

# **Metodologie pro Informační studia a knihovnictví 2**

**Modul III:**

**Popis a kontrola dat**

# Metodologie pro Informační studia a knihovnictví 2

## Modul 3: Popis, 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.

# 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:

<b>Jaké je Vaše vzdělání?</b>		<b>Četnost odpovědí</b>	<b>Relativní četnost</b>	<b>Validní relativní četnost</b>
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 %	
<b>Celkem</b>		<b>613</b>	<b>100,0 %</b>	

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

<b>Jaké je Vaše vzdělání?</b>	<b>Četnost odpovědí</b>	<b>Validní relativní četnost</b>
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 %
<b>Celkem</b>	<b>608</b>	<b>100,0 %</b>

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).

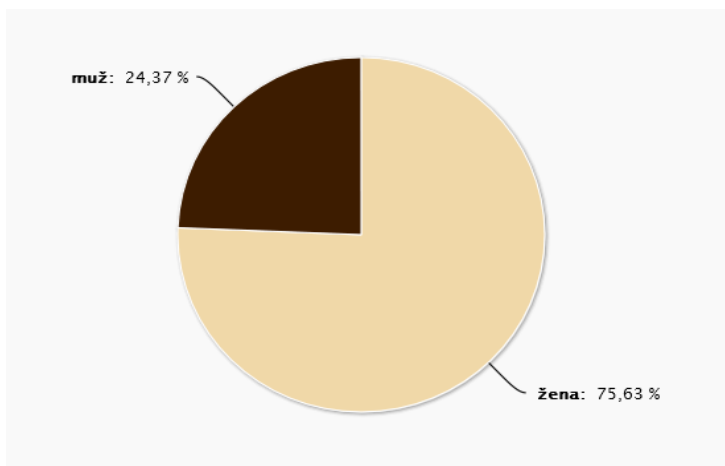
### **Grafy četností**

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

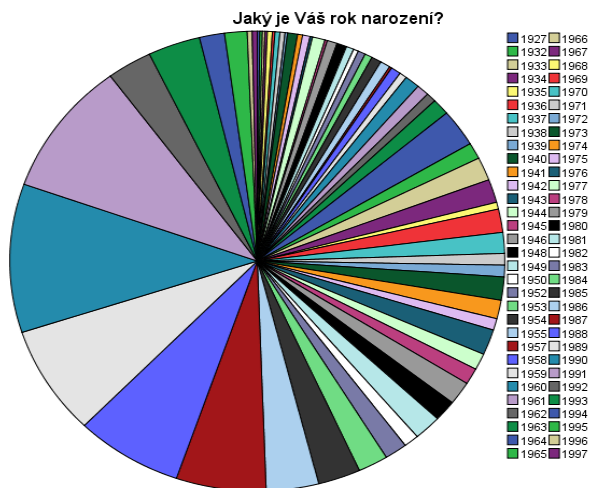
**Koláčový graf** je vhodný:

- 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:



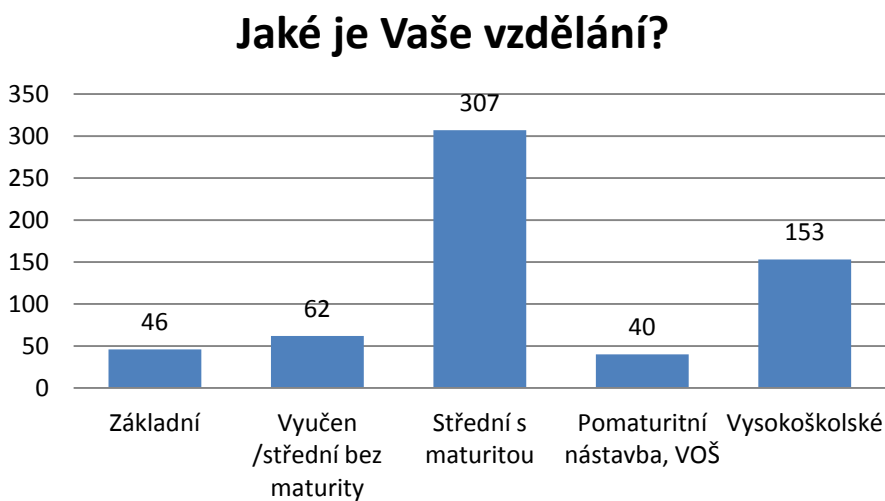
Příklad proměnné, kde NENÍ vhodné využít koláčový graf:



**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

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

## 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**“.

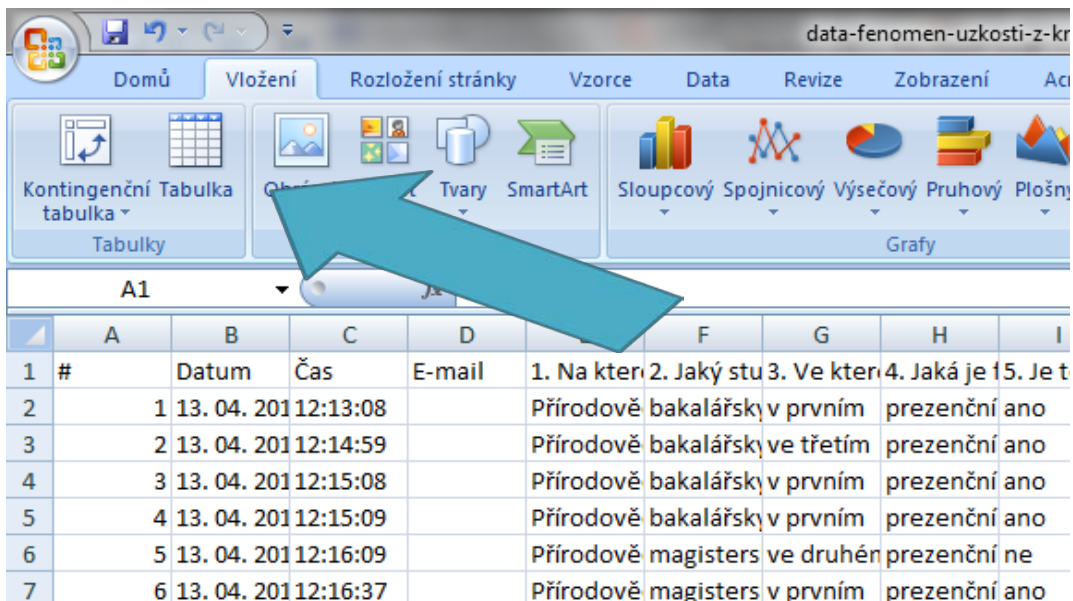
Stažení souboru ve formátu .xls:

The screenshot shows the Google Spreadsheets interface for a document titled "Metodologie2\_dataset\_2013". The "File" menu is open, and the "Download as" option is selected, which has opened a sub-menu. The sub-menu lists several export formats: Microsoft Excel (.xlsx), OpenDocument format (.ods), PDF document (.pdf), Comma Separated Values (.csv, current sheet), Plain text (.txt, current sheet), and Web page (.html, current sheet). The background spreadsheet shows a table with columns P and Q, and rows 1 through 571. The text "cizí řeči - učebnice pro samouky" is visible in cell Q4.

	P	Q	
1	2onlinechat	1_13skupina	
550	2	2	
551	2	4	
552	2	4	
553			
554	4	4	cizí řeči - učebnice pro samouky
555	3	4	
556	4	4	
557	3	3	
558	4	4	
559			
560			
561			
562			
563			
564			...cí na internetu
565			...cí na internetu
566			...cí na internetu
567			
568			
569	3	4	
570	4	4	3 2
571	4	4	4 2

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

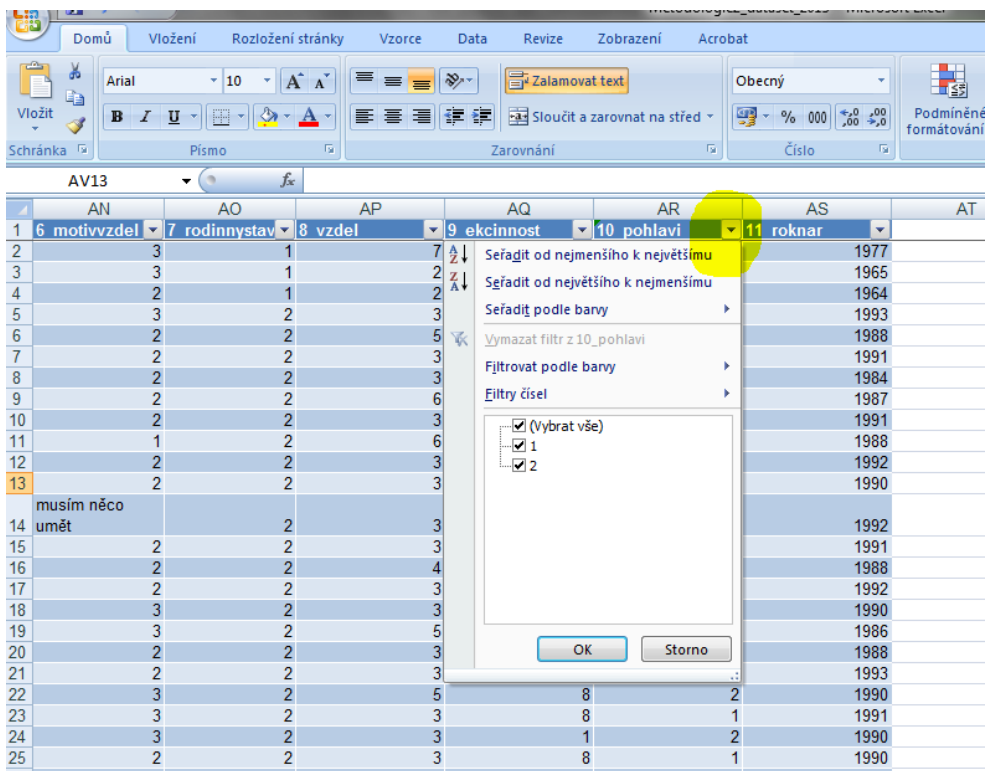




	A	B	C	D	F	G	H	I	
1	#	Datum	Čas	E-mail	1. Na které	2. Jaký stu	3. Ve které	4. Jaká je	5. Je t
2	1	13. 04. 201	12:13:08		Přírodově	bakalářsky	v prvním	prezenční	ano
3	2	13. 04. 201	12:14:59		Přírodově	bakalářsky	ve třetím	prezenční	ano
4	3	13. 04. 201	12:15:08		Přírodově	bakalářsky	v prvním	prezenční	ano
5	4	13. 04. 201	12:15:09		Přírodově	bakalářsky	v prvním	prezenční	ano
6	5	13. 04. 201	12:16:09		Přírodově	magisters	ve druhém	prezenční	ne
7	6	13. 04. 201	12:16:37		Přírodově	magisters	v prvním	prezenční	ano

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:

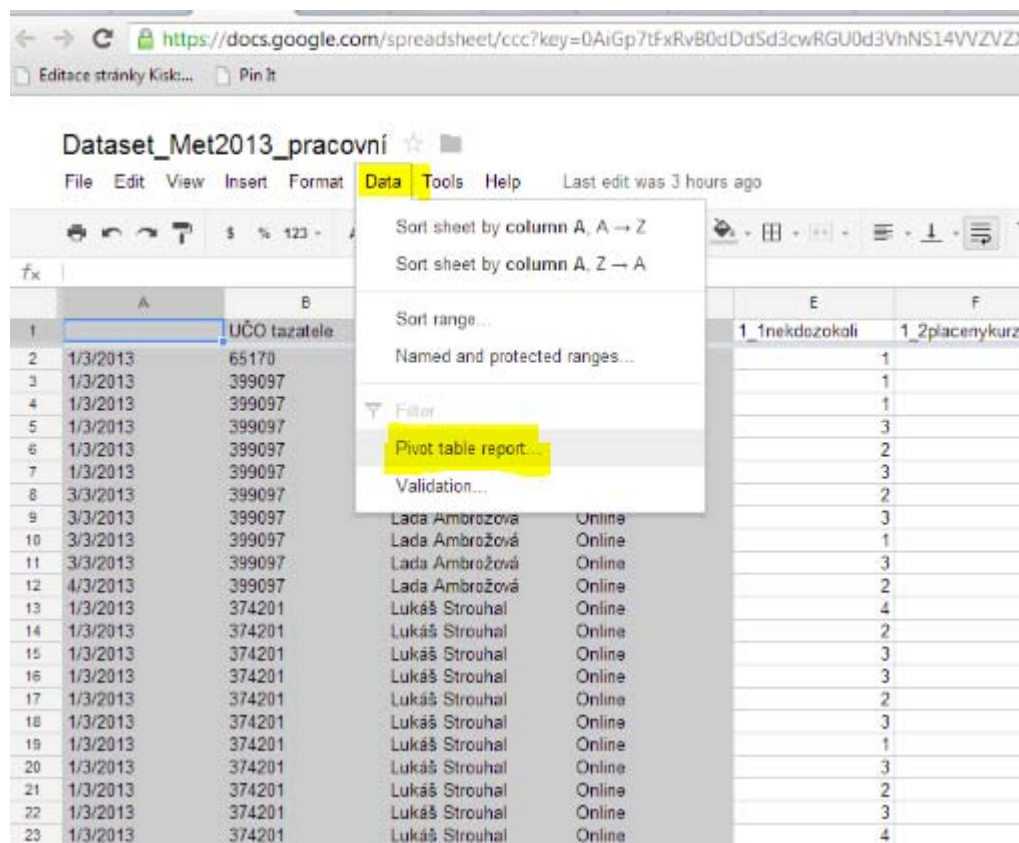


## ***Popis rozložení hodnot proměnných***

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

A	B
<b>Prodejce</b>	<b>Faktura</b>
Novák	15 000
Novák	9 000
Horák	8 000
Horák	20 000
Novák	5 000
Veselý	22 500
<b>Vzorec</b>	<b>Popis (výsledek)</b>
=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)

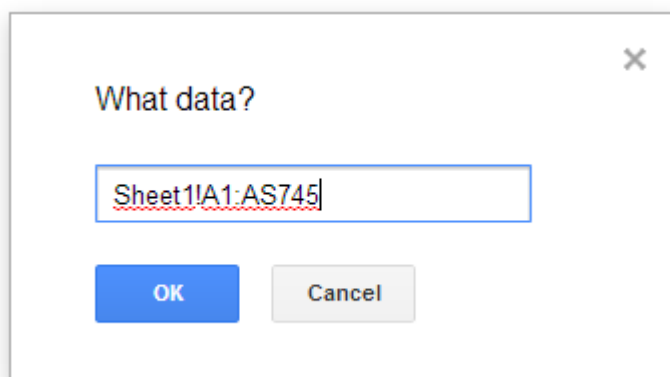
Příkaz COUNTIF nám spočítá výskyt konkrétní varianty hodnoty proměnné. Pro vytvoření tabulky četností je však užitečnější funkce „pivot tables“. Najdete ji v sekci „Data“.



The screenshot shows a Google Sheets interface with a spreadsheet titled "Dataset\_Met2013\_pracovni". The "Data" menu is open, and the "Pivot table report..." option is highlighted in yellow. The spreadsheet contains data with columns A, B, E, and F. Column A contains dates, column B contains IDs, column E contains "1\_1nekdozokoli", and column F contains "1\_2placenykurz".

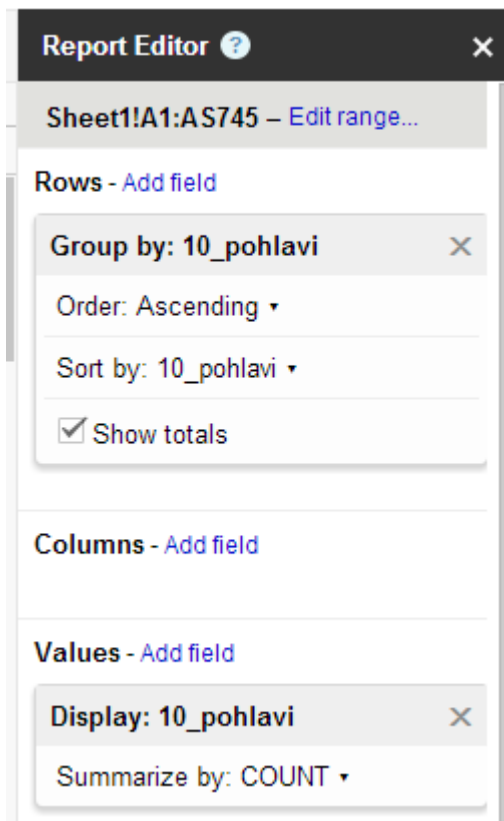
	A	B	E	F
1		ÚČO tazatele		
2	1/3/2013	65170		
3	1/3/2013	399097		
4	1/3/2013	399097		
5	1/3/2013	399097		
6	1/3/2013	399097		
7	1/3/2013	399097		
8	3/3/2013	399097		
9	3/3/2013	399097		
10	3/3/2013	399097		
11	3/3/2013	399097		
12	4/3/2013	399097		
13	1/3/2013	374201		
14	1/3/2013	374201		
15	1/3/2013	374201		
16	1/3/2013	374201		
17	1/3/2013	374201		
18	1/3/2013	374201		
19	1/3/2013	374201		
20	1/3/2013	374201		
21	1/3/2013	374201		
22	1/3/2013	374201		
23	1/3/2013	374201		

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



The dialog box titled "What data?" has a close button (X) in the top right corner. The text "What data?" is centered at the top. Below it is a text input field containing the range "Sheet1!A1:AS745". At the bottom of the dialog are two buttons: "OK" (blue) and "Cancel" (grey).

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“.



Chceme popsat  
proměnnou  
„Pohlaví“

Zajímají nás četnosti u  
jednotlivých hodnot  
proměnné  
„Pohlaví“

Zpracování v Google Spreadsheets může chvíli 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“).

<b>Jste:</b>	<b>Četnost odpovědí</b>	<b>Validní relativní četnost</b>
Muž	80	40 %
Žena	120	60 %
<b>Celkem</b>	<b>200</b>	<b>100 %</b>

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	13. 04. 2011	12:13:08	Přírodovědecká fakulta
2	13. 04. 2011	12:14:59	Přírodovědecká fakulta
3	13. 04. 2011	12:15:08	Přírodovědecká fakulta
4	13. 04. 2011	12:15:09	Přírodovědecká fakulta
5	13. 04. 2011	12:16:09	Přírodovědecká fakulta
6	13. 04. 2011	12:16:37	Přírodovědecká fakulta
7	13. 04. 2011	12:18:16	Přírodovědecká fakulta
8	13. 04. 2011	12:18:17	Přírodovědecká fakulta
9	13. 04. 2011	12:18:18	Přírodovědecká fakulta
10	13. 04. 2011	12:18:50	Přírodovědecká fakulta
11	13. 04. 2011	12:19:01	Přírodovědecká fakulta
12	13. 04. 2011	12:20:21	Přírodovědecká fakulta
13	13. 04. 2011	12:20:26	Přírodovědecká fakulta
14	13. 04. 2011	12:20:27	Přírodovědecká fakulta
15	13. 04. 2011	12:20:40	Přírodovědecká fakulta

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

**Do řádků přetáhneme proměnnou, kterou chceme popsat. Stejnou proměnnou přetáhneme i do políčka „Hodnoty“.**

Pro ukázkou 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 nastavba, VOŠ	5
5 - VŠ bakalářské	34
6 - VŠ magisterské	21
7 - VŠ doktorské	4
<b>Celkový součet</b>	<b>185</b>

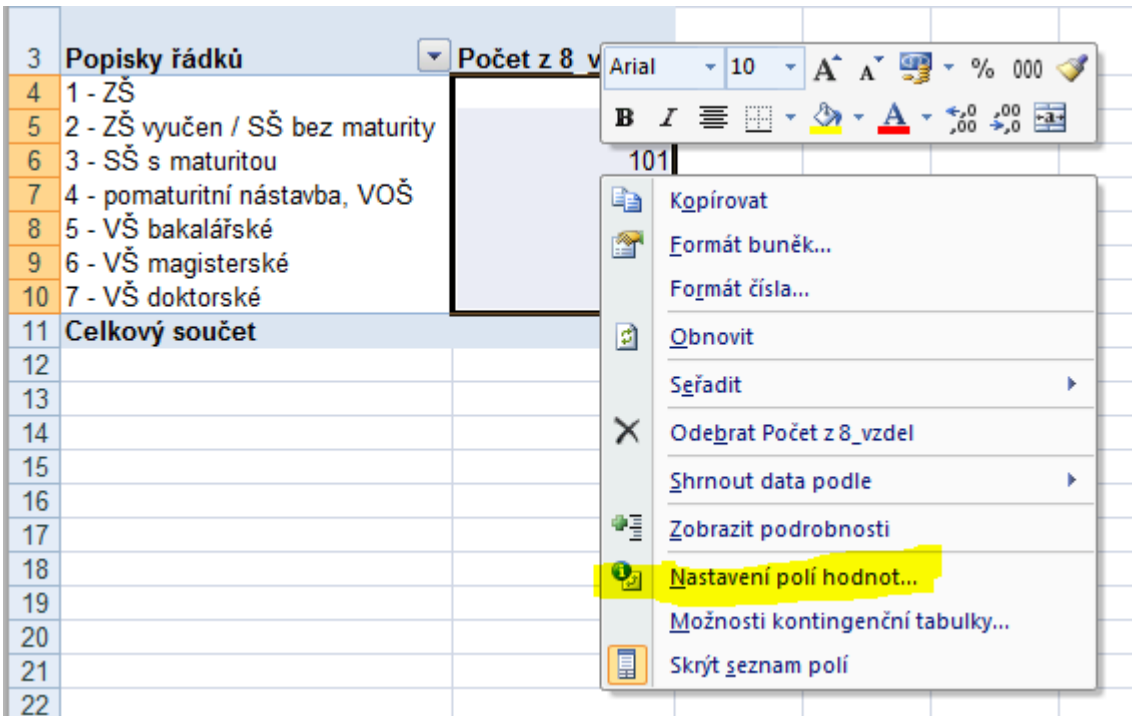
Tato tabulka ukazuje hodnoty výskytu jednotlivých variant odpovědí u proměnné „Dokončené vzdělání“

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:

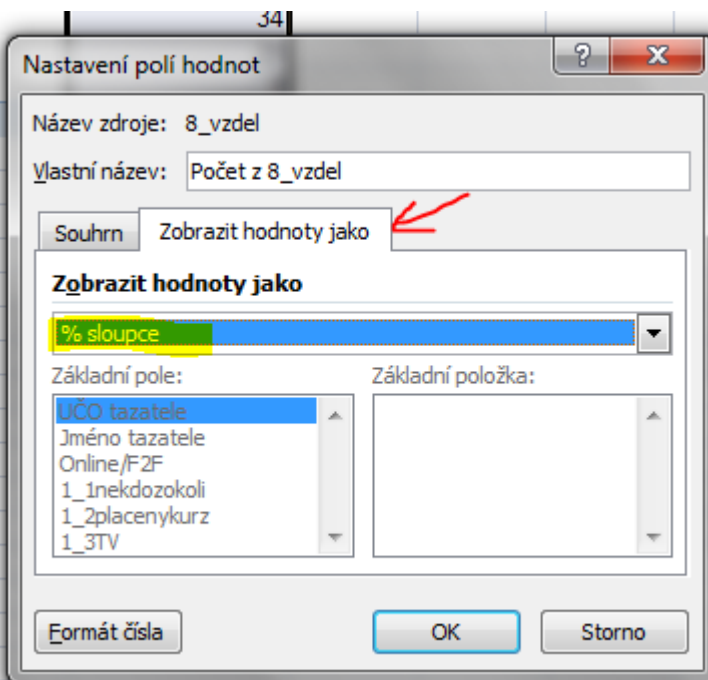
3	Popisky řádků	Počet z 8_vzdel
	Seřadit od A do Z	8
	Seřadit od Z do A	12
	Další možnosti řazení...	101
	Vymazat filtr z 8_vzdel	5
	Filtry popiseků	34
	Filtry hodnot	21
	<input checked="" type="checkbox"/> (Vybrat vše) <input checked="" type="checkbox"/> 1 - ZŠ <input checked="" type="checkbox"/> 2 - ZŠ vyučen / SŠ bez maturity <input checked="" type="checkbox"/> 3 - SŠ s maturitou <input checked="" type="checkbox"/> 4 - pomaturitní nastavba, VOŠ <input checked="" type="checkbox"/> 5 - VŠ bakalářské <input checked="" type="checkbox"/> 6 - VŠ magisterské <input checked="" type="checkbox"/> 7 - VŠ doktorské	4
		<b>185</b>

Zde můžeme „odškrtnout“ nevalidní hodnoty

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 nastavba, VOŠ	2,70%
5 - VŠ bakalářské	18,38%
6 - VŠ magisterské	11,35%
7 - VŠ doktorské	2,16%
<b>Celkový součet</b>	<b>100,00%</b>

## Minimální a maximální hodnoty

Minimální a maximální hodnoty lze rozpoznat už z popisu rozložení proměnných. 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)

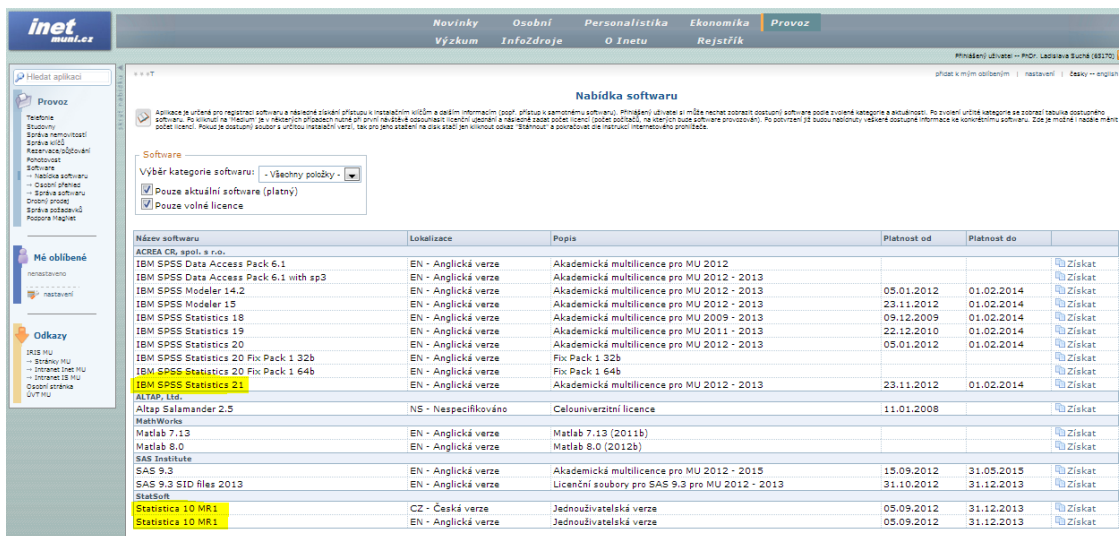
## ***Využívejte podpory a nápovědy!***

Pokud si nejste jistí provedením příkazu, využijte podpory [Microsoft Office](#) i [Google Spreadsheets](#). Na internetu lze najít také spoustu videotutorialů a návodů. V nejhorším případě pište na [sucha@phil.muni.cz](mailto:sucha@phil.muni.cz) ☺.

# Návod pro práci s SPSS

## Instalace programu

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



The screenshot shows the 'Nabídka softwaru' (Software Offer) page on the INET portal. The page features a navigation menu on the left with categories like 'Provoz', 'Mě oblíbené', and 'Odkazy'. The main content area displays a table of software licenses with columns for 'Název softwaru', 'Lokalizace', 'Popis', 'Platnost od', and 'Platnost do'. The table lists various software packages including IBM SPSS Data Access Pack 6.1, IBM SPSS Modeler 14.2, IBM SPSS Statistics 18, 19, 20, and 21, as well as MATLAB and SAS. The 'IBM SPSS Statistics 21' license is highlighted in yellow. A sidebar on the left contains a search bar and a list of software categories.

Název softwaru	Lokalizace	Popis	Platnost od	Platnost do	
ACRIA, s.r.o.					
IBM SPSS Data Access Pack 6.1	EN - Anglická verze	Akademická multilicence pro MU 2012			Získat
IBM SPSS Data Access Pack 6.1 with sp3	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013			Získat
IBM SPSS Modeler 14.2	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013	05.01.2012	01.02.2014	Získat
IBM SPSS Modeler 15	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013	23.11.2012	01.02.2014	Získat
IBM SPSS Statistics 18	EN - Anglická verze	Akademická multilicence pro MU 2009 - 2013	09.12.2009	01.02.2014	Získat
IBM SPSS Statistics 19	EN - Anglická verze	Akademická multilicence pro MU 2011 - 2013	22.12.2010	01.02.2014	Získat
IBM SPSS Statistics 20	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013	05.01.2012	01.02.2014	Získat
IBM SPSS Statistics 20 Fix Pack 1 32b	EN - Anglická verze	Fix Pack 1 32b			Získat
IBM SPSS Statistics 20 Fix Pack 1 64b	EN - Anglická verze	Fix Pack 1 64b			Získat
IBM SPSS Statistics 21	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2013	23.11.2012	01.02.2014	Získat
ALTAP, Ltd.					
Altap Salamander 2.5	NS - Nespecifikováno	Celouniverzitní licence	11.01.2008		Získat
MathWorks					
Matlab 7.13	EN - Anglická verze	Matlab 7.13 (2011b)			Získat
Matlab 8.0	EN - Anglická verze	Matlab 8.0 (2012b)			Získat
SAS Institute					
SAS 9.3	EN - Anglická verze	Akademická multilicence pro MU 2012 - 2015	15.09.2012	31.05.2015	Získat
SAS 9.3 SID files 2013	EN - Anglická verze	Licenční soubory pro SAS 9.3 pro MU 2012 - 2013	31.10.2012	31.12.2013	Získat
StatSoft					
Statistics 10 MR1	CZ - Česká verze	Jednoulivatelská verze	05.09.2012	31.12.2013	Získat
Statistics 10 MR1	EN - Anglická verze	Jednoulivatelská verze	05.09.2012	31.12.2013	Získat

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.

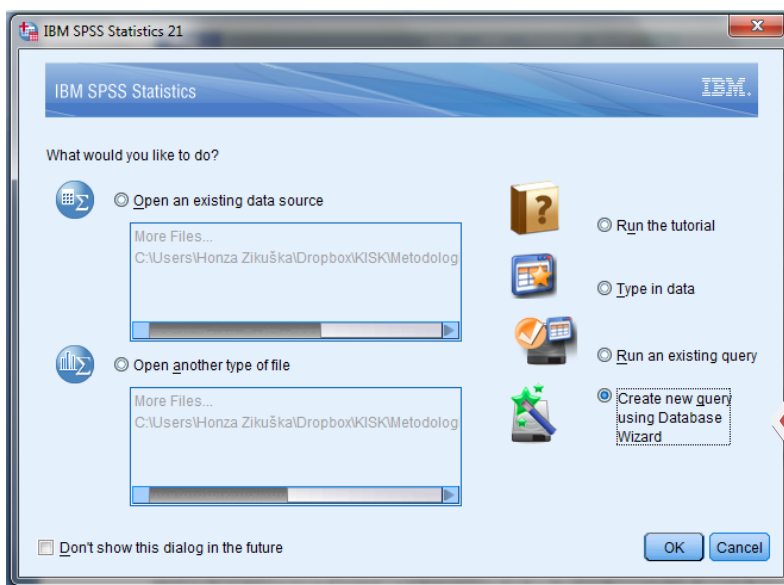
Dostupných je hned několik druhů licencí – doporučuji vybrat licenci **IBM SPSS Statistics 21** (nejnovější verze programu).

## Otevření souborů s daty

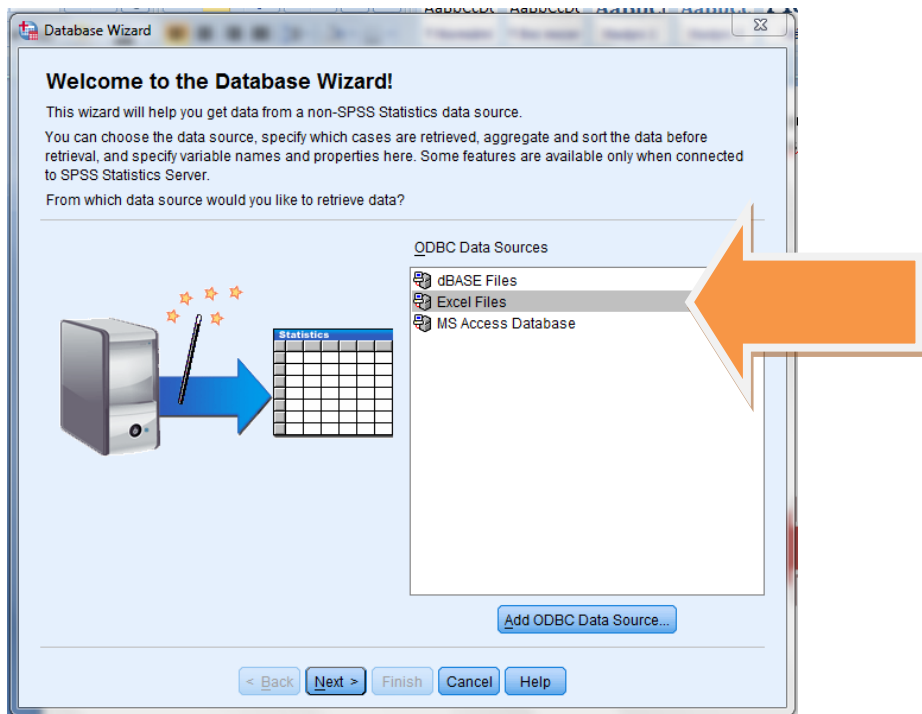
SPSS tedy máme nainstalované – najdete jej v nabídce Start nebo v přehledu vašich programů. Do SPSS můžete data dostat několika způsoby – ten nejjednodušší je přímé tvoření datasetu v SPSS. My ale budeme potřebovat pracovat s daty, která již máme ve formátu .xls.

Postupovat budeme následovně:

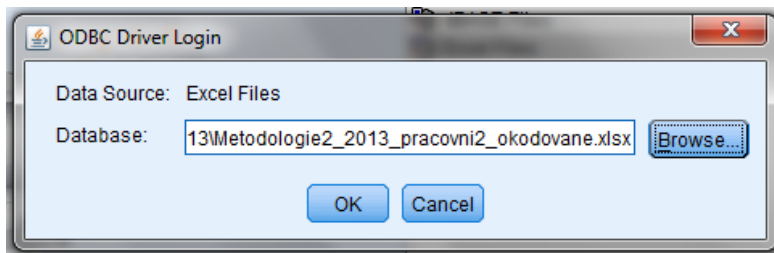
1. Uložíme si na své PC datový soubor ve formátu pro Excel (najdeme jej v ISu).
2. Pro převedení excelového souboru do souboru typu .sav spustíme „Database Wizzard“:



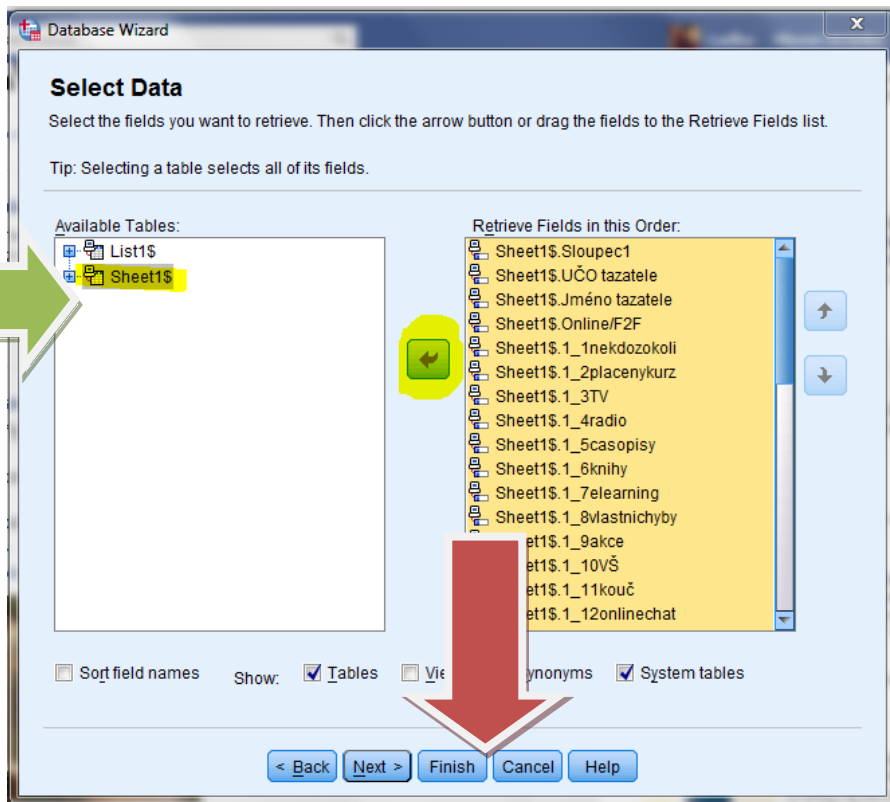
3. Z nabízených možností v dalším okně si vyberte „Excel files“:



4. Vyberte soubor ze svého PC:



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

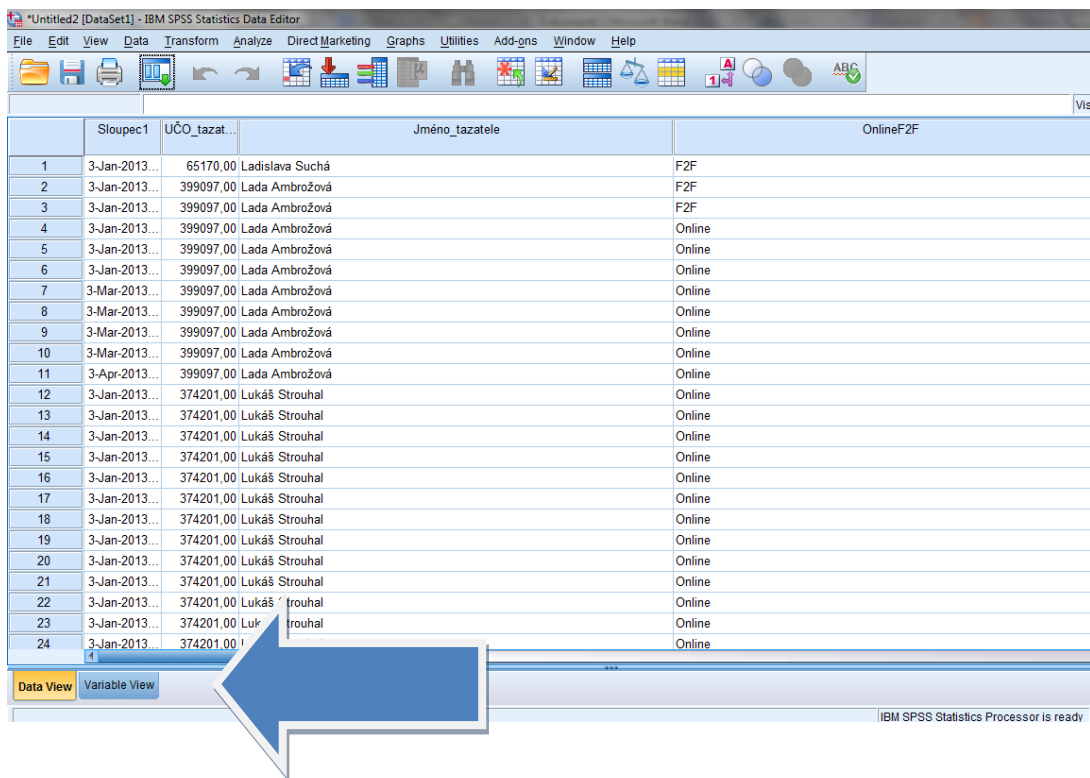


6. V počítači se vám otevrou dvě nová okna. Jedno přímo s datasetem a druhé je tzv. „Output“ – okno, kam se zapisují procesy a výsledky operací SPSS.

## 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.



The screenshot shows the IBM SPSS Statistics Data Editor interface. The main window displays a dataset with the following columns: Sloupec1, ÚČO\_tazat..., Jméno\_tazatele, and OnlineF2F. The data is presented in a table format with 24 rows. The 'Data View' button is highlighted in the bottom left corner, and a blue arrow points to it from the text below.

	Sloupec1	ÚČO_tazat...	Jméno_tazatele	OnlineF2F
1	3-Jan-2013...	65170,00	Ladislava Suchá	F2F
2	3-Jan-2013...	399097,00	Lada Ambrožová	F2F
3	3-Jan-2013...	399097,00	Lada Ambrožová	F2F
4	3-Jan-2013...	399097,00	Lada Ambrožová	Online
5	3-Jan-2013...	399097,00	Lada Ambrožová	Online
6	3-Jan-2013...	399097,00	Lada Ambrožová	Online
7	3-Mar-2013...	399097,00	Lada Ambrožová	Online
8	3-Mar-2013...	399097,00	Lada Ambrožová	Online
9	3-Mar-2013...	399097,00	Lada Ambrožová	Online
10	3-Mar-2013...	399097,00	Lada Ambrožová	Online
11	3-Apr-2013...	399097,00	Lada Ambrožová	Online
12	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
13	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
14	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
15	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
16	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
17	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
18	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
19	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
20	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
21	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
22	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
23	3-Jan-2013...	374201,00	Lukáš Strouhal	Online
24	3-Jan-2013...	374201,00	Lukáš Strouhal	Online

**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:

2. Považujete obor Informační studia a knihovnictví za perspektivní?

- velmi perspektivní
- spíše perspektivní
- spíše neperspektivní
- zcela neperspektivní
- nevím, nemohu odpovědět
- neodpověděl/a

- 1
- 2
- 3
- 4
- 1
- 2

Chybějící hodnoty (missing values)

Hodnoty proměnné okódované

Takto bude vypadat matice dat:

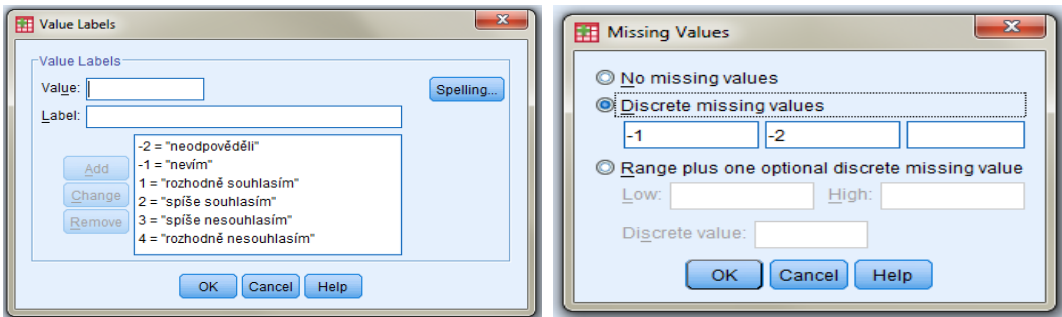
Q1\_prinos  
Studium na KISK hodnotím jako:

	q1_prinos	q2_perspektiva	q3_topoleneni	q4_znovoustud	q5_eal	q8_1_posl	q8_2_pr...	q8_3_tema	q8_4_pocetkurzu	q8_5_pra...	q9_1_navhkurzu	q9_2_na...	q9_3_na...	q10
1	1													
2	1													
3	2													
4	1	2	1	1	1									
5	1	2	2	1	1									
6	2	1	1	1	2									
7	1	1	2	1	1									
8	1	1	1	1	1									
9	2	2	1	1										
10	2	2	4	2	2									
11	2	2	4	2	2									
12	2	1	1	1	2									
13	1	2	2	1	1									
14	1	2	2	1	1									
15	2	1	1	1	1									
16	1	2	2	2	2									
17	2	2	2	2	1	2								
18	2	2	2	2	1	2								
19	2	2	2	2	1	2								
20	2	2	2	2	1	1								
21	2	2	2	2	1	2								
22	2	2	2	2	1	1								
23	2	2	2	1	1	1								

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 nezahrneme do dané analýzy – SPSS s nimi v konkrétních operacích nebude počítat.
- *Measure*: typ proměnné (nominální/ordinální/kardinální)

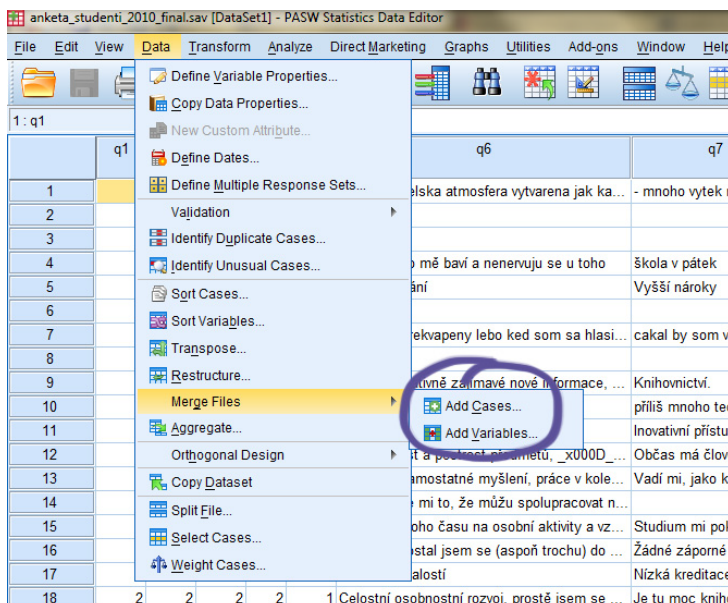


Ve studijních materiálech v ISu máte již datasey s popsányi proměnnými.

## 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

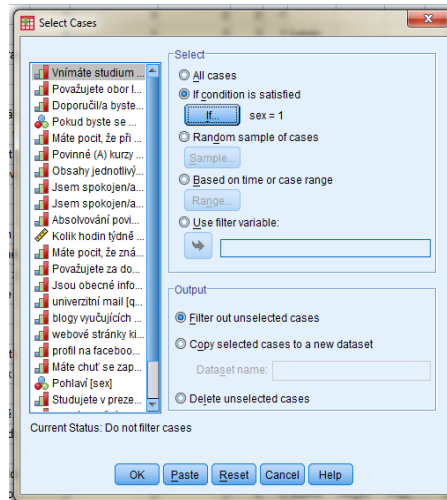




## Výběr případů

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 jakémkoliv 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

Co s missing values?

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 nastavba, VOS	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
	Babara 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 Šíková	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

Chybný zápis jména

Statistics

Pohlaví

N	Valid	613
	Missing	1

Pohlaví

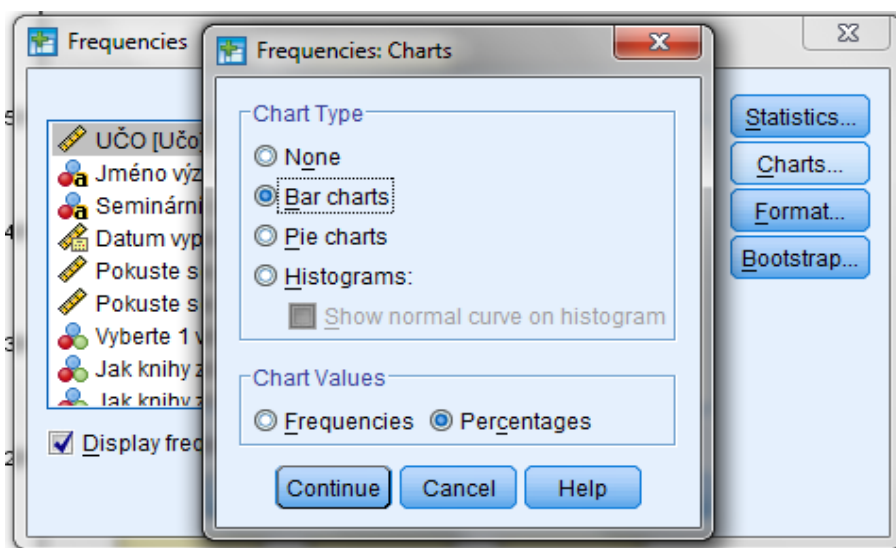
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Muž	279	45,4	45,5	45,5
	Žena	333	54,2	54,3	99,8
	7	1	,2	,2	100,0
	Total	613	99,8	100,0	
Missing	System	1	,2		
	Total	614	100,0		

Chyba: proměnná „Pohlaví“ by neměla nabývat hodnoty 7

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**.



## **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.