

3. Příprava dat, Excel



Klíčový význam korektního uložení získaných dat
Pravidla pro ukládání dat
Čištění dat před analýzou

Anotace



- Současná statistická analýza se neobejde bez zpracování dat pomocí statistického software. Předpokladem úspěchu je správné uložení dat v definované formě.
- Nejčastěji jde o databázové tabulky umožňující zpracování dat v celé škále různých aplikací.
- Neméně důležité je věnovat pozornost čištění dat předcházejícímu vlastní analýze. Každá chyba, která vznikne nebo není nalezena ve fázi přípravy dat, se promítne do všech dalších kroků a může zapříčinit neplatnost výsledků a nutnost opakování analýzy.

DATA – ukázka uspořádání datového souboru

Parametry (znaky)



Opakování

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	Report	Country	Site	Matrix	ampl_met	Paramete	Year	Month	Day	Mean	Unit	Value	LOQ	Note	Page	Backgrou	in report
2	CEEC	Armenia	Sevan, Tsc	Air	pas	o,p-DDE	2008	7	29		pg/m3	3	0,5		REC	yes	no
3	CEEC	Armenia	Sevan, Tsc	Air	pas	PCB 118	2008	7	29		pg/m3	3,2	0,5		REC	yes	yes
4	CEEC	Armenia	Artashat, i	Air	pas	p,p-DDD	2008	7	29		pg/m3	3,7	0,5		REC	yes	yes
5	CEEC	Kazakhsta	Borovoe	Air	pas	PeCB	2008	7	29		pg/m3	3,9	0,5		REC	yes	yes
6	CEEC	Armenia	Yerevan, [Air	pas	PCB 138	2008	7	29		pg/m3	4,4	0,5		REC	no	yes
7	CEEC	Armenia	Yerevan, [Air	pas	PCB 153	2008	7	29		pg/m3	4,4	0,5		REC	no	yes
8	CEEC	Kazakhsta	Borovoe	Air	pas	gamma-H	2008	7	29		pg/m3	9,4	0,5		REC	yes	yes
9	CEEC	Armenia	Sevan, Tsc	Air	pas	PCB 28	2008	7	29		pg/m3	9,6	0,5		REC	yes	yes
10	CEEC	Armenia	Artashat, i	Air	pas	PCB 153	2008	7	29		pg/m3	9,9	0,5		REC	yes	yes
11	CEEC	Armenia	Amberd, r	Air	pas	o,p-DDE	2008	7	29		pg/m3	10	0,5		REC	yes	yes
12	CEEC	Armenia	Yerevan, [Air	pas	p,p-DDD	2008	7	29		pg/m3	10,2	0,5		REC	no	yes
13	CEEC	Armenia	Artashat, i	Air	pas	PCB 138	2008	7	29		pg/m3	10,5	0,5		REC	yes	yes
14	WEOG	USA	Eagle Hart	Air	active	Mirex	1990	11	16		pg/m3	0,03				yes	IADN
15	WEOG	Canada	Alert	Air	active	HCB	1995			60,8	pg/m3				A1_69		
16	WEOG	USA	Eagle Hart	Air	active	Gamma-H	1990	11	16		pg/m3	0,777				yes	IADN
17	WEOG	USA	Eagle Hart	Air	active	Alpha-HCl	1990	11	16		pg/m3	1,482				yes	IADN
18	WEOG	USA	Eagle Hart	Air	active	p,p-DDE	1990	11	16		pg/m3	2,428				yes	IADN
19	WEOG	USA	Eagle Hart	Air	active	Dieldrin	1990	11	16		pg/m3	3,993				yes	IADN
20	WEOG	USA	Eagle Hart	Air	active	PCB 101	1990	11	16		pg/m3	5,036				yes	IADN
21	WEOG	USA	Eagle Hart	Air	active	PCB 52	1990	11	16		pg/m3	6,764				yes	IADN
22	WEOG	USA	Eagle Hart	Air	active	p,p-DDD	1990	11	16		pg/m3	11,442				yes	IADN
23	WEOG	USA	Eagle Hart	Air	active	PCB 44	1990	11	16		pg/m3	12,613				yes	IADN
24	WEOG	USA	Eagle Hart	Air	active	Gamma-H	1990	11	16		pg/m3	24,33				yes	IADN
25	WEOG	Canada	Alert	Air	active	HCB	1998			70	pg/m3				A1_10		
26	WEOG	USA	Eagle Hart	Air	active	Alpha-HCl	1990	11	16		pg/m3	268,831				yes	IADN
27	WEOG	USA	Eagle Hart	Air	active	Aldrin	1990	11	16		pg/m3	<LoQ				yes	IADN
28	WEOG	USA	Eagle Hart	Air	active	Aldrin	1990	11	16		pg/m3	<LoQ				yes	IADN
29	WEOG	USA	Eagle Hart	Air	active	Dieldrin	1990	11	16		pg/m3	<LoQ				yes	IADN
30	WEOG	USA	Eagle Hart	Air	active	p,p-DDD	1990	11	16		pg/m3	<LoQ				yes	IADN
31	WEOG	USA	Eagle Hart	Air	active	p,p-DDE	1990	11	16		pg/m3	<LoQ				yes	IADN

Zásady pro ukládání dat



- Správné a přehledné uložení dat je základem jejich pozdější analýzy.
- Je vhodné rozmyslet si předem jak budou data ukládána.
- Pro počítačové zpracování dat je vhodné ukládat data v tabulární formě.
- Nejvhodnějším způsobem je uložení dat ve formě databázové tabulky:
 - každý sloupec obsahuje pouze jediný typ dat, identifikovaný hlavičkou sloupce;
 - každý řádek obsahuje minimální jednotku dat (např. pacient, měření apod.);
 - je nepřípustné kombinovat v jednom sloupci číselné a textové hodnoty;
 - komentáře jsou uloženy v samostatných sloupcích;
 - u textových (kategoriálních) dat je nezbytné kontrolovat překlepy v názvech kategorií;
 - specifickým typem dat jsou kalendářní data u nichž je nezbytné kontrolovat, zda jsou uložena v korektním formátu (dle aplikace).
- Takto uspořádaná data je v tabulkových nebo databázových programech možné převést na libovolnou výstupní tabulku.
- Pro základní uložení a čištění dat menšího rozsahu je možné využít aplikací MS Office.

MS Excel

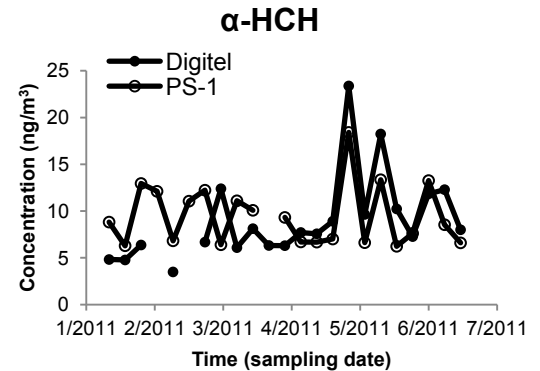


- Tabulkový procesor.
- První verze programu 30. 9. 1985 (Macintosh).
- Součást balíku kancelářských aplikací MS Office.
- Aktualizace každé 2 až 3 roky; nové funkce, rozšíření počtu řádků a sloupců, změna formátu.
- Nejnovější formát Office XML je zazipovaný XML dokument, přípona .xlsx.
- Aktuální verze 2016 umožňuje ukládat tabulku až o 1 048 576 řádcích a 16 384 sloupcích.
- Maximální velikost buňky je 32 767 znaků.
- Excel umožňuje práci se širokou škálou dalších formátů.

Možnosti MS Excel



- Správa a práce s tabulárními daty.
- Řazení dat, výběry z dat, přehledy dat.
- Formátování a přehledné zobrazení dat.
- Zobrazení dat ve formě grafů.
- Různé druhy výpočtů pomocí zabudovaných funkcí.
- Tvorba tiskových sestav.
- Makra – zautomatizování častých činností.
- Tvorba aplikací (Visual Basic for Applications).



17	10	2
18	12	3
19	5	4
20	8	5
21	4	8
22	7	9
23	9	11
24	suma součinů řádků	310
25		

P. bini	2	Pohlaví			
Počet z	Délka				
Číslo	ryby2	Číslo	rv1	Váha	?
1	1	<input type="checkbox"/>	(Zobrazit vše)		
2	2	<input checked="" type="checkbox"/>	68		
7	26	<input checked="" type="checkbox"/>	99		
8	106	<input checked="" type="checkbox"/>	102		
9	121	<input checked="" type="checkbox"/>	109		
10	160	<input checked="" type="checkbox"/>	112		
11	34	<input checked="" type="checkbox"/>	120		
12	45	<input checked="" type="checkbox"/>	173		
13	70	<input checked="" type="checkbox"/>	28		
14	72	<input checked="" type="checkbox"/>	29		
15	87	<input checked="" type="checkbox"/>	29		
16	Celkový součet	OK	Storno		
17					

Import a export dat



- **Import dat**
 - manuální zadávání;
 - import souborů – podpora importu ze starších verzí Excelu, textových souborů, databází apod.;
 - kopírování přes schránku Windows – vkládání z nejrůznějších aplikací – MS Office, Statistica, přímo z HTML apod.;
 - využití textových souborů jako kompatibilního formátu pro přenos dat mezi různými aplikacemi;
 - import z připojených databází (ODBC);
 - načítání dat z webové stránky.
- **Export dat**
 - ukládáním souborů ve formátech podporovaných jinými SW, časté jsou textové soubory, dbf soubory nebo starší verze Excelu;
 - přímé kopírování přes schránku Windows.

Import a export dat



- **Nejčastější datové formáty používané v MS Excel**
 - **.xlsx** – současný Office Open XML formát od verze MS Excel 2007;
 - **.xls** – starší binární varianta listů MS Excel (více verzí), stále používaná,
 - **.csv** – comma separated values, nejjednodušší tabulkový formát, 2 varianty,
 - **.dbf** – formát dBase, široce využívaný formát pro velké databáze;
 - **.db** – Paradox database, starší databázový systém;
 - **.slk** – SYmbolic LinK (SYLK) formát pro výměnu dat mezi aplikacemi Microsoft, neveřejný;
 - **.txt** – základní textový formát, často jediná možnost výměny dat s MS Excel.

Tipy a triky



- **Výběr buněk**

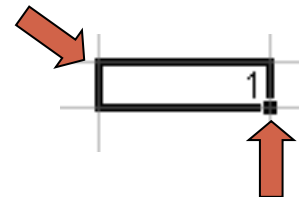
- CTRL+HOME – přesunutí na levý horní roh tabulky;
- CTRL+END – přesunutí na pravý dolní roh tabulky;
- CTRL+A – výběr celého listu;
- CTRL + klepnutí myší do buňky – výběr jednotlivých buněk ;
- SHIFT + klepnutí myší na jinou buňku – výběr bloku buněk;
- SHIFT + šipky – výběr sousedních buněk ve směru šipky;
- SHIFT+CTRL+END (HOME) – výběr do konce (začátku) oblasti dat v listu;
- SHIFT+CTRL+šipky – výběr souvislého řádku nebo sloupce buněk;
- SHIFT + klepnutí na objekty – výběr více objektů.

- **Kopírování a vkládání**

- CTRL+C – zkopírování označené oblasti buněk;
- CTRL+V – vložení obsahu schránky – oblast buněk, objekt, data z jiné aplikace;

- **Myš a okraje buňky**

- Chycení myší za okraj umožňuje přesun buňky nebo bloku buněk



- Při chycení čtverečku v pravém dolním rohu výběru je tažením možno vyplnit více buněk hodnotami původní buňky (ve vzorcích se mění relativní odkazy, je také možné vyplnění hodnotami ze seznamu – např. po sobě jdoucí názvy měsíců).

Ukotvení příček



- Umožňuje ukotvení libovolných řádků a sloupců pro pohodlné vkládání a prohlížení dat v tabulce.
- Umožňuje číst řádky/sloupce ze začátku tabulky i po přesunutí se dále.
- Záložka „Zobrazení“ → „Ukotvit příčky“.

- Nabízené možnosti:

- Ukotvit příčky – ukotví řádky nad označenou buňkou a sloupce vlevo od označené buňky.
- Ukotvit horní řádek.
- Ukotvit první sloupec.
- Ukotvení zrušíme opětovným odkliknutím možnosti ukotvení příček.

	F	G	H	
	poslední kontrola	pohlaví	nemocný	tíže
19	9.4.2010	muž		1
10	29.3.2010	muž		1

Databázová struktura dat v Excelu



Sloupce tabulky = parametry záznamů, hlavička udává obsah sloupce
– stejný údaj v celém sloupci

Jednotlivé záznamy
(taxon, lokalita,
měření, pacient atd.)



	A	B	C	D	E	F	G	H	I
1	Číslo	Značka	Společ	Pohlaví	Délka	Váha	P. anguillae	P. bini	
2	1	1	1	m	27,5	23,0	2	2	
3	2	2	2	f	34,0	62,5	0	2	
4	3	5	3	f	58,0	230,0	0	0	
5	4	6	4	f	42,0	155,0	0	0	
6	5	7	5	f	44,0	149,8	0	0	
7	6	8	6	f	56,0	323,0	0	1	
8	7	9	7	m	48,5	178,2	0	0	
9	8	10	8	f	30,5	47,7	4	6	
10	9	11	9	f	47,0	175,9	5	14	
11	10	12	10	f	40,0	85,1	5	9	
12	11	14	11	f	40,0	101,0	0	0	
13	12	15	12	f	31,0	84,0	15	9	
14	13	16	13	f?	22,0	9,0	0	0	
15	14	17	14	f	42,0	108,0	1	3	
16	15	18	15	f	44,0	130,0	0	0	
17	16	19	16	f	37,0	85,0	2	5	
18	17	20	17	f	50,0	212,0	1	8	

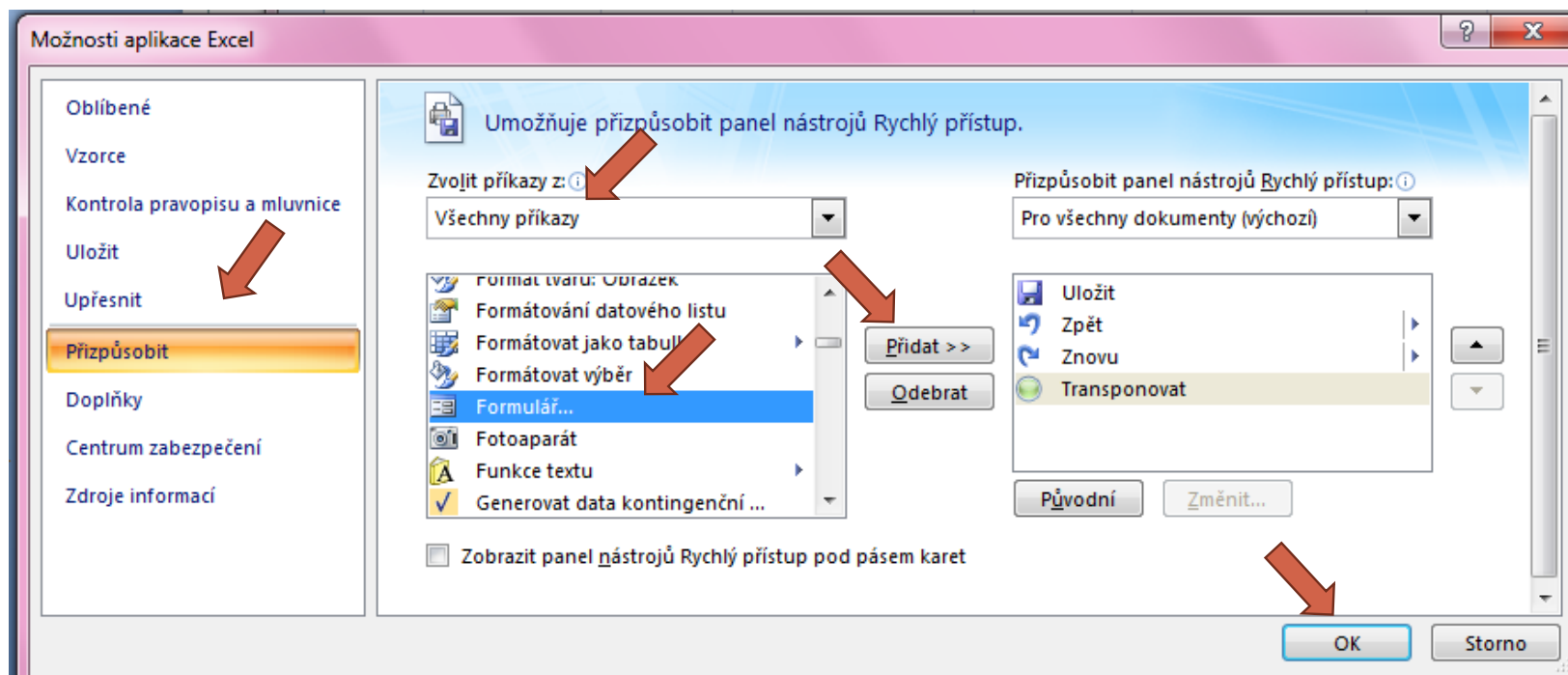
Excel neumožňuje pojmenování řádků a sloupců vlastními názvy.

Automatický zadávací formulář

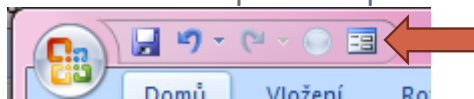


- Aplikaci automaticky zadávaného formuláře je nutné aktivovat

○ „Tlačítko Office“ → „Možnosti aplikace Excel“



- Automatický zadávací formulář spustíme pomocí nové ikonky na panelu nástrojů Rychlý přístup; dále stejné



Automatické seznamy



- Vytváří se z hodnot buněk v daném sloupci a umožňují vložit hodnotu výběrem ze seznamu již zadaných hodnot – usnadnění zadávání

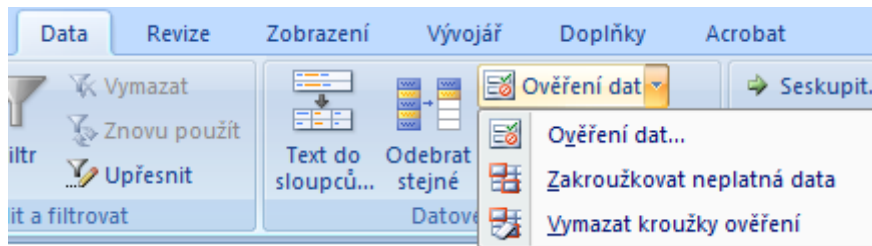
Sloupec z něž je seznam vytvořen a pro který platí

The image shows a spreadsheet with columns labeled 'Taxon', 'Abundance', 'Lokalita', and 'etc.'. A context menu is open over the 'Taxon' column, with the option 'Vybrat ze seznamu...' (Select from list...) highlighted. An arrow points from this menu option to a dropdown list of taxonomic names. The list includes: 'Piscicola geometra (Linnaeus, 1761)', 'Acanthocephallus lucii (Müller, 1776)', 'Apophallus mühlungi Jägerskiöld, 1899', 'Argulus foliaceus (Linnaeus, 1758)', 'Caryophyllaeides fennica (Schneider, 1902)', 'D. cabaleroi', 'D. crucifer Wagener, 1857', 'D. fallax Wagener, 1857', and 'D. nanus Dogiel et Bychowsky, 1934'. Another arrow points from the text 'Buňka, do níž se vloží vybraná hodnota' (Cell into which the selected value will be placed) to the 'Taxon' cell in the spreadsheet.

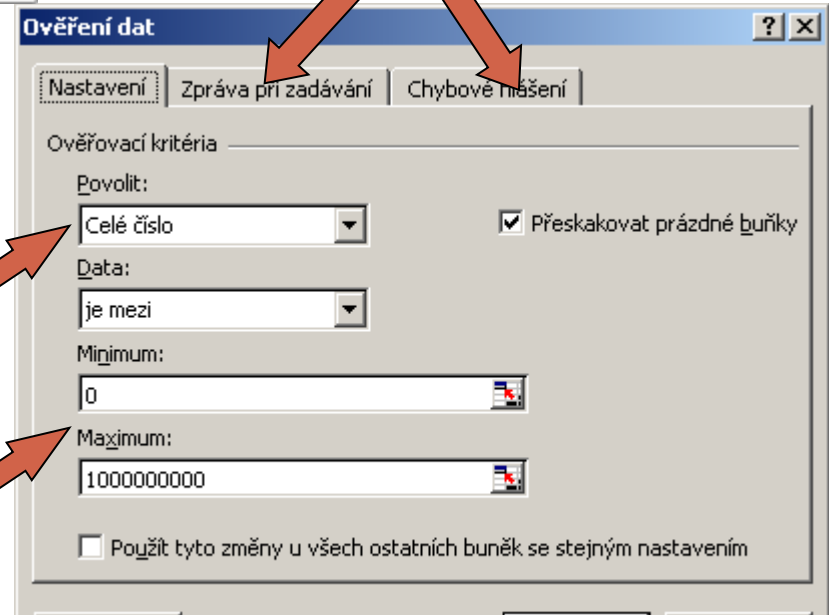
Automatická kontrola dat



- Umožňuje ověřit typ, rozsah nebo povolit pouze určitý seznam hodnot zadávaných do sloupce databázové tabulky



komunikace s uživatelem



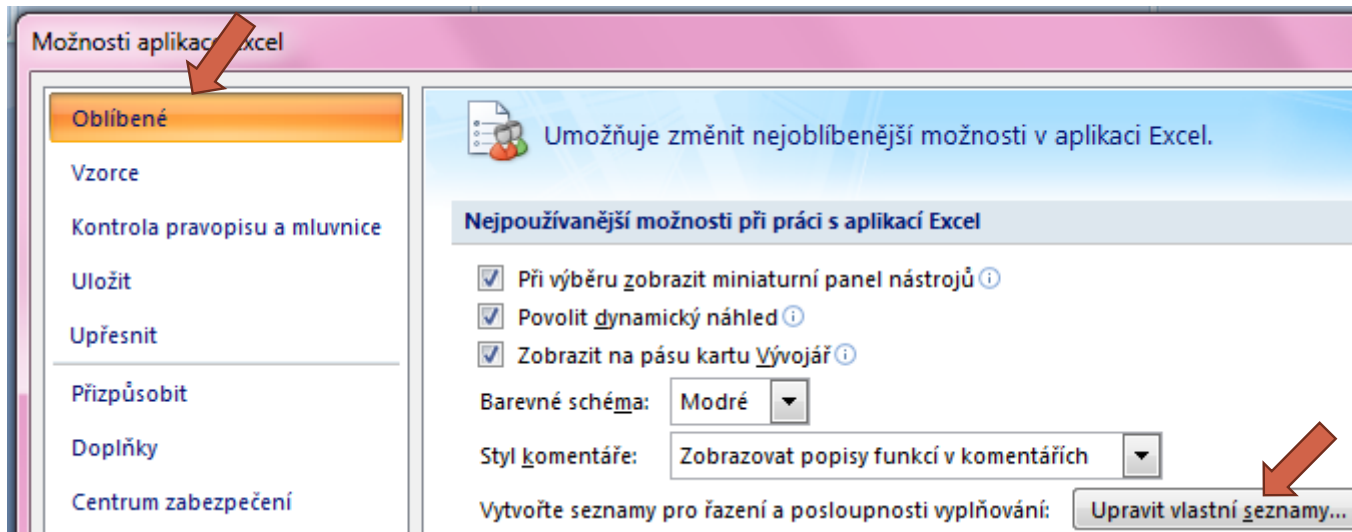
Co je povoleno – definiční obory čísel, seznamy, vzorce atd.

Rozsahy hodnot, načtení seznamů apod.

Seznamy



- „Tlačítko Office“ → „Možnosti aplikace Excel“



- Vlastní seznamy dále stejné (viz předchozí slide)

Řazení dat

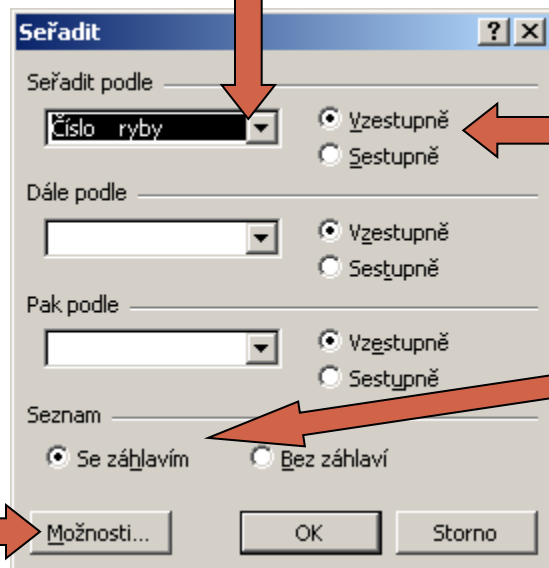
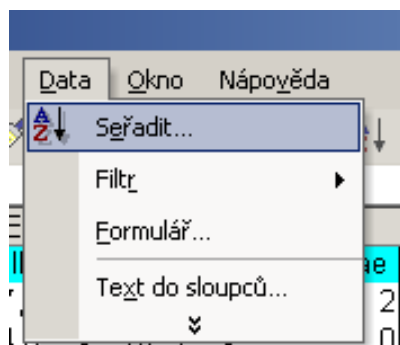


- Řazení dat je nejjednodušším způsobem jejich zpřehlednění, užitečným hlavně u menších/výsledkových tabulek



Zkontrolujte, zda seřazení nezničí vazby mezi buňkami = kontrola oblasti, kterou řadíte.

Podle čeho řadit



Směr řazení – vzestupně, sestupně

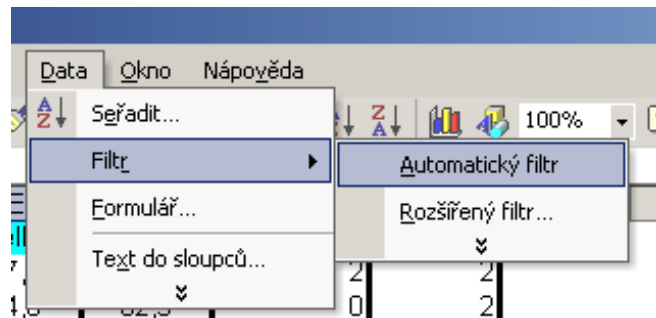
Využít první řádek oblasti jako záhlaví

Další možnosti – řazení řádků, řazení podle seznamu

Automatický filtr



- Pomocí automatického filtru je snadné vybírat úseky dat pro další zpracování na základě hodnot ve sloupcích databázové tabulky, výběr je možný i podle více sloupců (např. určitá skupina pacientů).
- Funkce automaticky rozezná hlavičky sloupců v souvislé oblasti buněk.
- Čísla filtrovaných řádků jsou zobrazena modře.
- **Výhodné pro čištění dat (vyhledávání překlepů, kombinace textu a čísel).**



Rozbalení seznamu hodnot nalezených ve sloupci



Výběr hodnot pro filtraci

	A	B	C	D	E
	Číslo	Značka	Společ	ohlav	Délka
1		1	1	(Vše)	27,5
2		2	2	(Prvních 10...)	34,0
3		5	3	(Vlastní...)	58,0
4	3	6	f	f?	42,0
5	4	7	m		44,0
6	5	8		f	56,0
7	6	9		m	48,5
8	7				

Rozšířený filtr



- Funguje podobně jako automatický filtr, ale seznam povolených hodnot není nutné vybírat ručně – je uveden v oblasti jinde na listu (nebo i na jiném listu).
- Podmínkou jsou shodná záhlaví filtrované oblasti a oblasti povolených hodnot.
- Prázdné buňky odpovídají prázdné podmínce – tj. je-li v oblasti povolených hodnot nějaká buňka prázdná, splní podmínku libovolná buňka filtrované oblasti.
- Čísla řádků filtrované oblasti jsou zobrazena modře.

Tlačítko Upřesnit na kartě Data

Upřesnit

Chcete-li omezit záznamy, které budou zahrnuty ve výsledné sadě dotazu, zadejte složitá kritéria.

D
ParazitZkratka
Gsp
DH

Rozšířený filtr

Akce

Přímě v seznamu

Kopírovat jina

Oblast seznamu:

Oblast kritérií:

Kopírovat do:

Bez duplicitních záznamů

OK Storno

Výběr oblasti cílových hodnot (přefiltrovaných)

Původní seznam včetně záhlaví

Oblast kritérií včetně záhlaví

Automatické dokončování hodnot buněk



- Vhodné pro textová pole; následně není nutné vypisovat celé slovo či slovní spojení, ale jen zvolit nabízené, již dříve použité slovo či slovní spojení
- Automatické dokončování hodnot buněk je nutné nastavit
 - „Tlačítko Office“ → „Možnosti aplikace Excel“



→ „Možnosti aplikace Excel“

