I. Příprava dat

Klíčový význam korektního uložení získaných dat

Pravidla pro ukládání dat

Čištění dat před analýzou

Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek, M. Cvanová

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	Backgrour	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, I	Air	pas	PCB 138	2008	7	29		pg/m3	4,4	0,5		REC	no	yes
7	CEEC	Armenia	Yerevan, I	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, I	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-HCH	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-HCH	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

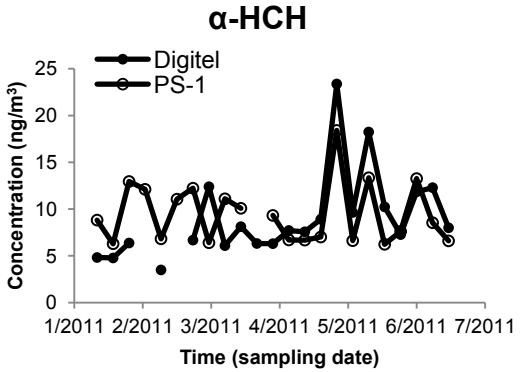
- Správné a přehledné uložení dat je základem jejich pozdější analýzy.
- Je vhodné rozmyslet si předem jak budou data ukládána.
- Pro počítačové zpracování dat je vhodné ukládat data v tabulární formě.
- Nejvhodnějším způsobem je uložení dat ve formě databázové tabulky:
  - každý sloupec obsahuje pouze 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.

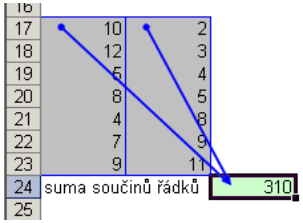
MS Excel

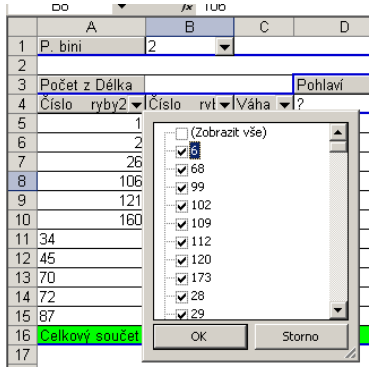
- Tabulkový procesor.
- První verze programu 30. 9. 1985 (Macintosh).
- Součást balíku kancelářských aplikací MS Excel.
- Aktualizace každé 2 až 3 roky; nové funkce, rozšíření počtu řádků a sloupců, změna formátu.
- Nejnovější formát Office XML je zazipovaný XML dokument, přípona .xlsx.
- Aktuální verze 2013 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 Aplications).







Import a export dat

- **Import dat**
  - manuální zadávání;
  - import – 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.

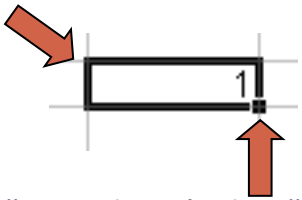
- **Export dat**
  - ukládáním souborů ve formátech podporovaných jinými SW, časté jsou textové soubory, dbf soubory nebo starší verze Excelu;
  - přímé kopírování přes schránku Windows.

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 jsoucí názvy měsíců.

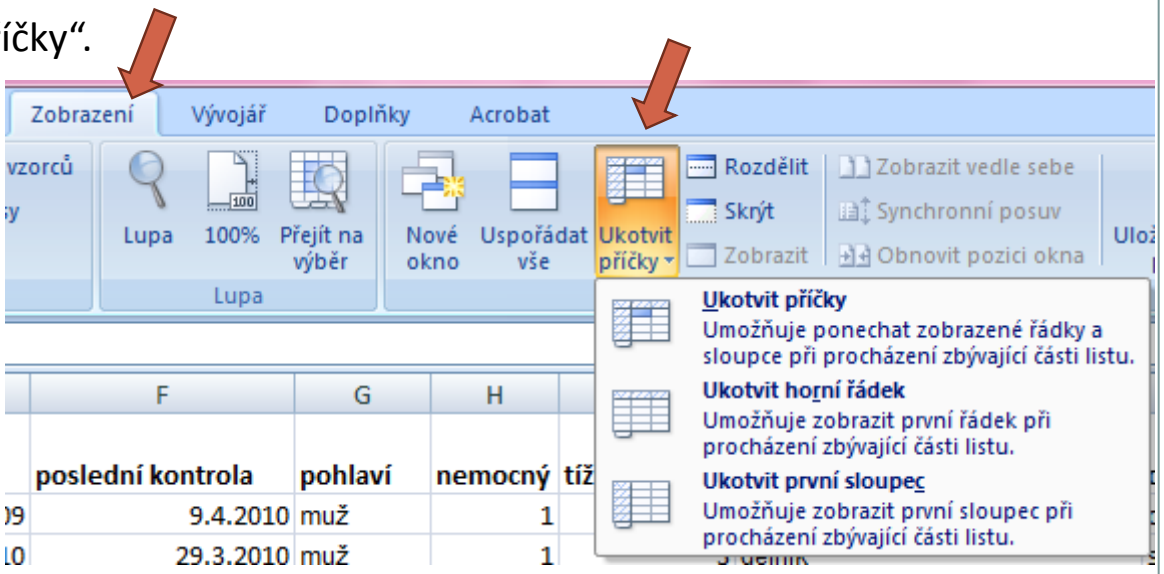
Ukotvení příček

- 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.

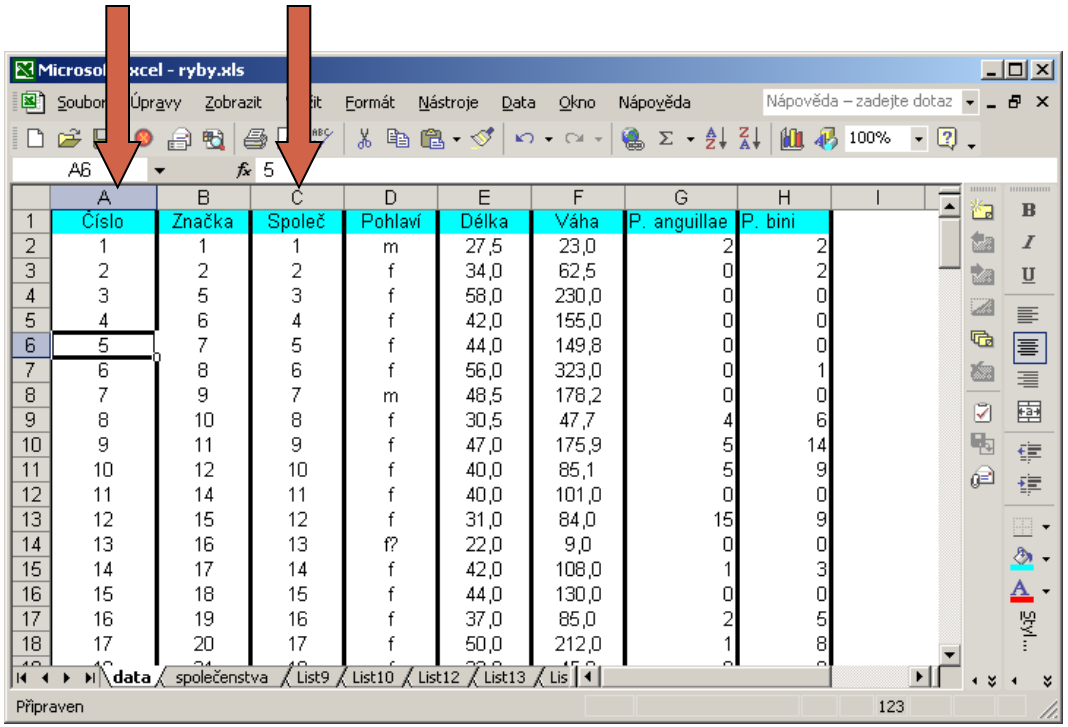
  - Ukotvení zrušíme opětovným odkliknutím možnosti ukotvení příček.



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.)



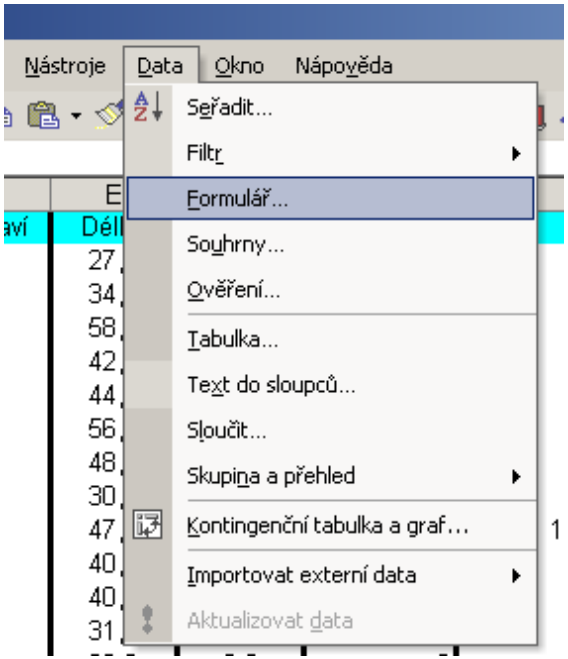
	A	B	C	D	E	F	G	H
1	Číslo	Značka	Společ	Pohlaví	Délka	Váha	P. anguillae	P. bini
2	1	1	1	m	27,5	23,0	2	2
3	2	2	2	f	34,0	62,5	0	2
4	3	5	3	f	58,0	230,0	0	0
5	4	6	4	f	42,0	155,0	0	0
6	5	7	5	f	44,0	149,8	0	0
7	6	8	6	f	56,0	323,0	0	1
8	7	9	7	m	48,5	178,2	0	0
9	8	10	8	f	30,5	47,7	4	6
10	9	11	9	f	47,0	175,9	5	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

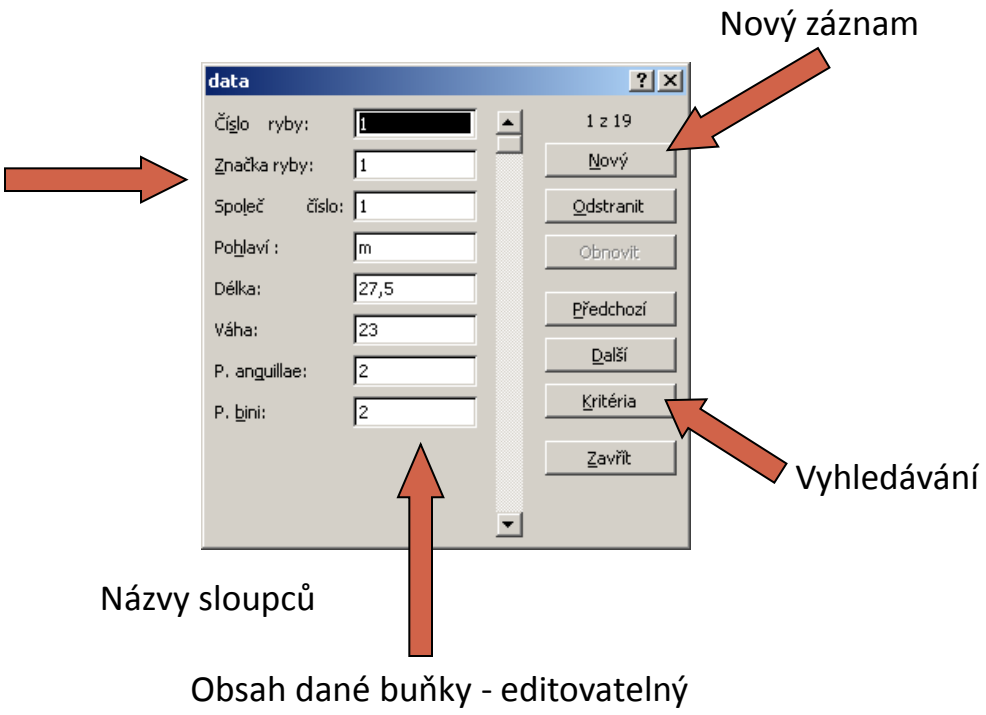
Excel neumožňuje pojmenování řádků a sloupců vlastními názvy.

Automatický zadávací formulář I.

- Slouží k usnadnění zadávání dat do databázových tabulek
- Načítá automaticky hlavičky sloupců jako zadávané položky

Microsoft Office 2003 a starší

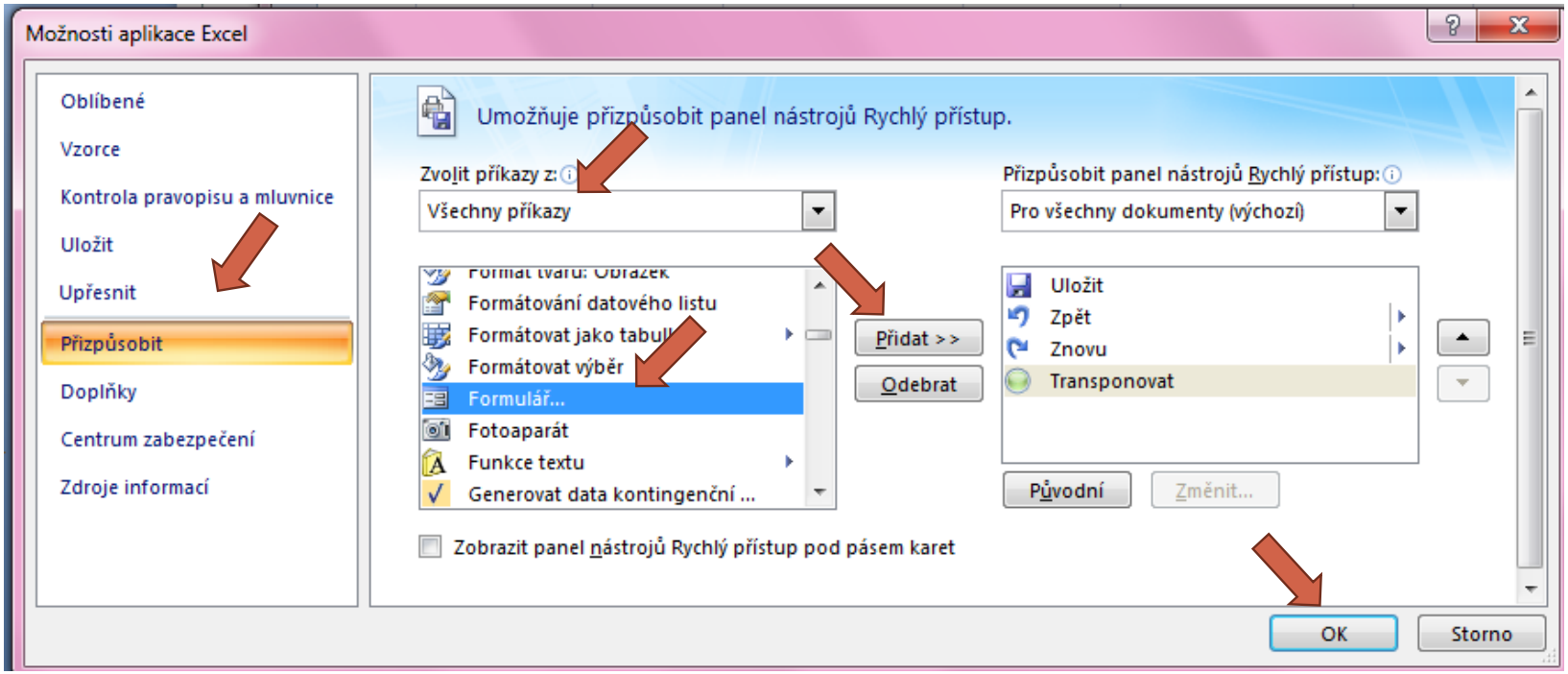




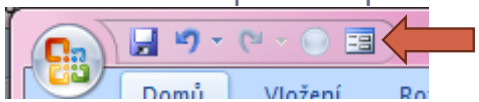
Automatický zadávací formulář II.

Microsoft Office 2007 a novější

- Aplikaci automaticky zadávaného formuláře je nutné aktivovat
  - „Tlačítko Office“  → „Možnosti aplikace Excel“

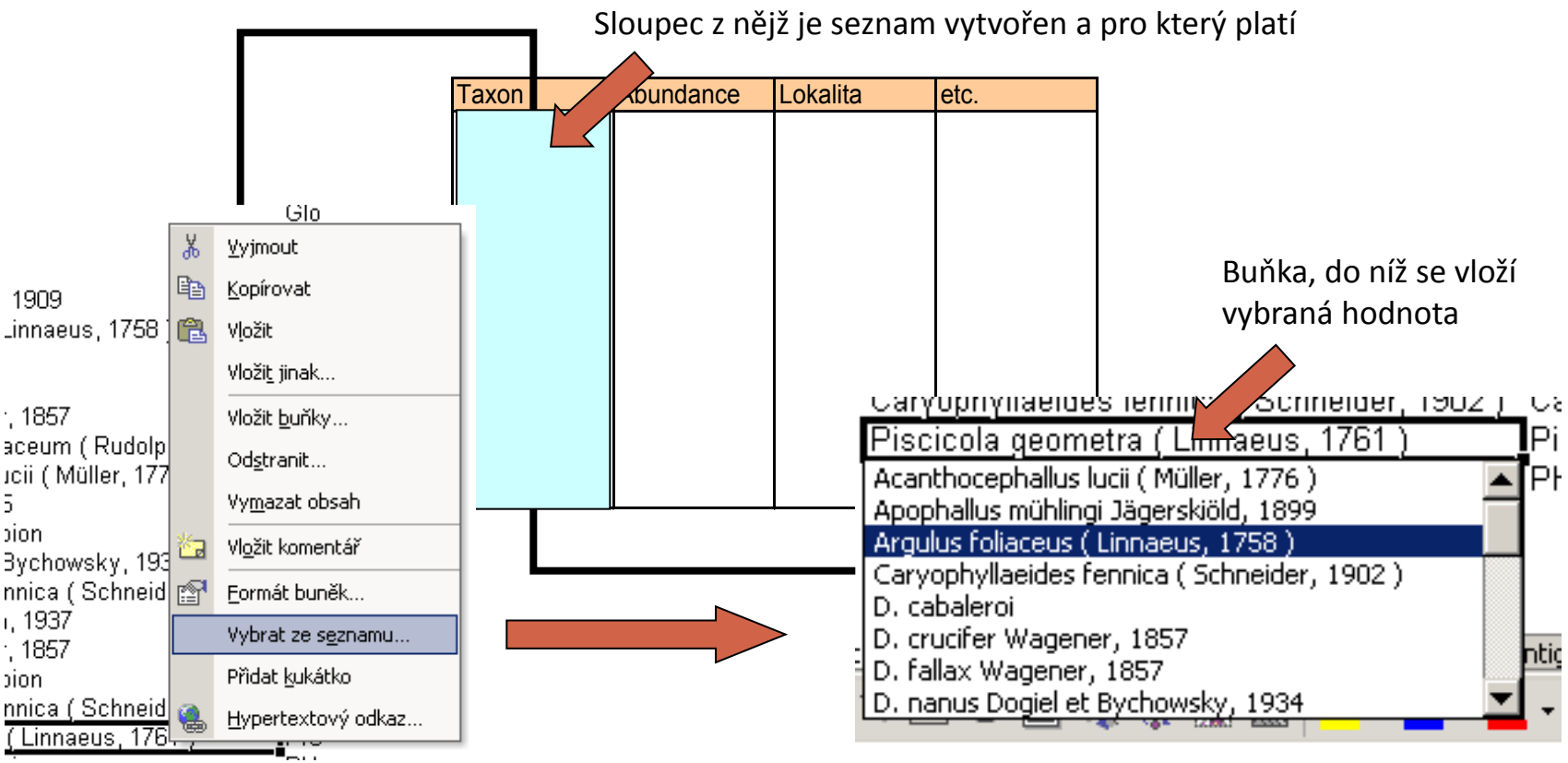


  - Automatický zadávací formulář spustíme pomocí nové ikonky na panelu nástrojů Rychlý přístup; dále stejné



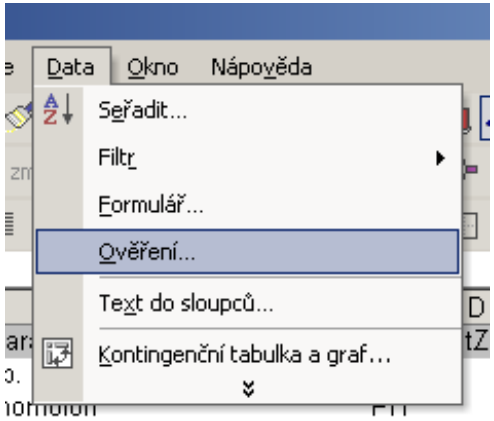
Automatické seznamy

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

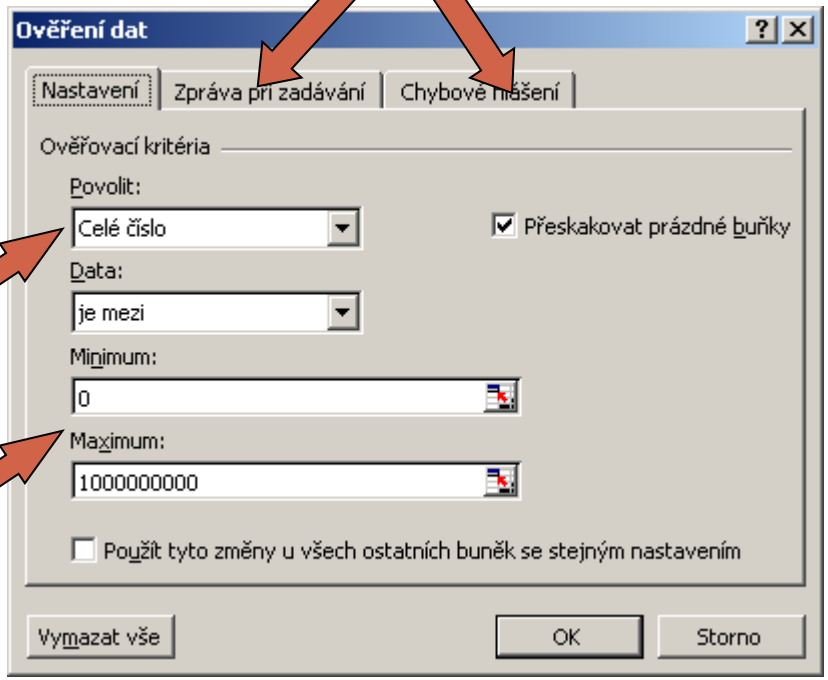


Automatická kontrola dat

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



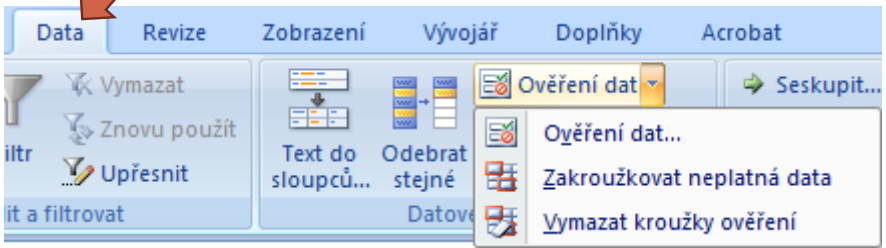
komunikace s uživatelem



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

Microsoft Office 2007

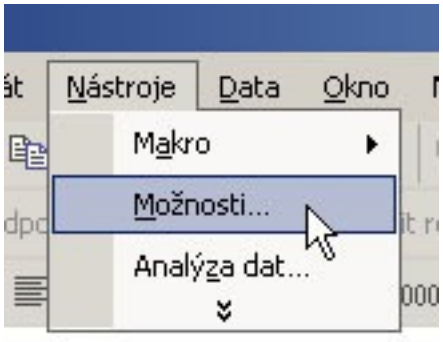
Rozsahy hodnot, načtení seznamů apod.



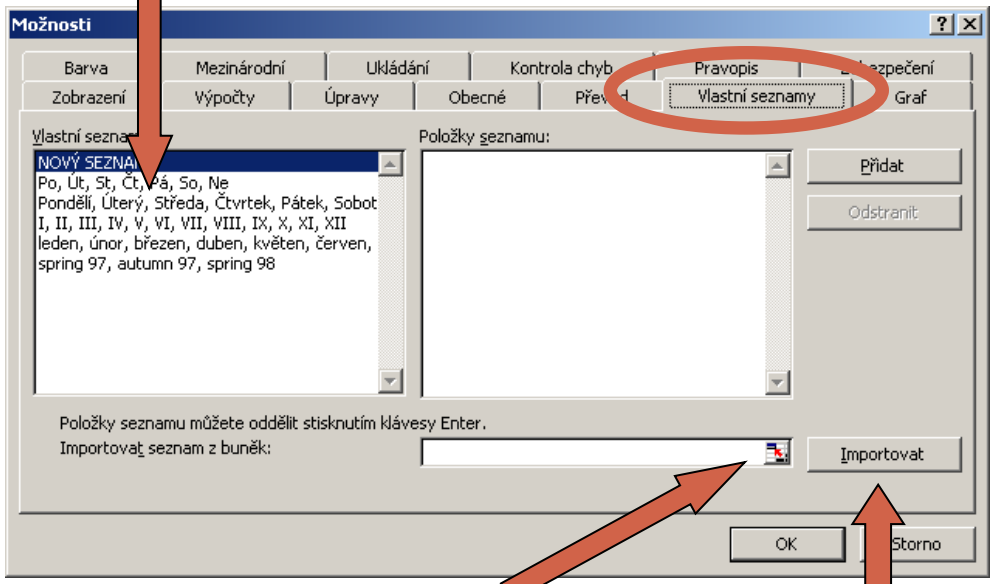
Seznamy I.

- Skupiny hodnot zachovávající logické pořadí, některé jsou zabudované (např. dny v týdnu, měsíce v roce), další je možné uživatelsky vytvořit, slouží pro účely řazení a automatického vyplňování dat

Microsoft Office 2003 a starší



Existující seznamy

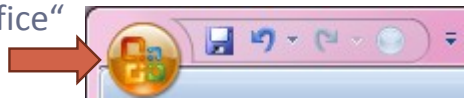


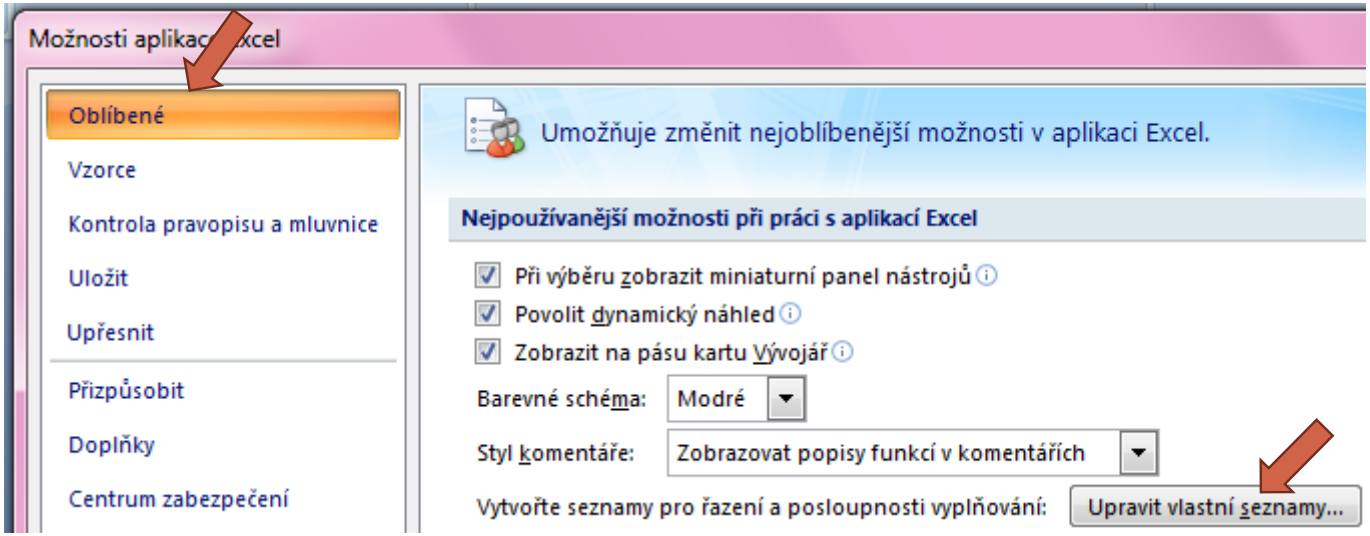
Výběr buněk pro nový seznam

Načtení nového seznamu

Seznamy II.

Microsoft Office 2007

- „Tlačítko Office“  → „Možnosti aplikace Excel“

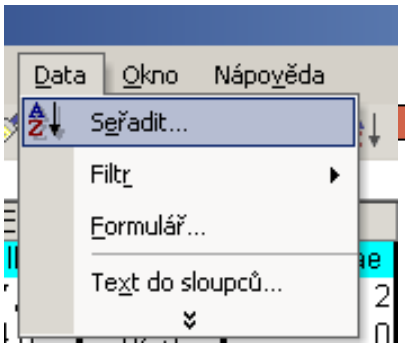


- Vlastní seznamy dále stejné (viz předchozí slide)

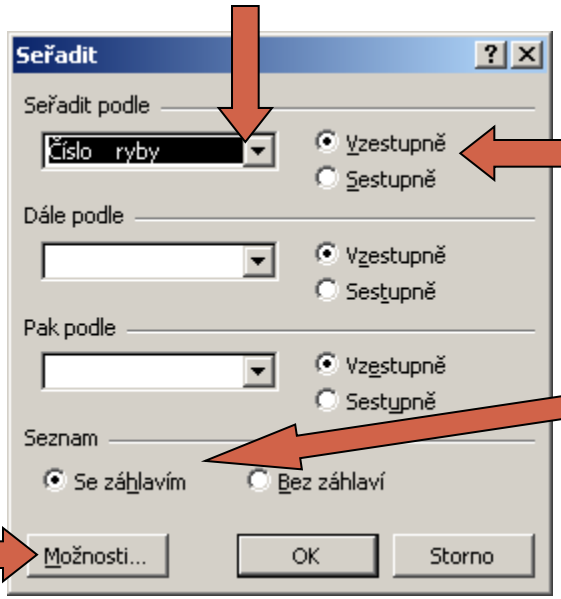
Řazení dat

- 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

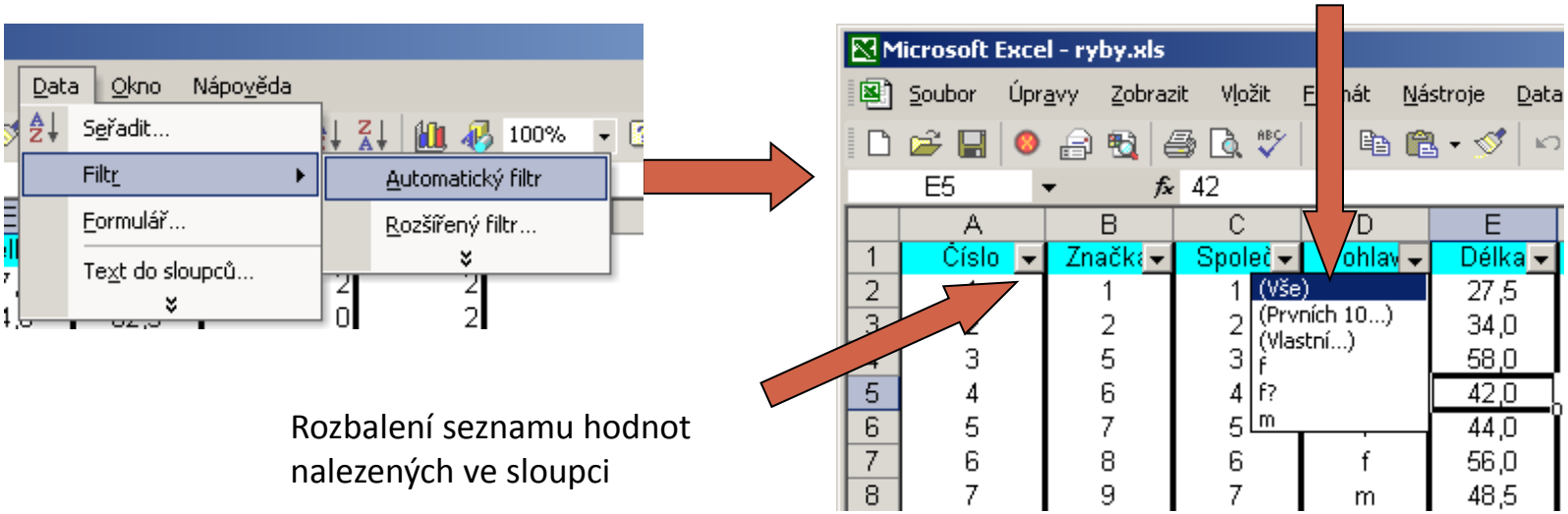
Využít první řádek oblasti jako záhlaví

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).**

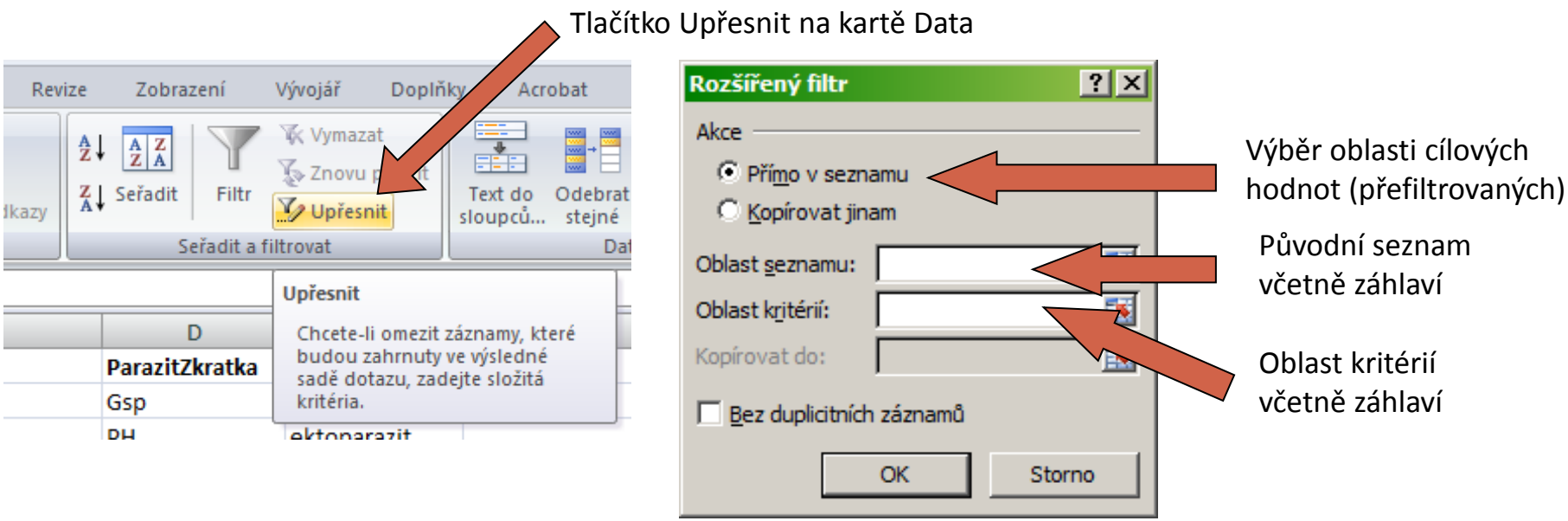
Výběr hodnot pro filtraci



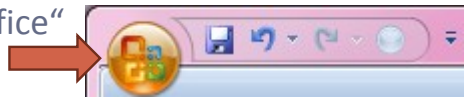
Rozbalení seznamu hodnot nalezených ve sloupci

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.



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“