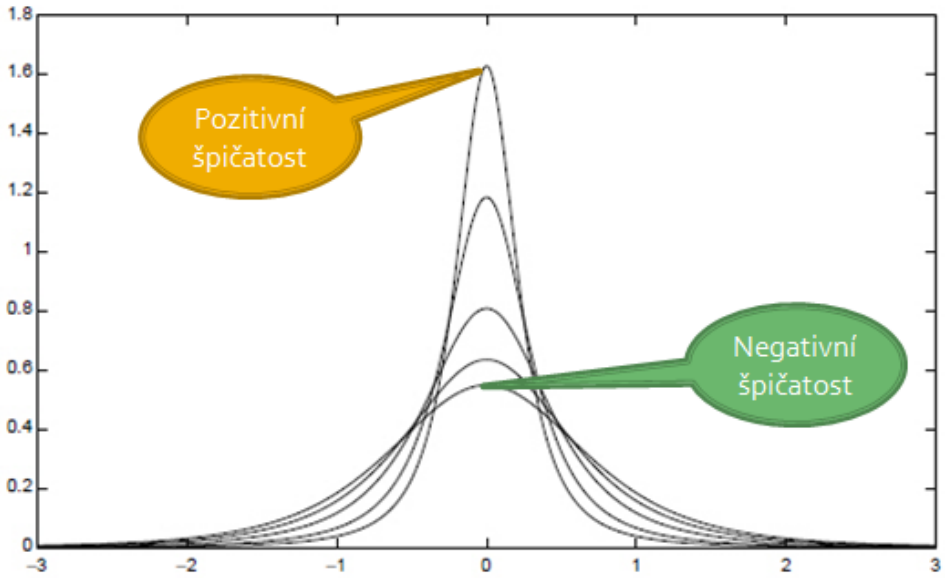
Spojitá data nabývají málokdy tzv. normálního rozložení. Při popisu dat si všímáme zešikmení a špičatosti dat.



Ad **šikmost**:

- **Symetrické (normální) rozložení** - aritmetický průměr, medián a modus mají stejné nebo velmi podobné hodnoty. (0)
- Pokud je aritmetický průměr větší než medián, který je zase větší než modus, znamená to, že je více případů menších než průměr a naše **rozložení je šikmé doprava**. (+)
- Třetí možností je, že je více případů větších než aritmetický průměr. Ten je pak menší než medián a ten je menší než modus. Naše **rozložení je šikmé doleva**. (-)

Špičatost zase udává, jak moc jsou data nakumulována v oblasti středních hodnot.



Šikmost a špičatost v Excelu

- příkaz **SKEW** (šikmost)
- příkaz **KURT** (špičatost)

Šikmost a špičatost v SPSS

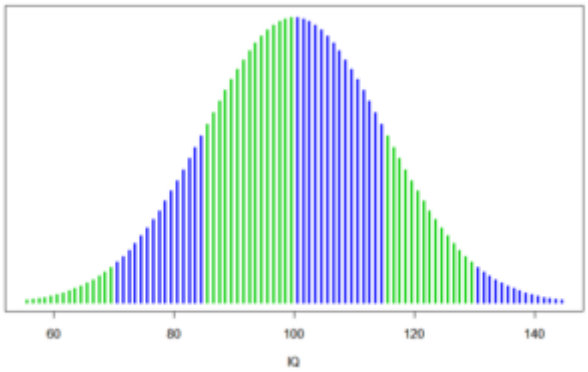
Analyze – Frequencies – Statistics (políčko Distribution).

Zobrazování kardinálních dat

Pro zobrazování kardinálních dat se používá několik možných grafů

Histogram

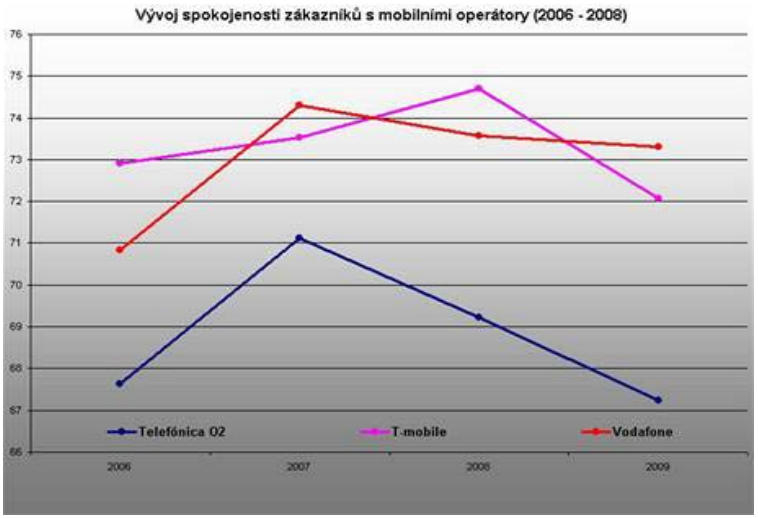
Histogram je podobný sloupcovému grafu, mezi jednotlivými sloupci ale nejsou mezery. Pracujete-li v Excelu, můžete využít klasický sloupcový graf.



Příklad histogramu – distribuce IQ v populaci (zdroj: IQscope.com)

Spojnicové grafy

Chcete-li ukázat, jak se hodnoty proměnné měnily v čase, je vhodné použít spojnicový graf.



Příklad využité spojnicového grafu – spokojenost s mobilními operátory 2006-2008

Bodové grafy

Bodové grafy zachycují jednotlivé hodnoty proměnných a využívají se v třídění druhého stupně jako zachycení toho, jak jedna proměnná ovlivňuje druhou (o tomto grafu více v dalších modulech).

Literatura

Hendl, J. *Přehled statistických metod analýzy dat*. Praha : Portál 2009

Levine, D. M., & Stephan, D. (2010). *Even you can learn statistics: A guide for everyone who has ever been afraid of statistics*. Upper Saddle River, N.J: FT Press.

Metodologie pro Informační studia a knihovnictví 2

Modul 6: Transformace proměnných

Co se dozvíte v tomto modulu?

- Jak vytvořit novou proměnnou pomocí rekódování?
- Jak vytvořit novou proměnnou pomocí aritmetických operací?
- Jak vytvořit novou proměnnou pomocí sčítání výskytů hodnot proměnné?
- Jak vytvořit novou proměnnou pomocí seřazení položek?

Obsah

Vytváření nových proměnných

Při analýze někdy potřebujeme zjistit a využít proměnné, které v dotazníku přímo nezkoumáme, ale můžeme si je snadno vytvořit z existujících proměnných. Například pro srozumitelnější analýzu může být výhodné kategorizovat si věkové skupiny. Nebo chceme z proměnné „*rok narození*" vytvořit srozumitelnější proměnnou „*věk*".

Obecně se s proměnnými dá dělat řada jednoduchých operací, my budeme využívat především:

- rekódování,
- aritmetické operace,
- kategorizace dle percentilů,
- sčítání výskytů,
- seřazení položek.

POZOR!!! Nové proměnné vždy vytváříme do nového sloupce – tak, abychom neztratili původní proměnné, kdybychom je ještě k něčemu potřebovali.

Rekódování

Rekódování nahrazuje kódy hodnot proměnných jinými kódy. Lze jej využít pro vytváření obecnějších kategorií u nominálních, ordinálních i kardinálních proměnných. Rekódování využíváme i pro otočení škály otázky (ve složitějších dotaznících bývají některé škály otočené, abychom udrželi respondentovu pozornost – při výsledné analýze, kdy vytváříme sumační indexy, je nutné reorientovat škály tak, aby byly všechny orientované jedním směrem).

Speciální příklad rekódování může být **kategorizace dle percentilů**. Ta nám umožní rozdělit respondenty na X stejně velkých skupin dle hodnot námi zvolené proměnné – například podle výše příjmů nebo dle věku (například bychom si chtěli respondenty rozdělit na „bohaté", střední třídu" a „chudé"). Za bohaté bychom považovali respondenty v horním kvartilu (percentil 75), za chudé respondenty v dolním kvartilu (percentil 25) a za střední třídu respondenty ve středních kvartilech.

Příklad: *V dotazníku máme otázku zjišťující ekonomickou aktivitu respondentů (otázka č. 9). Třídění v této otázce je však velmi jemné a nás zajímá pouze, zda existují rozdíly v přístupu ke vzdělávání mezi lidmi, kteří jsou zaměstnáni (do této kategorie si přiřadíme i pracující na částečný úvazek a OSVČ) a ostatními. Potřebujeme proto vytvořit novou proměnnou, která bude rozdělovat respondenty jen na dvě kategorie – na ty, kteří pracují (plus OSVČ)a na ostatní.*

Kategorie tedy bude nabývat nově dvou validních hodnot:

- *zaměstnán/a anebo OSVČ 1*
- *ostatní 2*

Hodnoty proměnné tedy rekódujeme takto:

- *1, 2, 3 → 1*
- *4, 5, 6, 7, 8 → 2*

Rekódování v Excelu

Na rekódování v Excelu existuje řada pluginů, ale je možné využít i běžný postup „Najít a nahradit" (CTRL+H). Je potřeba však dávat pozor na to, v jakém pořadí rekódujeme, aby nedošlo k dvojitému překódování.

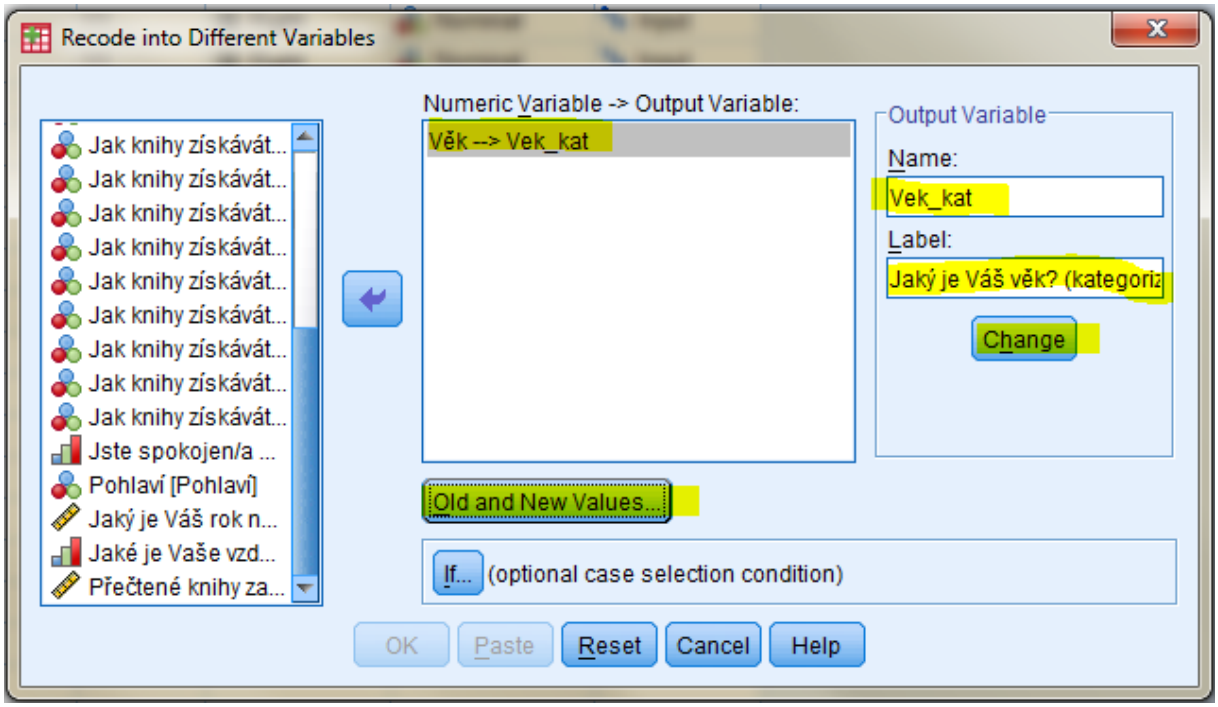
V našem příkladu by tedy byl postup následující:

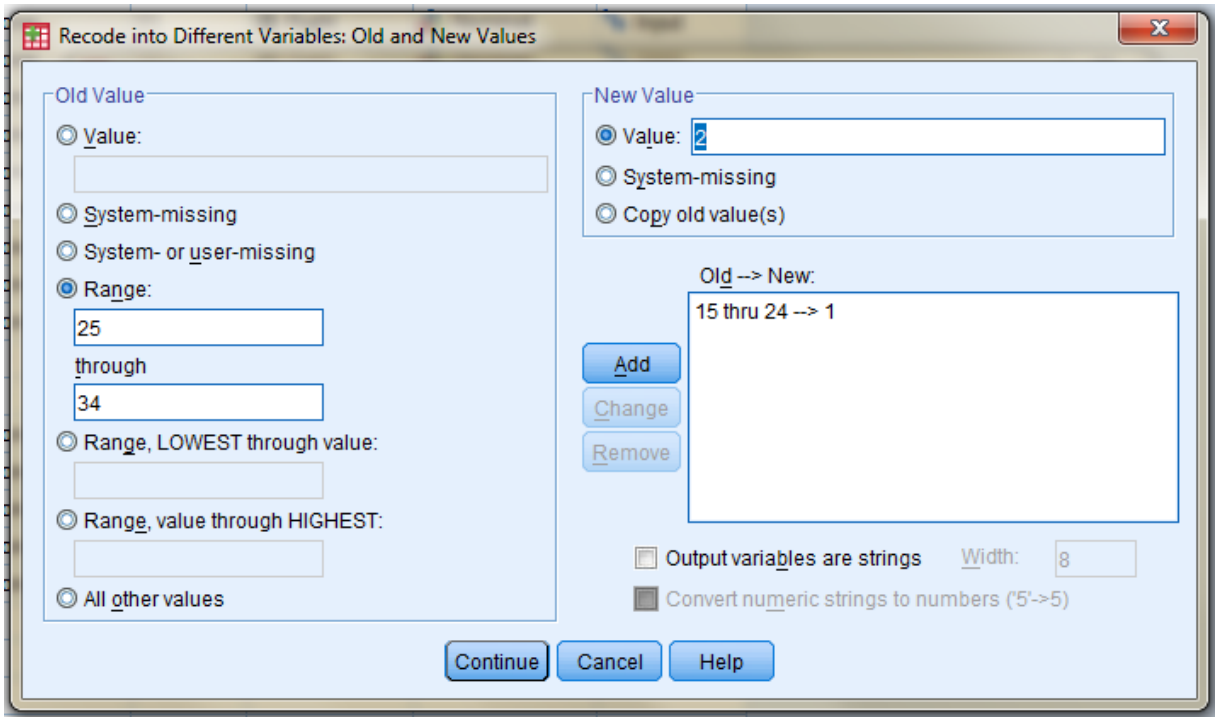
1. *Zkopírujeme si sloupec s proměnnou „ekonomická činnost" (9_ekcinnost) do nového sloupce, kde bude nová proměnná – pojmenujeme si ji třeba **9_ekcinnost2**.*
2. *Označíme si sloupec (dáváme pozor, abychom neměli označenou celou tabulku či na označení sloupce nezapomněli, v tom případě se nám překódují data z celého souboru).*
3. *Příkaz CTRL+H nám otevře dialogové okno, kde postupně budeme zadávat hodnoty k nahrazení. Vždy vybereme možnost „Nahradit vše".*
4. *Nahrazujeme v pořadí:*
    a. *Zaměstnaný/á na plný úvazek → Zaměstnán/a anebo OSVČ*
    b. *Zaměstnaný/á na částečný úvazek → Zaměstnán/a anebo OSVČ*
    c. *OSVČ → Zaměstnán/a anebo OSVČ*
    d. *Nezaměstnaný/á → Ostatní*
    e. *Na mateřské/rodičovské dovolené → Ostatní*
    f. *Nepracující důchodce → Ostatní*
    g. *V domácnosti → Ostatní*
    h. *Studující → Ostatní*
    i. *Jiné → Ostatní*

    Případně:
    a. *2 → 1*
    b. *3 → 1*
    c. *4 → 2*
    d. *5 → 2*
    e. *6 → 2*
    f. *atd… (hodnoty 1 zůstávají stejné)*

Rekódování v SPSS

Rekódování v SPSS je snadná operace. V záložce „Transform" zvolíme položku „Recode into different varialbles". V tabulce poté naklikáme jméno aoznačení nové proměnné a dále staré a nové hodnoty proměnné:





Aritmetické operace

Díky aritmetickým operacím lze snadno vytvářet nové proměnné, ze starých hodnot pomocí zadaného vzorce. Typickým příkladem je vytváření sumačních indexů – sčítání prostých nebo vážených hodnot stejných variant různých znaků.

Příklad: Pokud bychom chtěli zjišťovat celkovou spokojenost s knihovnou, kterou jsme si operacionalizovali jako sumu různých měr spokojeností (spokojenosti s personálem, spokojenosti s výběrem fondu a spokojenosti s online službami), sumační index bude tvořit průměrná míra spokojenosti v těchto dílčích oblastech (získáme ji součtem hodnot jednotlivých proměnných, který vydělíme jejich počtem).

Aritmetické operace v Excelu

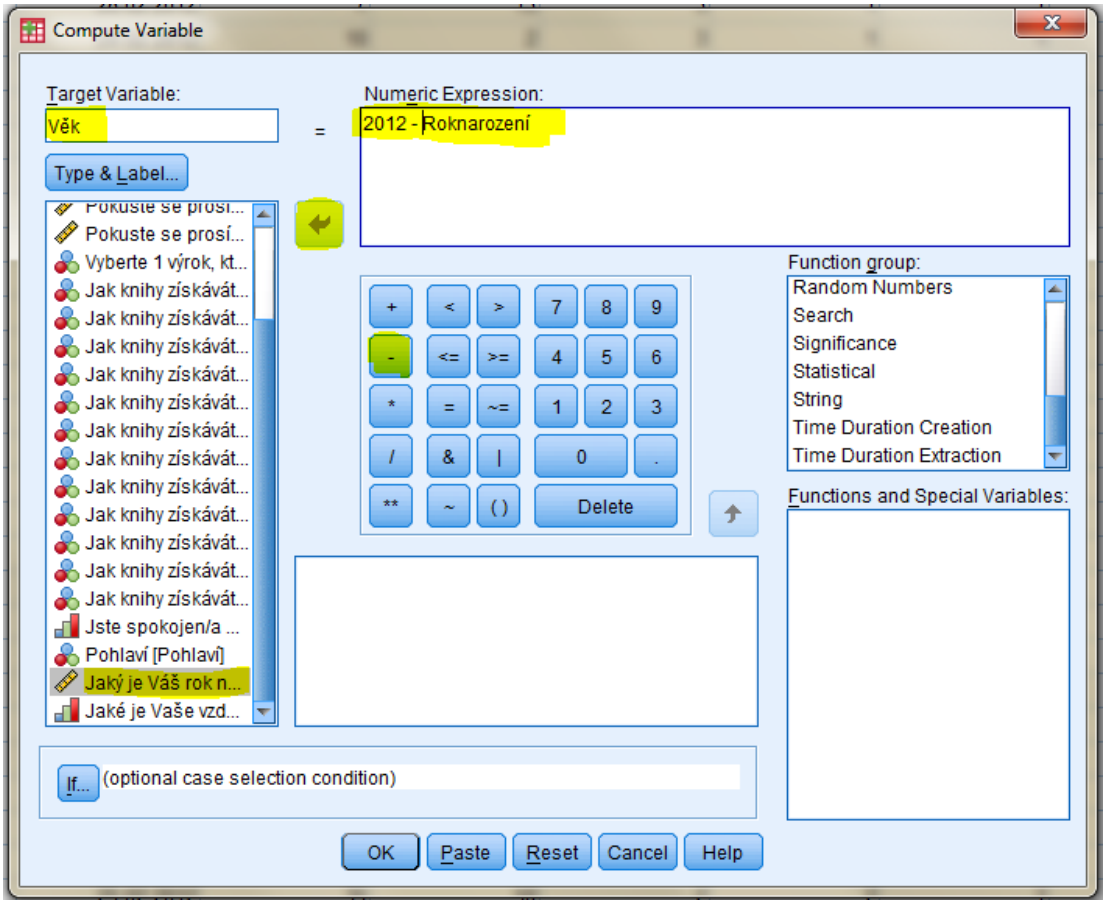
V Excelu zapisujeme vzorec přímo do tabulky.

Pokud máme např. v řádku 1 (respondent 1) proměnné spokojenost s personálem (sloupec A), spokojenost s výběrem fondu (sloupec B) a spokojenost s online službami (sloupec C), pak vzoreček pro celkovou spokojenost bude =(A1+B1+C1)/3.

Aritmetické operace v SPSS

V SPSS slouží k vytváření nových proměnných prostřednictvím aritmetických operací příkaz Compute (záložka Transform).

Níže je příklad vytvoření nové proměnné „věk" z proměnné „rok narození".



Sčítání výskytů

Sčítání výskytů je další ze způsobů jak vytvořit novou proměnnou. Někdy nás může zajímat, kolikrát respondenti například na různé otázky odpověděli „ano“.

Příklad: *V našem datasetu máme otázku, o jaké oblasti se respondenti zajímají. Nás jako výzkumníky ale může také zajímat, o kolik oblastí se průměrně lidé zajímají.*

Sčítání výskytů v Excelu

V Excelu slouží pro sčítání výskytů příkaz **COUNTIF**.

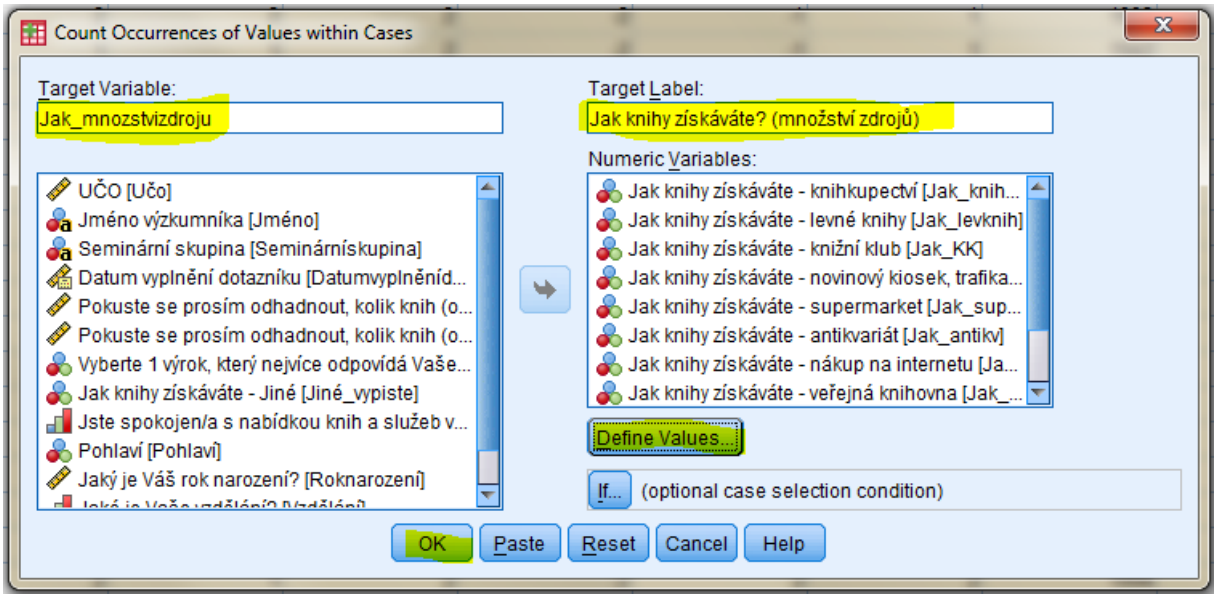
	A	B
1	**Salesperson**	**Invoice**
2	Buchanan	15,000
3	Buchanan	9,000
4	Suyama	8,000
5	Suyama	20,000
6	Buchanan	5,000
7	Dodsworth	22,500
	Formula	**Description (Result)**
	=COUNTIF(A2:A7,"Buchanan")	Number of entries for Buchanan (3)
	=COUNTIF(A2:A7,A4)	Number of entries for Suyama (2)
	=COUNTIF(B2:B7,"< 20000")	Number of invoice values less than 20,000 (4)
	=COUNTIF(B2:B7,">="&B5)	Number of invoice values greater than or equal to 20,000 (2)

Zdroj: Nápověda http://office.microsoft.com

Sčítání výskytů v SPSS

Příkaz COUNT v SPSS najdeme opět v záložce „Transform“.

Tabulka níže ukazuje výpočet počtu zdrojů, ze kterých respondenti získávají knihy (v dotazníku byla baterie otázek zaměřující se na různé zdroje knih, přičemž odpovědi byly vždy „ano“ nebo „ne“).



Seřazení položek

Vytváří novou proměnnou, kde řadí respondenty dle velikosti hodnoty proměnné. Například hledáte TOP 100 největších čtenářů či chcete najít 10 nejúspěšnějších studentů dle percentilu či 20 nejmladších respondentů. Seřazení položek je v tomto případě velmi elegantní řešení.

Příklad: V našem souboru chceme najít 10 nejstarších čtenářů. Budeme tedy pracovat se sloupečkem AS (rok narození). Do políčka u prvního respondenta (tedy v řádku 2) zapíšeme vzorec

=RANK(AS2;AS2:AS731;0)

Pokud máte dobře udělanou tabulku, Excel nám jej pravděpodobně přepíše na:

=RANK(Tabulka1[[#Tento řádek];[11_roknar]];[11_roknar];0)

Seřazení položek v Excelu

V Excelu na seřazení položek používáme příkaz **RANK**. RANK se zapisuje:

RANK(číslo;odkaz;pořadí)

- Číslo je číslo, jehož pořadí hledáte.
- Odkaz je matice nebo odkaz na seznam čísel. Nečíselné hodnoty jsou ignorovány.
- Pořadí je číslo určující, zda se budou hodnoty třídit vzestupně či sestupně.

Pokud je pořadí rovno 0 nebo není zadáno, určuje se v aplikaci Microsoft Excel pořadí čísla jako v sestupném seznamu. Pokud je pořadí jakákoliv nenulová hodnota, určuje se v aplikaci Microsoft Excel pořadí čísla jako ve vzestupném seznamu.

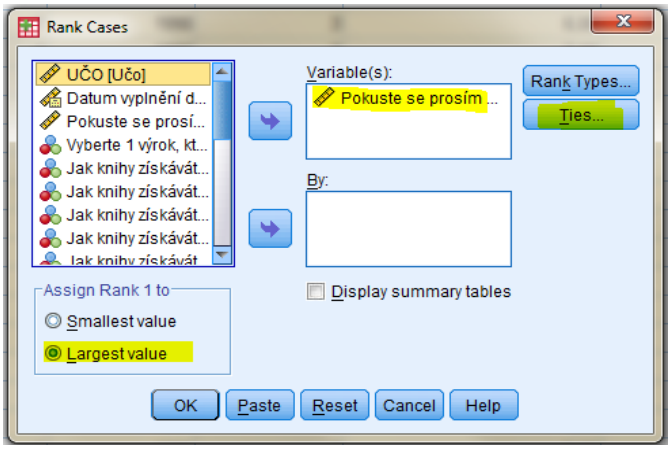
	A
1	Data
2	7
3	3,5
4	3,5
5	1
6	2

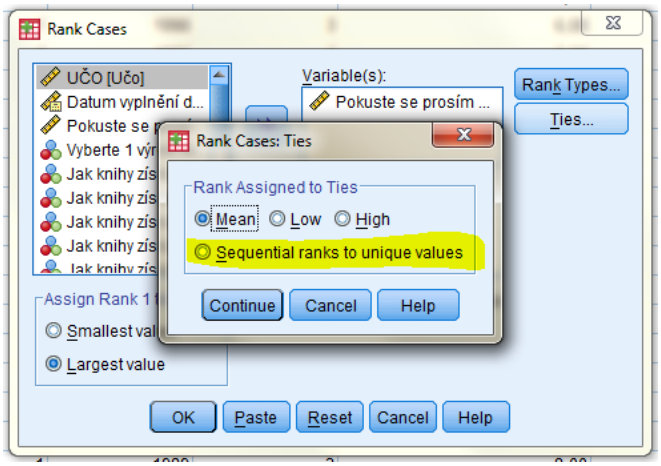
Vzorec	Popis (výsledek)
=RANK(A3;A2:A6;1)	Pořadí čísla 3,5 ve výše uvedeném seznamu (3)
=RANK(A2;A2:A6;1)	Pořadí čísla 7 ve výše uvedeném seznamu (5)

Zdroj: http://office.microsoft.com

Seřazení položek v SPSS

V Excelu na seřazení položek používá příkaz Rank Cases (opět v záložce Transform).





Metodologie pro Informační studia a knihovnictví 2

Modul 7: Třídění druhého stupně. Kontingenční tabulky

Co se dozvíte v tomto modulu?

- Co je třídění druhého stupně
- Jak vytvořit a interpretovat kontingenční tabulku

Obsah

Třídění druhého stupně

Doposud jsme se zabývali jen **popisem jednotlivých proměnných** – prováděli jsme tzv. třídění prvního stupně. Často jsou pro nás ale mnohem zajímavější data, která vzniknou tzv. **tříděním druhého stupně**, ve kterém se porovnávají dvě proměnné.

Třídění druhého stupně se používá například:

- chceme-li zjistit, zda odpovídali různě muži a ženy,
- chceme-li zjistit, zda jsou rozdíly v odpovědích respondentů dle věku,
- chceme-li zjistit, zda jsou rozdíly v odpovědích respondentů dle vzdělání,
- chceme-li zjistit, zda jsou rozdíly v odpovědích respondentů dle postojů k jinému problému.

Pro třídění druhého stupně se používá speciální tabulka četností – tzv. **kontingenční tabulka** (v Excelu funkce pivot table, v SPSS Crosstabs).

Příklad: Chceme zjistit, zda existují rozdíly v tom, jak na otázku po využívání knih ve vzdělávání odpovídali muži a ženy. Takovouto tabulku dostaneme, pokud si spočítáme pouze absolutní četnosti.

		muž	žena	celkem
1_6knihy	Je to má první volba	65	87	152
	Často	112	153	265
	Příležitostně	108	121	229
	Nikdy	39	36	75
Celkem		324	397	721

Mužů a žen bylo ale v souboru rozdílné množství!! ***Abychom mohli odpovědi porovnat, potřebujeme znát relativní četnosti!!!***

			muž	**žena**	**celkem**
1_6knihy	Je to má první volba	Absolutní četnosti	65	87	152
		Relativní četnosti	20,1%	21,9%	21,1%
	Často	Absolutní četnosti	112	153	265
		Relativní četnosti	34,6%	38,5%	36,8%
	Příležitostně	Absolutní četnosti	108	121	229
		Relativní četnosti	33,3%	30,5%	31,8%
	Nikdy	Absolutní četnosti	39	36	75
		Relativní četnosti	12,0%	9,1%	10,4%
Total		Absolutní četnosti	324	397	721
		Relativní četnosti	100,0%	100,0%	100,0%

Z tabulky můžeme vyčíst, že ***rozdíly mezi tím, jak odpovídali muž a ženy, nejsou výrazné*** *– pohybují se v jednotkách procent. Zatímco u mužů jsou knihy prví volbou v 20,1 procentech případů, u žen je to 21,9 %.*

Tvorba a interpretace kontingenční tabulky

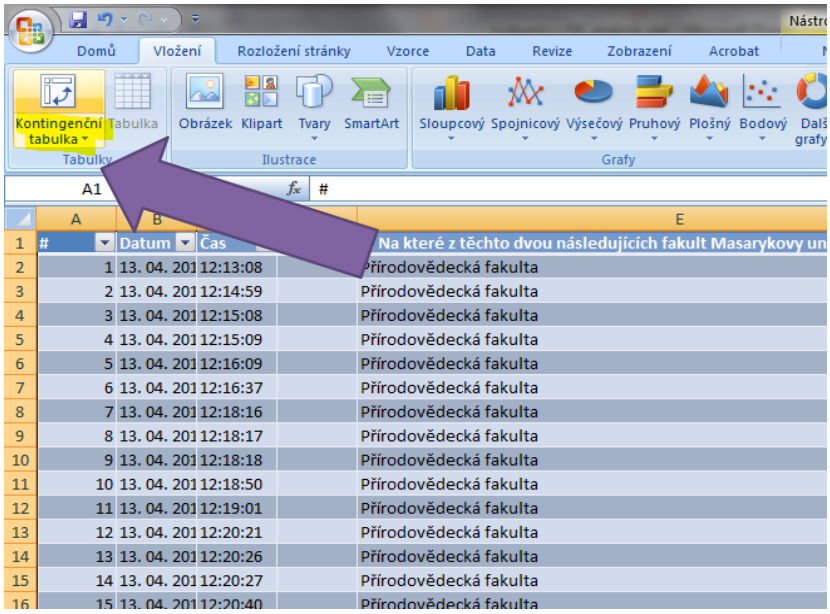
Vytvoření kontingenční tabulky je velmi jednoduché, je však třeba mít na paměti několik základních pravidel:

1. Důležité jsou pro nás **relativní četnosti**. Absolutní hodnoty jsou závislé na zastoupení jednotlivých skupin respondentů ve výběrovém vzorku.
2. Musíme určit, kterou proměnnou považujeme za **závislou** a kterou za **nezávislou**.
3. **Je-li nezávislá proměnná ve sloupcích, porovnáváme sloupcová procenta. Je-li nezávislá proměnná v řádcích, porovnáváme řádková procenta.**

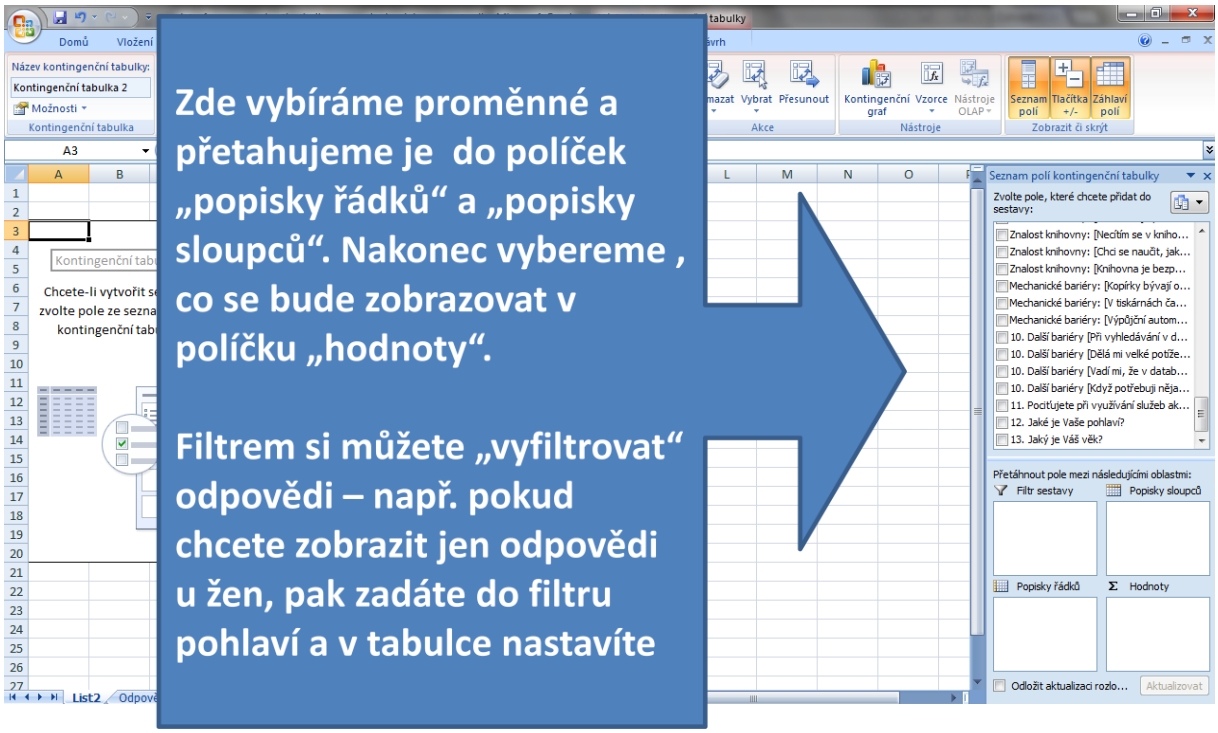
Pozor!
V bodě 3 se velmi často chybuje

Kontingenční tabulka v Excelu

V Excelu budeme opět používat funkci Pivot tables (Kontingenční tabulka).



Nejprve si musíme vybrat, jaké proměnné se budou zobrazovat v řádcích a jaké ve sloupcích. Neexistuje jednoznačný úzus (např. nezávislé proměnné v řádcích). Pokud má jedna z proměnných mnoho variant, je vhodné ji umístit do řádků (umístění do sloupců by znesnadňovalo orientaci v tabulce, případně by se tabulka musela umístit do listu s horizontální orientací.

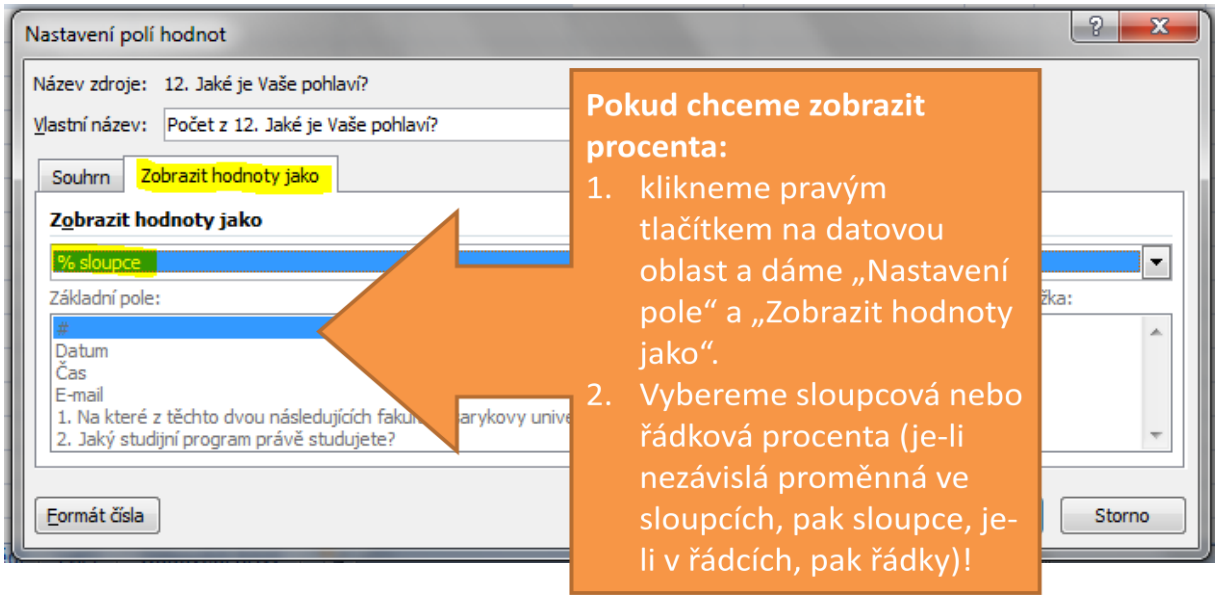


Pokud v datech zůstaly nevalidní hodnoty (missing values), je možné je pro analýzu vyřadit.

Počet z 12. Jaké je Vaše pohlaví?	Popisky sloupců		
Popisky řádků	muž	žena	Celkový součet
jednou měsíčně	110	408	518
jednou týdně	152	461	613
jednou za dva týdny	111	407	518
méně často	113	323	436
několikrát do týdne	194	511	705
nikdy	9	16	25
Celkový součet	**689**	**2126**	**2815**

Zde vybereme, jaké hodnoty proměnné se mají zobrazovat (vyřadíme nevalidní hodnoty - např. pokud nechceme zobrazovat, nezahrneme odpověď „nevím“, nebo vynechanou odpověď)

V tuto chvíli máme tabulku s absolutními četnostmi. Potřebujeme však tabulku, kde budou uvedeny i **četnosti relativní**. Kliknete pravým tlačítkem na datovou oblast a nastavíme si obrazení polí hodnot. Podle toho, kde máme nezávislou proměnnou, vybereme řádková či sloupcová procenta.



Jiný příklad kontingenční tabulky: Jak často navštěvují knihovnu prezenční a kombinovaní studenti?

Počet z 4. Jaká je forma Vašeho studia?	Popisky sloupců		
Popisky řádků	kombinovaná	prezenční	Celkový součet
několikrát do týdne	7,94%	26,26%	25,04%
jednou týdně	7,41%	22,77%	21,74%
jednou za dva týdny	12,17%	18,81%	18,37%
jednou měsíčně	32,80%	17,37%	18,40%
méně často	34,92%	14,18%	15,57%
nikdy	4,76%	0,61%	0,89%
Celkový součet	**100,00%**	**100,00%**	**100,00%**

Příklad kontingenční tabulky – zde vidíme výrazné rozdíly ve frekvenci návštěv knihovny u prezenčních a kombinovaných studentů

Kontingenční tabulka v SPSS

V SPSS jsou kontingenční tabulky v záložce Analyze → Descriptive Statistics →Crosstabs. Zobrazení řádkových a sloupcových procent se nastavuje v nabídce „Cells“. SPSS umí generovat tabulku obsahující jak absolutní, tak relativní četnosti.

Metodologie pro Informační studia a knihovnictví 2

Modul 9: Úvod do induktivní statistiky

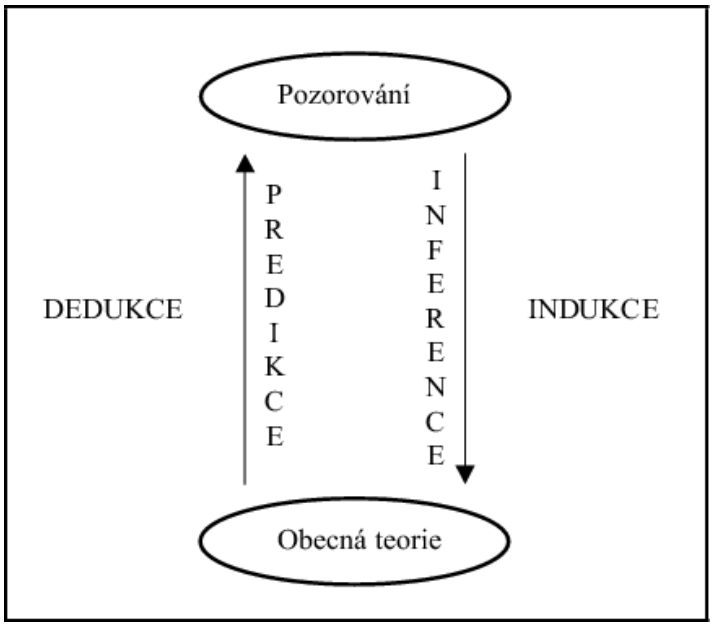
Obsah

Induktivní statistika

Dostáváme se nyní k nové kapitole statistického zpracování dat – k zobecňování na populaci. Dosud naše výpočty vypovídaly vždy jen o našich respondentech – vzorku, který neodpověděl na naše otázky. Cílem výzkumů je ale často vztáhnout výsledky na celou výzkumnou populaci, kterou vzorek zastupuje

Připomeňme si rozdíly mezi deskriptivní a induktivní statistikou:

- **Deskriptivní statistika**: popisuje rozložení četností naměřených proměnných.
- **Statistická indukce**: umožňuje zkoumat vztahy mezi proměnnými a zobecňovat výsledky na základní populaci**.**



Zdroj obrázku: http://new.euromise.org/czech/tajne/ucebnice/html/html/node3.html

Kdy můžeme zobecňovat?

Na úvod je důležité si říci, že zobecňování na populaci si nemůžeme automaticky dovolit v každém výzkumu. **Vzorek totiž musí být reprezentativní vzhledem k populaci**. Toho lze docílit různými způsoby, základním způsobem, se kterým ale počítá statistická indukce je **prostý náhodný výběr**.

Teorie statistické indukce – tedy zobecňování formou zjišťování statistické významnosti - je vyvinuta pro případy velkých reprezentativních náhodných výběrů z velkých základních souborů.

Rabušic a Soukup (2007) říkají:

„Značná část českých sociálních vědců, nemluvě o značné proporci studentů, je posedlá statistickou významností. Testy statistické signifikance v jejich povědomí (neboť tak „pochopili" smysl testováni v kurzech statistiky) slouží jako všemocné zaklinadlo. Jsou přesvědčeni, že bez testů statistických hypotéz není možné získat vědecky relevantní poznatky. Domnívají se, že tyto testy musí aplikovat na všechny

výsledky bez ohledu na to, zdali jejich data pocházejí z pravděpodobnostního (náhodného) výběru, vyčerpávajícího zjišťováni (z cenzu) nebo výběru nenáhodného (kvótního, záměrného, samovýběru). Jsou přesvědčeni, že testy významnosti jim řeknou, co je v datech důležitého, prostřednictvím nalezené statistické signifikance se snaží prokazovat těsnost vztahu dvou proměnných. Nic z toho ovšem statistická významnost neumí."

Logika statistické indukce

Přestože z úvodních řádků vyplývá, že statistickou indukci není možné aplikovat na značnou část výzkumů, které se v praxi realizují, je přesto dobré seznámit se s její logikou.

Základem statistické indukce je **testování statistických hypotéz**, přesněji řečeno zejména testování tzv. nulové hypotézy. Hypotéza je výrok o vztahu proměnných.

- **Nulová hypotéza** předpokládá stav neexistence rozdílu (tj. předpokládá stav shody) mezi proměnnými/skupinami v populaci. (Arbuthnott, 1710)
- **Alternativní hypotéza** předpokládá existenci rozdílu (na základě teorie definujeme předpoklady o rozdílech mezi jednotlivými skupinami v populaci)

Příklady nulových hypotéz:

- H0: Neexistuje rozdíl mezi rozložením proměnných ve vzorku a v populaci.
- H0: Neexistuje vztah mezi časem věnovaným internetu a pohlavím.
- Ho: Neexistuje rozdíl mezi průměrným příjmem mužů a žen zaměstnaných v knihovnách.

Příklady alternativních hypotéz:

- H0: Existuje rozdíl mezi rozložením proměnných ve vzorku a v populaci.
- H1: Neexistuje vztah mezi časem věnovaným internetu a pohlavím.

  H1a: Muži tráví na internetu více času než ženy. (Abychom si mohli dovolit formulovat takto orientovanou hypotézu, měli bychom mít podklady v předchozích výzkumech). NEBO
  H2b: Ženy tráví na internetu více času než muži. (Abychom si mohli dovolit formulovat takto orientovanou hypotézu, měli bychom mít podklady v předchozích výzkumech).

  H0: Neexistuje rozdíl mezi průměrným příjmem mužů a žen zaměstnaných v knihovnách.
  H1a: Muži zaměstnaní v knihovnách mají vyšší příjem než ženy. (Abychom si mohli dovolit formulovat takto orientovanou hypotézu, měli bychom mít podklady v předchozích výzkumech).

Pokud data neodpovídají H0, nulovou hypotézu zamítáme. Zamítnutí nulové hypotézy ovšem samo o sobě většinou nestačí k přijetí hypotézy alternativní.

Pro přijetí či zamítnutí nulové hypotézy je klíčová hladina **statistické významnosti**.

Statistická významnost je pravděpodobnost, s jakou bychom – za předpokladu platnosti nulové hypotézy – mohli obdržet data odporující nulové hypotéze. (*Soukup 2010)*

➔ Je-li statistická významnost nízká, nulová hypotéza nejspíš neplatí.

Zlaté pravidlo pro induktivní statistiku:

- Vysoká hodnota testu statistické významnosti (tj. $\alpha > 0{,}05$) → rozdíl není statisticky významný → **držíme nulovou hypotézu.**
- Nízká hodnota testu statistické významnosti (tj. $\alpha \leq 0{,}05$) → rozdíl je statisticky významný → **zamítáme nulovou hypotézu**.

Princip většiny statistických testů spočívá v tom, že se výsledky naměřených hodnot porovnávají s teoretickým modelem jejich rozložení – z něj jsou odvozeny tzv. kritické hodnoty testu (Reichel 2009). Pro různé druhy hypotéz existuje řada **testovacích kritérií**.

Proč nelze jednoduše zobecnit ze vzorku na populaci aneb zobecňování průměrů

Představte si, že zkoumáme populaci magisterských studentů knihovnictví. Chceme vidět, jak se měnil nějaký konkrétní ukazatel – třeba jejich váhu v kilogramech. Dejme tomu, že je studentů celkem 200. Náš vzorek je 15 studentů (víme už, že takový vzorek by byl velmi malý, ale pro tento příklad si jej ponechme).

Populační průměr sledované vlastnosti je 69,63. Pokaždé, kdy náhodně vybereme nějaký vzorek 15 studentů, dostaneme poněkud odlišné výsledky:

Číslo měření	***Průměr***	***St. odchylka***	***Minimum***	***Medián***	***Maximum***	***Rozpětí***
1.	*66,12*	*9,21*	*47,2*	*65*	*87*	*39,8*
2.	*73,3*	*12,48*	*52,4*	*71,1*	*101,1*	*48,7*
3.	*68,67*	*10,78*	*54*	*69,1*	*85,4*	*31,4*
4.	*69,95*	*10,57*	*54,5*	*68*	*87,8*	*33,3*

Takto bychom mohli pokračovat a při každém výběru bychom dostali poněkud jiné výsledky. Nyní vidíme, že z jednoho měření nelze jednoduše zobecnit průměr – každý výběr je zatížen tzv. **výběrovou chybou**.

Výběrová chyba je chyba, která vyplývá z faktu, že neměříme populaci, ale vzorek. Velikost výběrové chyby vychází především z distribuce vlastnosti v populaci. Pokud je populace homogenní vzhledem k vybranému kritériu, výběrová chyba bude pravděpodobně menší. Výběrová chyba také bude klesat s velikostí vzorku. Vzorek 50 studentů bude mít pravděpodobně nižší výběrovou chybu než vzorek 15 studentů.

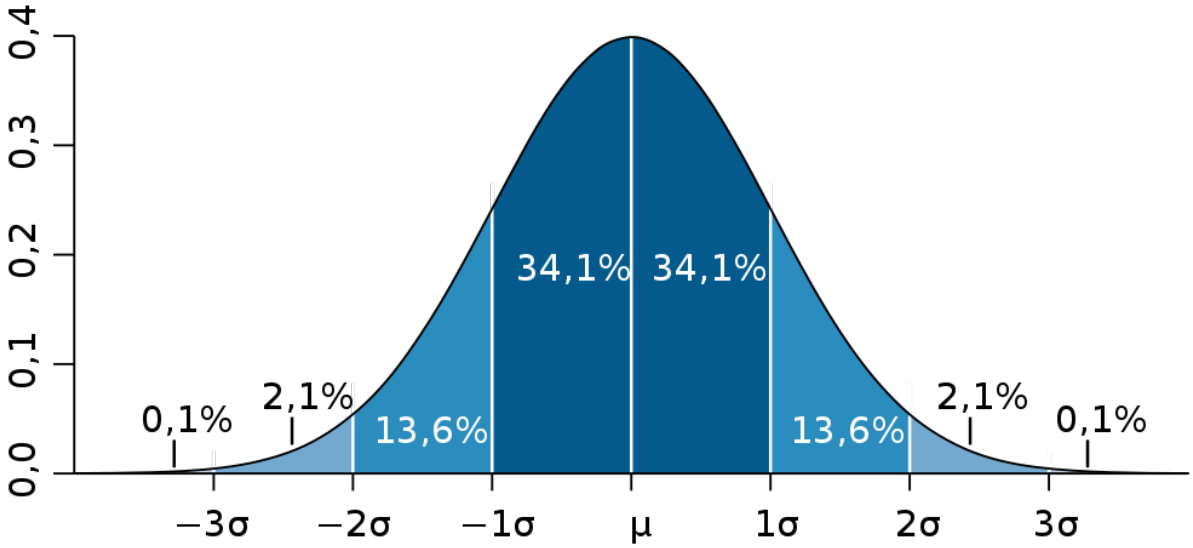
Jak se vypořádat s výběrovou chybou? Musíme pochopit, že ze vzorku nemůžeme se 100%pravděpodobností usuzovat na výsledek (průměr) celé populace. O výsledku tedy můžeme hovořit jen jako o odhadu v rámci určitého intervalu a s určitou mírou jistoty.

Je jasné, že čím nižší míra jistoty, tím menší může být interval, ve kterém se spolehlivě průměr nachází v populaci, a naopak: pokud chceme mít vysokou míru jistoty, interval bude větší.

Nejčastěji volíme **interval spolehlivosti 95 % nebo 99 %**. To znamená, že o naměřeném výsledku můžeme s 95% (respektive 99%) spolehlivostí tvrdit, že se nachází v daném intervalu.

K výpočtu horní a spodní hranice interval spolehlivosti nám pomůže znalost velikosti směrodatné odchylky.

Na obrázku vidíme normální rozložení hodnot v populaci. V intervalu jedné směrodatné odchylky od průměru na obou stranách leží 68,2 % všech naměřených hodnot. V intervalu dvou směrodatných odchylek už leží 95 % a v intervalu tří směrodatných odchylek leží 99 % naměřených hodnot.



Výpočet intervalu spolehlivost v Excelu

V Excelu pro výpočet intervalu spolehlivosti používáme příkaz CONFIDENCE. Podrobný popis použití příkazu najdete zde.

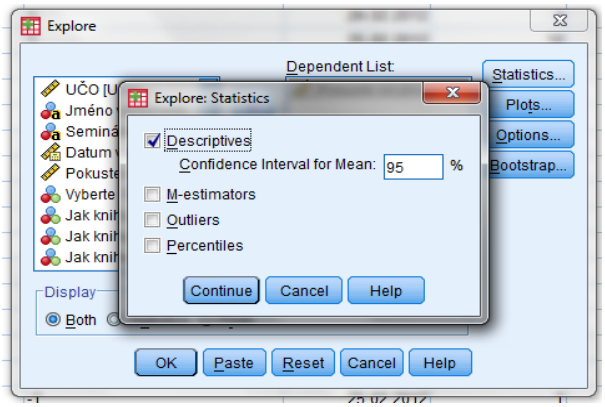
K výpočtu potřebujeme znát:

- **koeficient spolehlivosti** (0,05 pro 95% interval spolehlivosti a 0,01 pro 99% interval spolehlivosti),
- **směrodatnou odchylku v populaci**,
- **velikost výběrového souboru.**

V praxi ale většinou neznáme hodnoty průměru v populaci či výši směrodatné odchylky. Proto byly vyvinuty postupy realizovatelné při využití standardní odchylky naměřeného průměru – tzv. T-rozložení a T-test.

Výpočet intervalu spolehlivost v SPSS

V SPSS používáme záložku Explore, kde si na kartě Statistics upravíme velikost intervalu spolehlivosti:



SPSS vrátí informace o horní a spodní hranici intervalu spolehlivosti.

Hypotézy o shodě dvou populačních průměrů

Pro vyhodnocování hypotézy o shodě dvou průměrů používáme tzv. T-test.

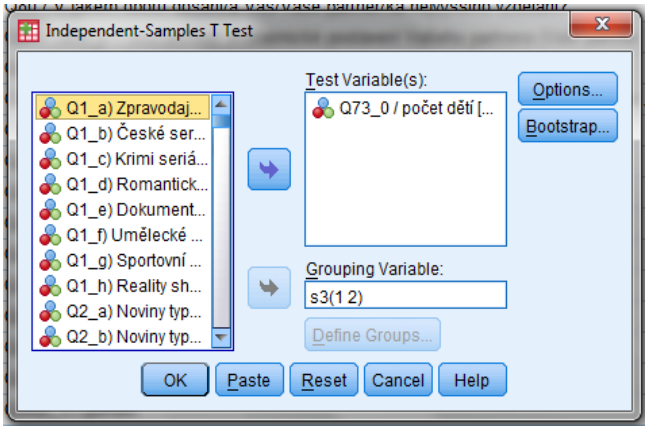
- **Studentův t-test** (William Gosset)
  - *směrodatná odchylka (s), která sama podléhá variabilitě výběru, již nemusí být spolehlivým odhadem populační směrodatné odchylky* (zdroj)
  - Pro nás relevantní: Independent Samples T-test

Např. zkoumáme vztah mezi pohlavím a počtem dětí (v populaci třicátníků) – příklad pracuje s daty z výzkumu Distinkce a hodnoty 2008 (viz Studijní materiály v ISu).

Nulová a alternativní hypotéza:

- *H0: Neexistuje rozdíl mezi počtem dětí u skupin podle pohlaví.*
- *Ha: Existuje rozdíl mezi počtem dětí u skupin podle pohlaví.*

Postup v SPSS: Analyze – Compare Means – Independent Samples T-test



Podíváme se, jaké rozdíly jsme naměřili na vzorku:

Group Statistics

	pohlaví	N	Mean	Std. Deviation	Std. Error Mean
Q73_0 / počet dětí	muž	492	,78	,915	,041
	žena	529	1,13	,939	,041

Existují rozdíly i v populaci?

Interpretujeme test ve dvou krocích:

1. *podíváme se na výsledky F testu o shodě variací*
   - *Signifikance u F > 0,05 → použijeme T-testu pro případ EQUAL VARIANCES ASSUMED*
   - *Signifikance u F < 0,05 → použijeme T-testu pro případ EQUAL VARIANCES NOT ASSUMED*
2. *v příslušném sloupci čteme významnost*

Je-li menší než 0,05, nulovou hypotézu o shodě populačních průměrů lze zamítnout – rozdíl pravděpodobně existuje i v populaci

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Q73_0 / počet dětí	Equal variances assumed	,531	,467	-6,091	1019	,000	-,354	,058	-,468	-,240
	Equal variances not assumed			-6,097	1016,790	,000	-,354	,058	-,468	-,240

Porovnávání více populačních průměrů

Opět si vše ukážeme na příkladu z výzkumu Distinkce a hodnoty 2008.

Např. zkoumáme vztah mezi vzděláním a počtem dětí

Nulová a alternativní hypotéza:

- *H0: Neexistuje rozdíl mezi počtem dětí u jednotlivých vzdělanostních skupin.*
- *Ha: Existuje rozdíl mezi počtem dětí u jednotlivých vzdělanostních skupin.*

*Nejprve si zjistíme rozdíly v **naměřených** průměrech: Analyze – **Compare Means***

Porovnání průměrů ukazuje, že v naměřených hodnotách jsou rozdíly. Jsou však rozdíly i v populaci?

Q73_0 / počet dětí

Q27 / Jaké je Vaše ...	Mean	N	Std. Deviation
základní, bez vyučení	1,19	48	1,214
střední s vyučením	1,07	316	,978
střední bez maturity	,80	104	,885
střední s maturitou	,97	386	,921
vyšší odborné (pomaturitní studium)	,89	36	,854
vysokoškolské bakalářské	,75	36	,806
vysokoškolské magisterské, inženýrské	,71	86	,824
vysokoškolské doktorské	,89	9	1,054
Total	,96	1021	,944

- *1. krok: Analyze – **One way** ANOVA*
- *Options: Descriptives*

Descriptives

Q73_0 / počet dětí

	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
základní, bez vyučení	48	1,19	1,214	,175	,83	1,54	0	4
střední s vyučením	316	1,07	,978	,055	,96	1,18	0	4
střední bez maturity	104	,80	,885	,087	,63	,97	0	3
střední s maturitou	386	,97	,921	,047	,87	1,06	0	3
vyšší odborné (pomaturitní studium)	36	,89	,854	,142	,60	1,18	0	3
vysokoškolské bakalářské	36	,75	,806	,134	,48	1,02	0	2
vysokoškolské magisterské, inženýrské	86	,71	,824	,089	,53	,89	0	3
vysokoškolské doktorské	9	,89	1,054	,351	,08	1,70	0	3
Total	1021	,96	,944	,030	,90	1,02	0	4

- *2. krok: **statistika F** a její **signifikance***

ANOVA

Q73_0 / počet dětí

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	16,466	7	2,352	2,669	,010
Within Groups	892,887	1013	,881		
Total	909,354	1020			

Podíl variability mezi skupinami (between groups) a variability uvnitř skupin (within groups), konkrétně jejich průměrů součtu druhých mocnin směrodatných odchylek. Pokud platí nulová hypotéza, že rozdíly mezi průměry jsou nulové, musí být obě průměrné hodnoty druhých mocnin podobné a jejich vzájemný poměr (F) tedy musí být blízko 1.

Hodnota je signifikantní (menší než 0,05). Pravděpodobnost podržet nulovou hypotézu je nízká (0,01) → **zamítáme** (tj. průměry v populaci nejsou stejné)

- *3. krok: Chceme vědět, mezi kterými skupinami **statisticky významný rozdíl** existuje*

Post Hoc Tests

Multiple Comparisons

Q73_0 / počet dětí
Bonferroni

(I) Q27 / Jaké je Vaše nejvyšší dosažené vzdělání?	(J) Q27 / Jaké je Vaše nejvyšší dosažené vzdělání?	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
základní, bez vyučení	střední s vyučením	,115	,145	1,000	-,34	,57
	střední bez maturity	,389	,164	,494	-,12	,90
	střední s maturitou	,221	,144	1,000	-,23	,67
	vyšší odborné (pomaturitní studium)	,299	,207	1,000	-,35	,95
	vysokoškolské bakalářské	,438	,207	,974	-,21	1,09
	vysokoškolské magisterské, inženýrské	,478	,169	,134	-,05	1,01
	vysokoškolské doktorské	,299	,341	1,000	-,77	1,37
střední s vyučením	základní, bez vyučení	-,115	,145	1,000	-,57	,34
	střední bez maturity	,275	,106	,274	-,06	,61
	střední s maturitou	,106	,071	1,000	-,12	,33
	vyšší odborné (pomaturitní studium)	,184	,165	1,000	-,33	,70
	vysokoškolské bakalářské	,323	,165	1,000	-,19	,84
	vysokoškolské magisterské, inženýrské	,363*	,114	,042	,01	,72
	vysokoškolské doktorské	,184	,317	1,000	-,81	1,18
střední bez maturity	základní, bez vyučení	-,389	,164	,494	-,90	,12
	střední s vyučením	-,275	,106	,274	-,61	,06

SPSS potvrdilo statisticky významný rozdíl pouze mezi skupinou SŠ s vyučením a VŠ mgr/ing. U ostatních skupin nemůžeme s jistotou říci, že rozdíl existuje i v populaci

Zobecňování výsledků třídění druhého stupně (kontingenčních tabulek)

Druhým příkladem zobecňování z naměřených hodnot na populaci je zobecňování výsledků třídění druhého stupně kategorizovaných dat.

Příklad: Chceme vědět, jak se liší frekvence čtení u skupin podle vzdělání. Formulujeme nulovou a alternativní hypotézu:

- ➔ *H0: Neexistuje rozdíl ve frekvenci čtení mezi skupinami třicátníků s různým vzděláním.*
- ➔ *Ha: Existuje rozdíl ve frekvenci čtení mezi skupinami třicátníků s různým vzděláním.*

Uděláme si kontingenční tabulku (už ji umíme od modulu 7):

Jaké je vaše vzdělání? * Četbě knih (rec) Crosstabulation

			Četbě knih (rec)				
			Několikrát týdně nebo denně	Jednou za měsíc až jednou týdně	Několikrát za rok	Vůbec ne	Total
Jaké je vaše vzdělání?	ZŠ (i nedokončené)	Count	3	11	10	23	47
		% within Jaké je vaše vzdělání?	6,4%	23,4%	21,3%	48,9%	100,0%
	SŠ/VOŠ	Count	172	292	168	200	832
		% within Jaké je vaše vzdělání?	20,7%	35,1%	20,2%	24,0%	100,0%
	VŠ	Count	57	44	17	12	130
		% within Jaké je vaše vzdělání?	43,8%	33,8%	13,1%	9,2%	100,0%
Total		Count	232	347	195	235	1009
		% within Jaké je vaše vzdělání?	23,0%	34,4%	19,3%	23,3%	100,0%

Vidíme poměrně zajímavé rozdíly! Můžeme je zobecnit?

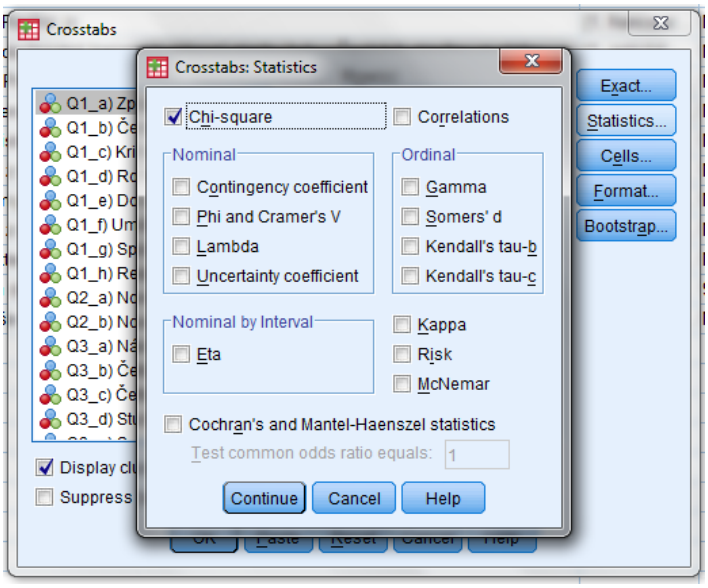
Pro zobecňování rozdílů u kategorizovaných proměnných se používá jako testovací kritérium tzv. test nezávislosti chí kvadrát ($\chi 2$).

Chí-kvadrát je založený na srovnávání naměřených a očekávaných proměnných

- **Očekávaná četnost**: počet jednotek, který by do dané kategorie spadl při náhodném rozložení
- **Naměřená četnost:** počet jednotek, které jsme v dané kategorii ve vzorku naměřili
- **Reziduál**: rozdíl mezi OČ a NČ
- **Adjustované reziduály**: koeficient determinace (AR mají přibližně normální rozložení s průměrem 0 a standardní odchylkou 1)

Chí kvadrát v SPSS

Chí-kvadrát – Analyze – Crosstabs: Statistics



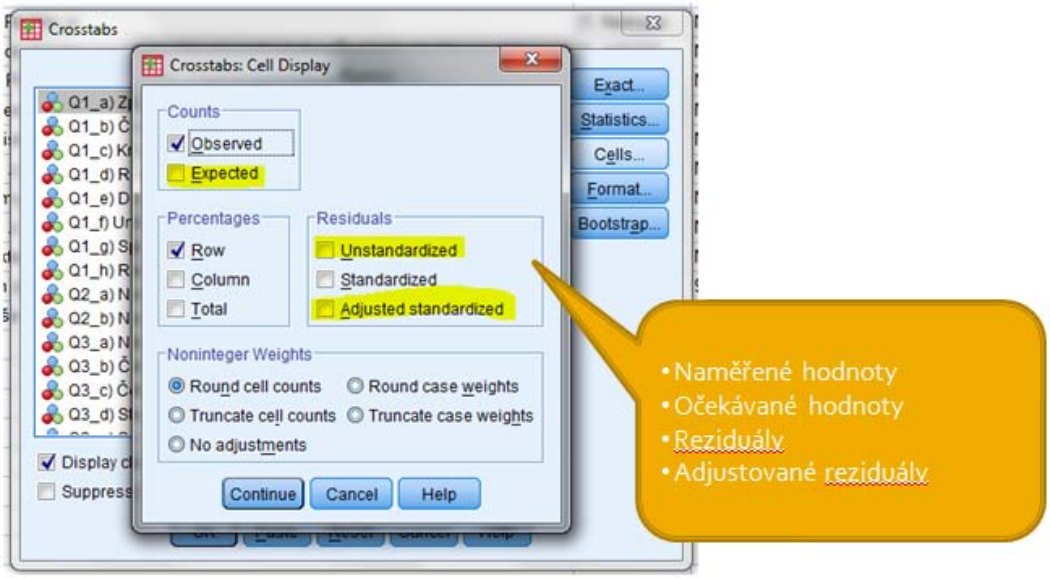
Chi-Square Tests

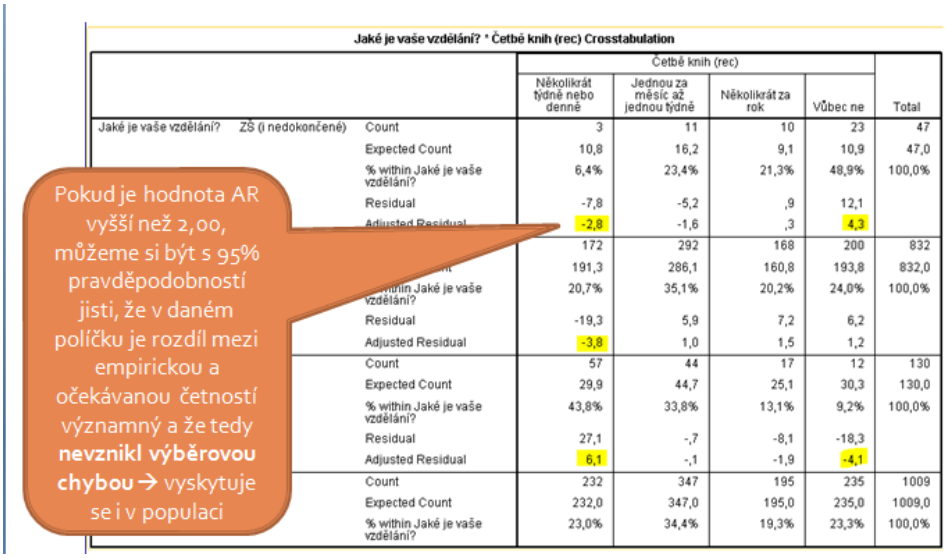
	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	61,503[a]	6	,000
Likelihood Ratio	59,263	6	,000
Linear-by-Linear Association	54,887	1	,000
N of Valid Cases	1009		

a. 0 cells (,0%) have expected count less than 5. The minimum expected count is 9,08.

Hodnota významnosti o → **zamítáme hypotézu o neexistenci rozdílu v populaci**

Pozor! Chí-kvadrát se dá použít jen pokud více než 20 % políček má očekávanou četnost menší než 5 a minimální očekávaná četnost nesmí být menší než 1





Jaké je vaše vzdělání? * Četbě knih (rec) Crosstabulation

		Četbě knih (rec)				
		Několikrát týdně nebo denně	Jednou za měsíc až jednou týdně	Několikrát za rok	Vůbec ne	Total
Jaké je vaše vzdělání? ZŠ (i nedokončené)	Count	3	11	10	23	47
	Expected Count	10,8	16,2	9,1	10,9	47,0
	% within Jaké je vaše vzdělání?	6,4%	23,4%	21,3%	48,9%	100,0%
	Residual	-7,8	-5,2	,9	12,1	
	Adjusted Residual	-2,8	-1,6	,3	4,3	
	Count	172	292	168	200	832
	Expected Count	191,3	286,1	160,8	193,8	832,0
	% within Jaké je vaše vzdělání?	20,7%	35,1%	20,2%	24,0%	100,0%
	Residual	-19,3	5,9	7,2	6,2	
	Adjusted Residual	-3,8	1,0	1,5	1,2	
	Count	57	44	17	12	130
	Expected Count	29,9	44,7	25,1	30,3	130,0
	% within Jaké je vaše vzdělání?	43,8%	33,8%	13,1%	9,2%	100,0%
	Residual	27,1	-,7	-8,1	-18,3	
	Adjusted Residual	6,1	-,1	-1,9	-4,1	
	Count	232	347	195	235	1009
	Expected Count	232,0	347,0	195,0	235,0	1009,0
	% within Jaké je vaše vzdělání?	23,0%	34,4%	19,3%	23,3%	100,0%

Pokud je hodnota AR vyšší než 2,00, můžeme si být s 95% pravděpodobností jisti, že v daném políčku je rozdíl mezi empirickou a očekávanou četností významný a že tedy **nevznikl výběrovou chybou** → vyskytuje se i v populaci

Literatura:

Reichel, J. 2009. Kapitoly metodologie sociálních výzkumů. Praha: Grada.

Soukup, P. 2010. „Nesprávné užívání statistické významnosti a jejich možná řešení.“ Data a výzkum – SDA Info 4(2): 77–104.

SOUKUP, Petr - RABUŠIC, Ladislav. Několik poznámek k jedné obsesi českých sociálních věd - statistické významnosti. Sociologický časopis. 2007, roč. 43, č. 2, s. 379-395. ISSN 0038-0288.

Metodologie pro Informační studia a knihovnictví 2

Modul 10: Vizualizace výsledků. Nástroje pro statistickou analýzu a vizualizace

Obsah

1. Nástroje pro statistickou analýzu

V předchozích hodinách jsme pracovali s Excelem, případně s SPSS. Nástrojů pro statistickou analýzu je ale celá řada:

1.1.SPSS

SPSS jsme si již představili. Jedná se o **jeden z nejrobustnějších nástrojů pro statistickou analýzu**. Pro studenty FF MU je dostupná akademická licence MU.

- Webové stránky: http://www-01.ibm.com/software/cz/analytics/spss/
- Program lze stáhnout po přihlášení v INETu.

1.2.Statistica

Dalším komerčním programem pro statistickou analýzu je **Statistica**. Pro studenty MU je také k dispozici akademická licence.

- Webové stránky: http://www.statsoft.cz/
- Program lze stáhnout po přihlášení v INETu.

Existuje i řada dalších komerčních programů pro statistiku (Stata, MATLAB…). Pojďme se ale podívat na nástroje, které lze využívat jako **open source**.

1.3.PSPP

PSPP je **open source alternativa k SPSS**. Rozhraní je velmi podobné SPSS, program umožňuje provádět deskriptivní statistiku, T-testy, lineární regresii atd.

- Webové stránky: http://www.gnu.org/software/pspp

1.4.R

R je programovací jazyk a prostředí pro statistické zpracování dat. V současnosti je R asi **nejrozšířenější statistický open source**.

- Webové stránky: http://www.r-project.org

1.5.SOFA Statistics

SOFA je stejně jako R open source program, ale na rozdíl od R, jehož největší výhodou je robustnost a využitelnost , SOFA klade důraz na uživatelskou přívětivost.

- Webové stránky: http://www.sofastatistics.com

2. Nástroje pro vytváření infografik

2.1.Infogr.am

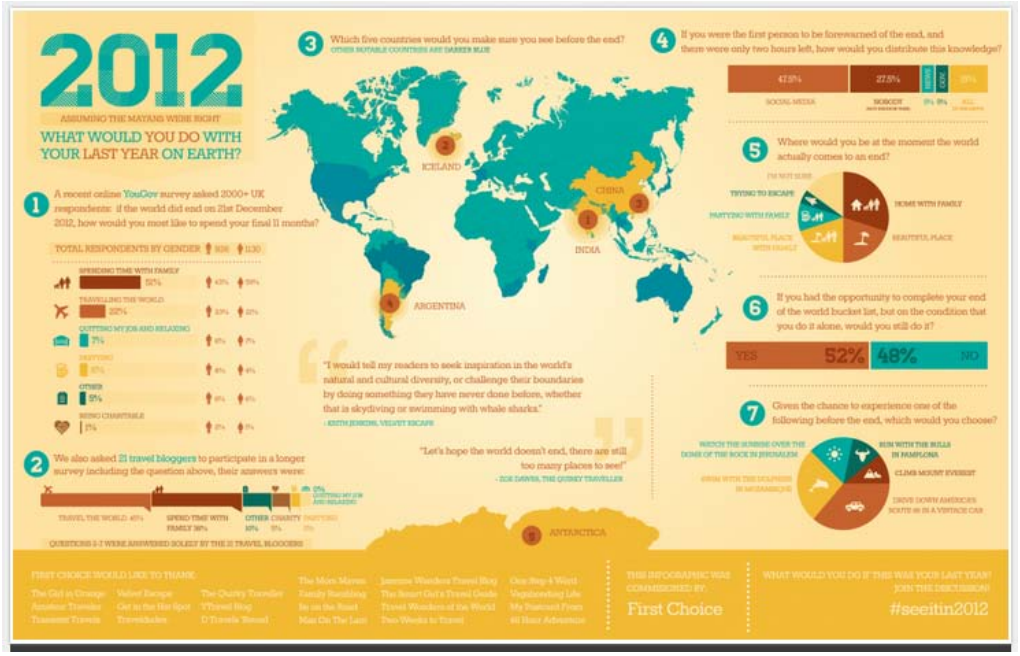
Pravděpodobně nejúspěšnější aplikace pro vytváření infografik. Uživatelsky přívětivá, se skvělou grafikou, výsledky lze snadno stáhnout i sdílet.

- Webová stránka: http://infogr.am

2.2.Visual.ly

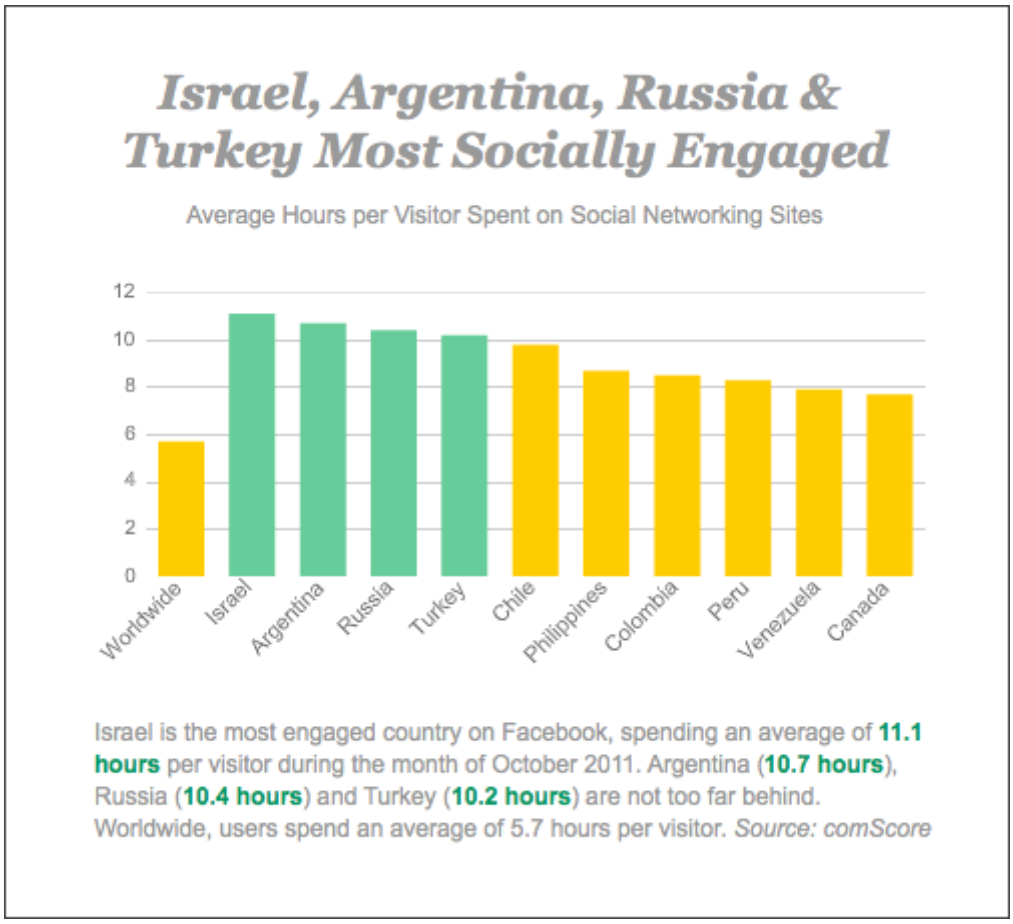
Aplikace pro vytváření vizualizací a infografik.

- http://visual.ly



2.3.iCharts

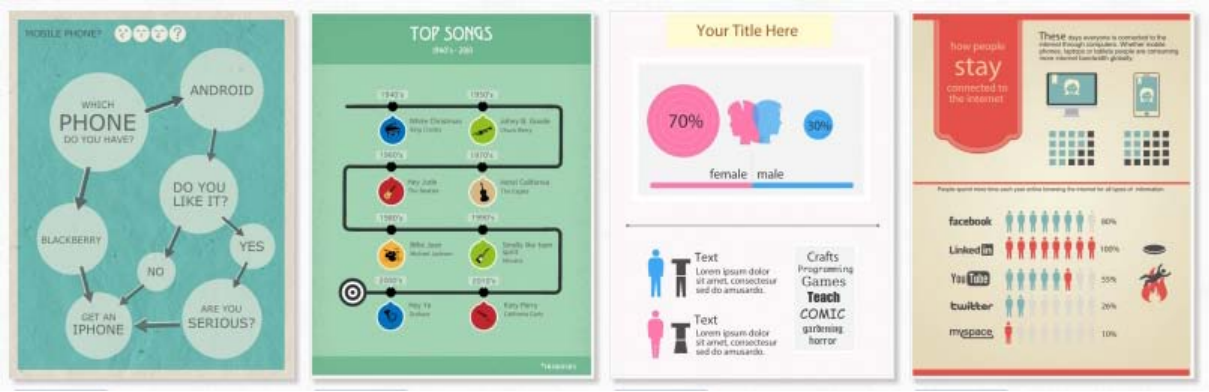
Různé typy grafů, podpora storytellingu, snadné sdílení grafů.



2.4.Easel.ly

Aplikace založená na širokém výběru vizuálních témat.

Webové stránky: http://www.easel.ly



3. Galerie infografik pro inspiraci

- http://visual.ly/
- http://visualizing.org
- http://www.informationisbeautiful.net/

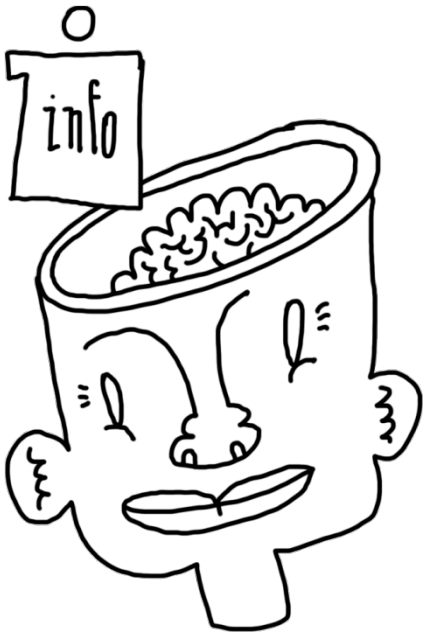
Metodologie pro Informační studia a knihovnictví 2

Modul 10: Analýza a interpretace v kvalitativním výzkumu

Obsah

1. Kvalitativní výzkum a jeho charakteristiky

V tomto modulu si představíme některé obecné zásady pro interpretaci a vyhodnocování kvalitativních výzkumů. Než se dostaneme k tématu vyhodnocování, je dobré si připomenout charakteristiky kvalitativního zkoumání.



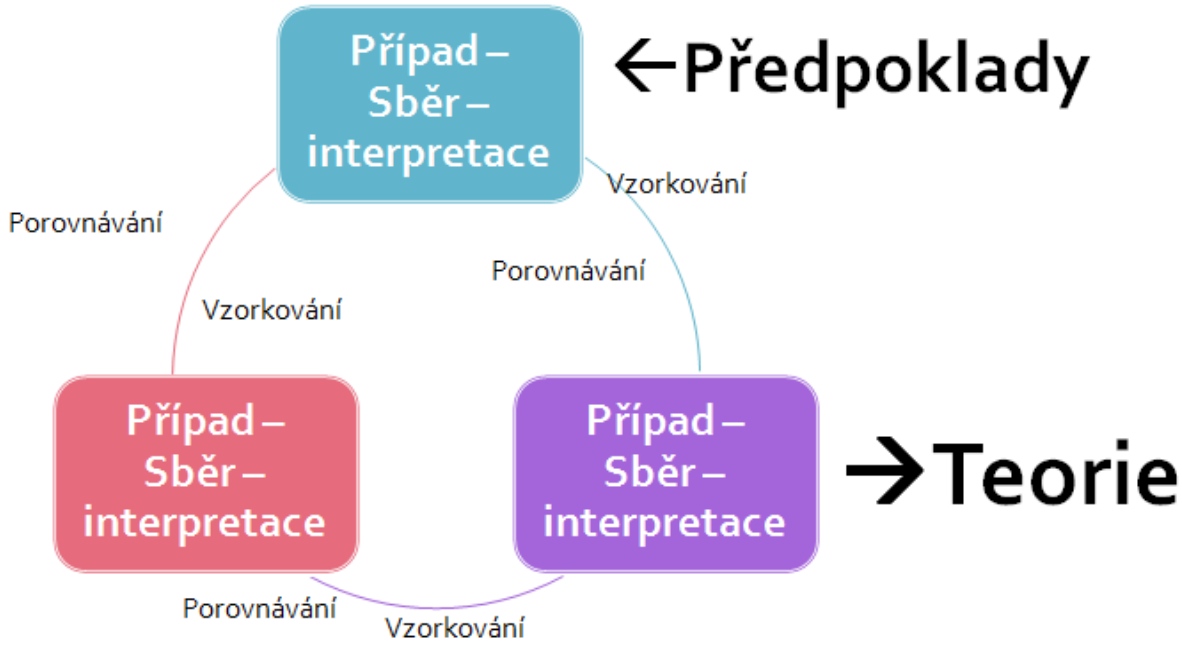
Základní charakteristiky kvalitativních výzkumů podle Creswella (2009):

- **Přirozené prostředí**. Výzkumníci zkoumají lidi v prostředí, ve kterém informanti zažívají zkoumané problémy/jevy. Interakce v přirozeném prostředí probíhají většinou tváří v tvář, participaci nejsou přenášeni nikam do laboratoře, prostředníkem mezi výzkumníkem a participantem není ani žádný nástroj (např. dotazník).
- **Hlavním výzkumným nástrojem je sám výzkumník.** Výzkumníci sbírají data o chování participantů prostřednictvím pozorování chování, studium dokumentů, hloubkové rozhovory. Mohou postupovat na základě určitého předem připraveného protokolu, ale jsou to výzkumníci, kdo řídí výzkum, nikoliv protokol.
- **Více zdrojů dat.** Kvalitativní výzkumníci využívají kombinace metod (studium dokumentů, rozhovory, pozorování…). Data jsou vyhodnocována napříč metodami, tvoří se kategorie.
- **Induktivní logika výzkumu.** Kvalitativní výzkumníci tvoří kategorie a kódy „zezdola nahoru" – organizováním a kategorizováním pozorovaných jevů (kategorie nejsou předem dány, ale vyplývají ze zjištění výzkumníků).
- **Významy určují účastníci výzkumu.** Výzkumníkovi jde o to pochopit, jaké výzkumy věcem a jevům přikládají účastníci výzkumu, nikoliv o přiřazení „objektivních" významů či významů, které najde výzkumník v literatuře.
- **Plán výzkumu se vyvíjí v průběhu zkoumání.** Na rozdíl od kvantitativního šetření, v kvalitativním výzkumu není plán sběru a analýzy dat předem znám. Rozhodnutí o

metodách sběru dat, počtu participantů atd. se může změnit v průběhu výzkumu ve chvíli, kdy narazíme na nové informace.

- **Role teorie.** Creswell zmiňuje „teoretické brýle“, kterými se výzkumníci dívají na realitu. Teorie nás vede k tomu, co pozorovat (např. generové, kulturní či třídní rozdíly).
- **Výzkum je interpretativní.** Spíše než o analýze dat hovoříme v kvalitativním výzkumu o interpretaci – interpretaci významů vytvářejí jak účastníci výzkumu, tak výzkumníci i čtenáři výzkumné zprávy Všichni do svých interpretací vnášejí svůj pohled založený na „backgroundu“ – kultuře, osobní historii atd.
- **Holistický přístup.** V kvalitativním výzkumu jde o vytvoření komplexního obrazu studovaného problému. Jde se do hloubky, problém je zkoumán z mnoha perspektiv.

Připomeňme si i schéma procesu kvalitativního výzkumu:



Na jednoduchém schématu můžeme vidět, že kvalitativní výzkum je spíše cyklický proces, ve kterém se neustále rozhodujeme o vzorku, metodách sběru dat, počtu zkoumaných případů.

2. Strategie kvalitativního výzkumu

Různí autoři hovoří o různých strategiích, které se uplatňují v kvalitativním výzkumu. Klasické dělení, které vytvořil Tesch (1990), hovoří o několika desítkách druhů kvalitativních výzkumů:

- action research
- ethnographic kontent analysis
- interpretive interactionism
- case study
- interpretive human studies
- clinical research
- ethnography
- life history study
- cognitive antropology
- ethnography of communication
- naturalistic inquiry
- collaborative enquiry
- oral history
- content analysis
- ethnomethodology
- panel research
- dialogical research
- ethnoscience
- participant observation
- conversation analysis
- experiential psychology
- participative research
- Delphi study
- field study
- phenomenography
- descriptive research
- focus group research
- phenomenology
- direct research
- grounded theory
- qualitative evaluation
- discourse analysis
- hermeneutics
- structural ethnography
- document study
- heuristic research
- symbolic interactionism
- ecological psychology
- holistic enthnography
- transcendental realism
- educational connoisseurship and criticism
- imaginal psychology
- intensive evaluation
- transformative research
- educational ethnography

Hrubší dělení Creswella (2007) zmiňuje pět základních přístupů ke kvalitativnímu výzkumu:

1. **narativní přístup,**
2. **fenomenologii,**
3. **etnografický přístup,**
4. **případové studie,**
5. **zakotvenou teorii.**

(Připomeňte si prezentaci z minulého semestru, kdy jsme si na příkladu teorie úzkosti z knihoven od Sharon Bostick demonstrovali rozdíly mezi jednotlivými přístupy).

V návrhu kvalitativního výzkumu by se mělo objevit, jakou výzkumnou strategii volíme (Creswell 2010):

- Jaký design výzkumu plánujeme? (např. narativní studii / fenomenologickou studii / etnografický přístup?)
- Jaké jsou charakteristiky tohoto přístupu? (Měli bychom čtenáře seznámit se strategií formou definice, využití atd.)
- Proč volíme tuto strategii? V návrhu bychom také měli povysvětlit, proč je tato strategie vhodná.

- Jak vybraná strategie ovlivní typ výzkumných otázek, způsob sběru dat, analýzu a interpretaci?

3. Role výzkumníka v kvalitativním výzkumu

V **kvantitativním výzkumu** je výzkumník jakoby neviditelný. Předpokládá se, že osoba výzkumníka nemůže nijak ovlivnit výsledná data ani jejich analýzu. V **kvalitativním výzkumu** je to jinak. Výzkumník je neoddělitelný od výzkumného přístupu, předpokládá se, že osoba výzkumníka zásadním způsobem ovlivňuje výzkum. Proto by ve zprávě o kvalittaivím výzkumu měly zaznít i informace o samotném zkoumajícím:

- Jaké má výzkumník předchozí zkušenosti s tématem a vztah k tématu?
- Jaké má výzkumník vztahy s informanty a místem výzkumu?
- Jak probíhalo vyjednávání o výzkumu v daných podmínkách? Jaké prostředí bylo vybráno pro výzkum a proč? Jaké aktivity se děly po dobu výzkumu? Rušil výzkum tyto aktivity nebo probíhal na pozadí? Byly využiity nějaké osoby (tzv. gatekeepeři) pro vstup do prostředí, seznámení se s informanty atd.?
- Jaké etické problémy mohly vyvstat během výzkumu?

4. Sběr dat

Data jsou v kvalitativním výzkumu získávána obvykle jednou ze základních metod:

1. **Pozorování** – základní způsoby pozorování se liší podle toho, zda účastníci výzkumu o pozorovateli ví a zda pozorovatel vstupuje do interakcí):
    a. Úplný pozorovatel
    b. Participant jako pozorovatel
    c. Pozorovatel jako participant
    d. Úplná participant
2. **Rozhovory**
    a. Rozhovor tváří v tvář
    b. Telefonický rozhovor
    c. Mailový/chatový rozhovor atd.
    d. Focus group (skupinový rozhovor)
3. **Studium dokumentů**
    a. Veřejné dokumenty (noviny, kroniky, zápisy z jednání)
    b. Soukromé dokumenty (deníky, korespondence…)
4. **Studium audiovizuálních materiálů**
    a. Fotografie
    b. Videozáznamy
    c. Filmy

d. Umělecké objekty atd…

5. Výzkumné protokoly

V kvalitativním výzkumu hrají velkou roli tzv. **výzkumné protokoly**. Protokoly slouží k záznamu pozorování, rozhovorů atd. Především jsou záznamem o tom, jak probíhal výzkum ¨- výzkumník si do nich zapisuje nejen přímou interakci s participanty, výsledky pozorování a odpovědi na jeho otázky v rozhovorech, ale především tzv. field notes – terénní poznámky:

- ***Popisné poznámky*** – jak vypadalo prostředí, kde se odehrával výzkum, jaké jsou charakteristiky participantů – věk, demografie, histirie atd…
- ***Reflexivní poznámky*** – osobní myšlenky, spekulace, vztahy s participanty a vše, co by ze strany výzkumníka mohlo zasahovat do výzkumu.

6. Analýza a interpretace dat

Konečně se dostáváme k samotné **analýze kvalitativních dat**. Jak už byla výše řečeno, analýza kvalitativních dat je především **proces, který se vyvíjí spolu s výzkumem**. Sběr dat, reflexe výzkumných otázek, tvorba poznámek, analýza a interpretace probíhají zároveň – na základě probíhající analýzy a interpretace může dojít k rozhodnutí o dalším sběru dat či rozšíření základny účastníků.

Způsoby analýzy kvalitativních dat se mění podle výzkumné strategie. Creswell hovoří o několika způsobech:

- **Základním způsobem** je sběr a analýza kvalitativních dat, hledání perspektiv a témat a popis cca 4-5 témat, které se v analýze objeví.
- **Zakotvená teorie** jde dále než za tento jednoduchý způsob – postupuje v několika krocích: otevřené kódování, axiální kódování, selektivní kódování. Zakotvená teorie pracuje se třemi úrovněmi: datovými úryvky, které kóduje a kódy poté slučuje do kategorií.
- **Případové studie** a **etnografické výzkumy** využívají detailní popisy, které poté analyzují
- **Fenomenologické studie** hledají významové jednotky.
- **Narativní výzkum** využívá strategii převyprávění příběhů účastníků výzkumu prostřednictvím strukturálních nástrojů.

Creswell (2010) vizualizuje kvalitativní přístup k analýze dat jako soubor několika kroků (v praxi není posloupnost striktní, kroky mohou probíhat zároveň).

1. Syrová data

Přepisy, poznámky, obrázky, fild notes, veškeré materiály...

2. Organizace a příprava dat na analýzu

Optická kontrola dat, přepisy, organizace dat dle typu atd.

3. Podívat se na všechna data

Prvním krokem je získání obecného přehledu o všech datech - o hloubce, důvěryhodnosti, myšlenkách atd...

4. Kódování dat

kódování je procedura, ve které jsou jednotlivé materiály označovány a kategorizovny do segmentů podle objevujících se témat, přístupů, opakujících se vzorců... Více ke kódování dále pod tabulkou.

5. Témata a popis

Fáze popisu celého výzkumu: prostředí, ve ketrém se výzkum odehrává, participanti, události a aktivity probíhající počas výzkumu. Výběr **témat** či kategorií vyplývajících z kódování - tato témata jsou hlavním výstupem kvalitativního výzkumu (obvykle strukturují i výzkumnou zprávu tak, že tvoří podnadpisy k jednotlivým podkapitolám analýzy).

6. Vztahy mezi tématy

Promyšlení, jak budou reprezentovány vztahy mezi tématy. Mohou být vloženy narativní pasáže, vizualizační techniky, tabulky, nákresy, mapy témat, modely...

7. Interpretace významů

Finální krok analýzy. Vztažení výsledků analýzy k existujícímu předešlému výzkumu. Zasazení do kontextu. Srovnání s jinými přístupy a závěry. Formulace nových teorií nebo srovnání se stávajícími teoriemi.

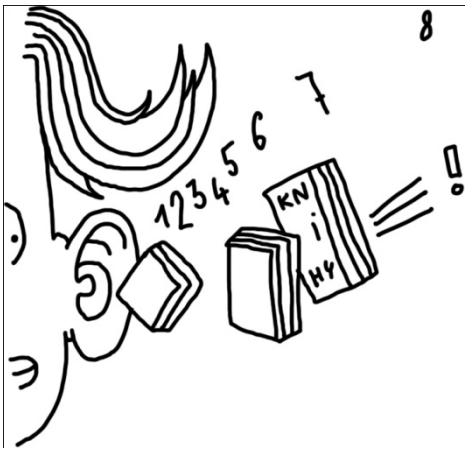
Kódování dat je klíčovou procedurou v analýze a interpretaci kvalitativních dat. Tesch (1990) navrhuje postupovat v osmi krocích:

1. Udělat si celkový obrázek – nejprve si přečíst pečlivě všechny poznámky a přepisy.
2. Analyzovat jednotlivé dokumenty a ptát se „ o čem to je?“ Zapisovat si na okraj poznámky.
3. Po analýze několika dokumentů vytvořit seznam všech témat, které se v poznámkách objevily. Kategorizovat podobná témata do jedné skupiny. Identifikovat hlavní témata, vedlejší a marginální témata.
4. Se seznamem témat se vrátit zpět k datům. Vzniklé kategorii používat jako kódy. Zjistit, zda tyto kódy „fungují“.
5. Promyslet, jak redukovat množství kódů – lze je sloučit do kategorií? Promyslet i vztahy mezi kategoriemi (lze vytvořit např. pojmové mapy).
6. Roztřídit data dle kódů a provést analýzu uvnitř jednotlivých kategorií.
7. Pokud je to nutné (kódy nejsou vhodně zvolené), překódovat soubor dat.

Pro účely kódování by si výzkumník měl zřídit kódovací knihu – tzv. codebook.

Kódovat lze systémem **tužka-papír**, ale i pro analýzu kvalitativních dat dnes existují nástroje a aplikace na PC. Nejznámějšími jsou pravděpodobně:

- Atlas.ti
- RQDA (open source alternativa založená na jazyku R)
- QDA Miner (další free nástroj)
- MAXqda
- QSR NVivo
- HyperRESEARCH



7. Kvalita v kvalitativním výzkumu

Kvalita v kvantitativním výzkumu je posuzována obvykle dvěma ukazateli: validitou a reliabilitou. V kvalitativním výzkumu je situace obdobná (i když někteří autoři navrhují jiné ukazatele), odlišné jsou ale strategie, jakými se reliability a validity dosahuje:

Strategie pro zvyšování reliability výzkumu:

- Kontrola přepisů (nenastala někde chyba?)
- Kontrolou kódů (neposunul se význam kódů?)
- Kontrolou komunikace mezi kódy a výzkumníky v případě teamové práce

Strategie pro zvyšování validity výzkumu:

- Triangulace (kombinace výzkumných metod)
- Follow-up s participanty (předložení výsledků výzkumu účastníkům – odpovídá to jejich realitě?)
- Bohatý popis okolností výzkumu
- Osvětlení pohledů, které do studie vnáší sám výzkumník (reflexivní přístup)
- Peer-review (externí audit, expertní zhodnocení jiným odborníkem)