MUNI
MED

Přednáška II.
Data, jejich popis a vizualizace

Jak vznikají data?

– Záznamem skutečnosti…

… **kterou chceme dále studovat** → smysluplnost?

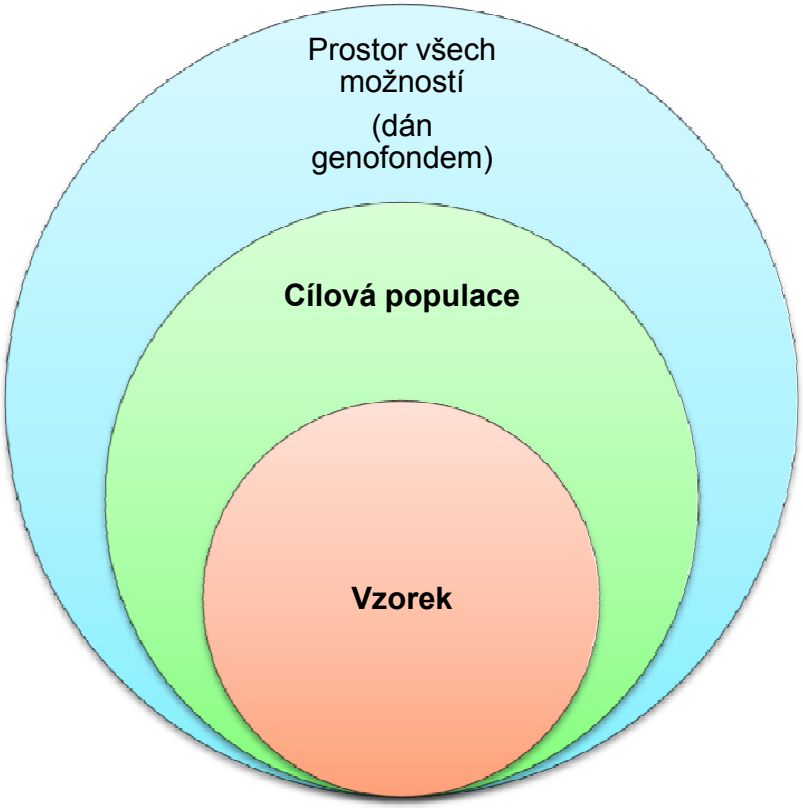
(krevní tlak, glykémie × počet srdcí, počet domů)

… **více či méně dokonalým** → kvalita?

(variabilita = informace + chyba)

Cílová populace, výběrová populace

- **Cílová populace** – skupina subjektů, o které chceme zjistit nějakou informaci. Odpovídá základnímu prostoru Ω.
- **Experimentální vzorek** neboli **výběrová populace** – podskupina cílové populace, kterou pozorujeme, měříme a analyzujeme. Jakékoliv výsledky chceme zobecnit na celou cílovou populaci. **Výběrová populace musí svými charakteristikami odpovídat cílové populaci (reprezentativnost)**. Toho můžeme docílit náhodným, ale i záměrným výběrem.



MUNI MED

Proč je popis a vizualizace dat třeba?

- Chceme **zpřehlednit** pozorovaná data – ve vhodných grafech.
- Chceme **zachytit** případné odlehlé a **extrémní** body nebo nečekané, **nelogické** hodnoty.
- Chceme **popsat** naměřené hodnoty.
- Chceme vypočítat vhodné sumární statistiky, které budou pozorovaná data dále **zastupovat** při prezentaci, srovnáních apod. Chceme pozorovanou informaci „uložit“ v zástupných statistikách, použití všech pozorovaných dat je nepraktické až nemožné.

MUNI
MED

Jaké jsou výstupy popisné analýzy?

- Obecně neformální, jde o **shrnutí pozorovaného** a ne o formální testování.
- **Vztahují se pouze na pozorovaná data** (respektive na experimentální vzorek).
- Mohou sloužit jako **podklad pro stanovení hypotéz**.

MUNI
MED

Příprava dat

- Současná statistická analýza se neobejde bez zpracování dat pomocí statistických software. Předpokladem úspěchu je správné uložení dat ve formě „databázové“ tabulky umožňující jejich zpracování v libovolné aplikaci.
- Neméně důležité je věnovat pozornost čištění dat předcházející 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.

Reálná data

E	F	G	H	I	J	K	L	M	N	O	P
SEX	NHL_STUP	DG_1	DATUM_DG	IPI	LDH	B2M	KS	RT OD	RT_DO	STAV	ZEMREL
F	DLCL	DLCL	28.04.99	0	5.7	1.5	I			KR	
F	DLCL	DLCL	03.11.99	1	13.3	NA	II			ZTR	
F	difusní velkobuněčný B-lymfom	DLCL	19.01.00	2	11.1	2.5	III			EX	31.01.01
F	difusní lymfom z velkých bb	DLCL	27.04.00	0	8.3	2.3	I	12.09.00	13.10.00	KR	
M	centroblastický B-lymfom	DLCL	13.11.00	3	12.6	2.6	III			KR	
M	DLCL	DLCL	15.03.01	0	7.1	3.0	II	25.06.01	18.07.01	KR	
M	DLCL	DLCL	19.04.01	0	5.6	0.2	I			KR	
F	DLCL, bude 2. čtení	DLCL	29.08.01	20	17.9	1.9	II			EX	07.09.02
F	B-velkobuněčný	DLCL	17.10.01	0	9.6	2.1	III	xx.04.02		KR	
M	DLBCL	DLCL	07.02.02	0	8.4	5.6	I			KR	
F	DLCL	DLCL	15.02.02	0	6.5	1.4	II	27.05.02	14.06.02	KR	
Ž	FCL/DLCL	DLCL	20.02.02	0	8.3	1.3	I			EX	18.05.02
M	DLCL	DLCL	07.06.02	0	6.7	NA	I	22.08.03	20.09.03	PR	
M	difusní velkobuněčný B lymfom	DLCL	25.10.02	1	8.2	2.5	III			KR	
M	DLBCL	DLCL	31.01.03	1	13.8	1.8	II	plánovaná		KR	
F	DLBCL	DLCL	06.08.03	2	9.2	1.7	III			KR	
F	DLBCL	DLCL	05.09.03	1	7.3	1.7	III			KR	
F	DLCL	DLCL	03.03.99	1	8.8	1.5	I	20.07.99	16.08.99	KR	
M	DLCL	DLCL	17.08.00	1	8.8	2.0	I	27.02.01	26.03.01	KR	
M		DLCL	Motol	2	8	2.7	III			KR	
M	DLCL	DLCL	19.02.01	1	9.8	2.4	II			KR	
M	DLCL	DLCL	13.03.01	1	16.1	2.0	I	24.10.01	21.11.01	KR	
F	difuzní B-lymfom, HG	DLCL	15.06.01	0	5.7	3.2	II	26.11.01	21.12.01	KR	
F		DLCL		1	11.4	2.0	I			EX	08.01.05
F	difusní velkobuněčný B lymfom	DLCL	01.07.02	2	32.0	6.0	I	28.01.03	10.02.03	EX	27.5.2003
M		DLCL	Motol	0	5.2	1.9	I	21.1.2003	20.2.2003	KR	
	DLBCL	DLCL	07.02.03	0	5.9	2.3	I			PR	
F	DLBCL	DLCL		3	10.6	1.25	IV			KR	
M	DLBCL	DLCL	28.04.99	1	8.4	2.2	II			KR	15.11.02
M	DLBCL	DLCL	05.05.99	2	23.3	4.1	IV			EX	14.06.00

MUNI MED

MUNI
MED

Příprava dat, MS Excel

Datová tabulka
Zásady správné tvorby dat
Možnosti MS Excel

Ukázka datového souboru

Parametry (znaky)

Základní jednotka dat

	A	AC	AD	AE	AF	AG	AH	AI	AJ
1	ID	obvod_pasu_po	obvod_boku_po	WHR_po	WHR_riziko_po	syst_tlak_po	diast_tlak_po	hypertenze_po	cholesterol_po
2	1	86,7	98,6	0,88	1	103	68	0	3,82
3	2	70,3	82,9	0,85	1	118	75	0	6,18
4	3	61,2	88,3	0,69	1	114	74	0	3,90
5	4	81,6	87,3	0,93	1	127	73	0	5,06
6	5	89,2	104,2	0,86	3	135	99	1	6,24
7	6	74,2	100,1	0,74	1	111	81	0	3,44
8	7	114,2	108,5	1,05	3	136	80	0	4,17
9	8	65,1	82,5	0,79	1	118	98	1	2,87
10	9	81,5	79,0	1,03	3	116	87	0	4,20
11	10	75,9	124,9	0,61	1	125	82	0	4,12
12	11	89,8	111,6	0,80	2	108	84	0	2,83
13	12	67,8	87,9	0,77	1	120	64	0	4,71
14	13	95,7	92,7	1,03	3	169	112	1	4,67
15	14	86,1	88,5	0,97	2	111	82	0	4,91
16	15	86,4	101,2	0,85	1	108	79	0	4,99
17	16	68,1	86,3	0,79	1	122	81	0	4,83
18	17	65,0	77,9	0,83	1	139	81	0	4,17
19	18	92,5	72,8	1,27	3	143	77	1	5,89
20	19	69,4	81,6	0,85	2	130	90	0	5,90
21	20	93,7	90,8	1,03	3	136	83	0	4,64
22	21	71,4	83,5	0,86	3	123	83	0	4,75
23	22	95,0	95,4	1,00	3	141	91	1	4,

MUNI MED

Zásady pro ukládání dat

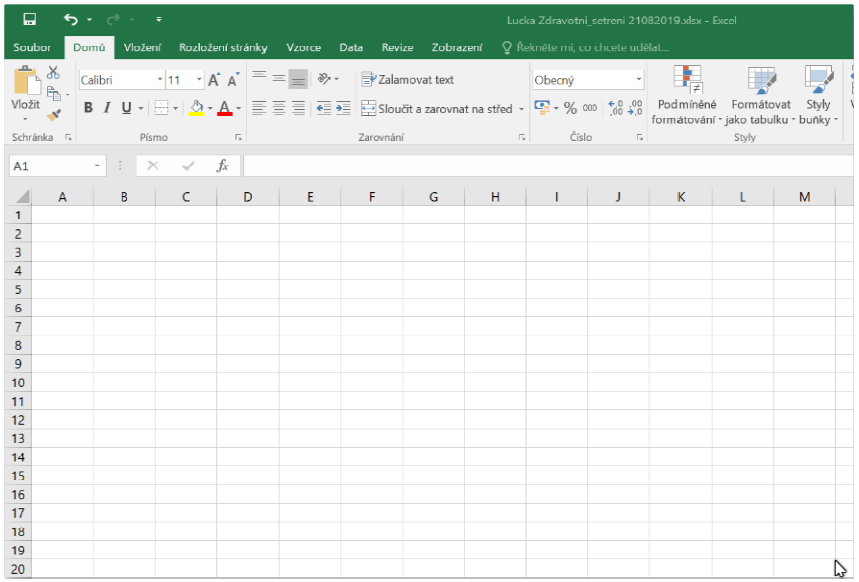
- Správné a přehledné uložení dat je základem jejich pozdější analýzy.
- Je vhodné rozmyslet si předem jak budou data ukládána.
- Pro počítačové zpracování dat je nezbytné ukládat data v tabulární formě.
- Nejvhodnějším způsobem je uložení dat ve formě databázové tabulky.
- Takto uspořádaná data je v tabulkových nebo databázových programech možné převést na libovolnou výstupní tabulku.
- Pro základní uložení a čištění dat menšího rozsahu je možné využít aplikací MS Excel.

Zásady pro ukládání dat

- Každý sloupec obsahuje pouze jediný typ dat, identifikovaný hlavičkou sloupce;
- Každý řádek obsahuje minimální jednotku dat (např. pacient, jedna návštěva pacienta apod.);
- Je nepřípustné kombinovat v jednom sloupci číselné a textové hodnoty;
- Komentáře jsou uloženy v samostatných sloupcích;
- U textových dat je nezbytné kontrolovat překlepy v názvech kategorií;
- Specifickým typem dat jsou data, u nichž je nezbytné kontrolovat, zda jsou uloženy v korektním formátu.

MS Excel

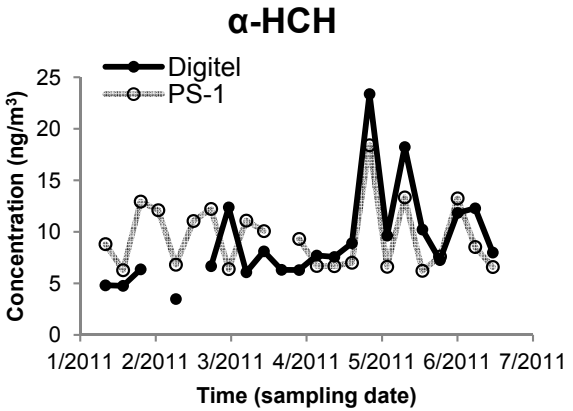
- Tabulkový procesor.
- Aktualizace každé 2 až 3 roky; nové funkce, rozšíření počtu řádků a sloupců, změna formátu.
- Starší formát: .xls, novější: .xlsx.
- Aktuální verze 2016 umožňuje ukládat tabulku
  o 1 048 576 řádcích
  a 16 384 sloupcích.

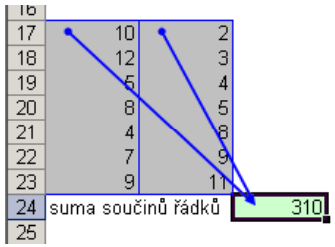


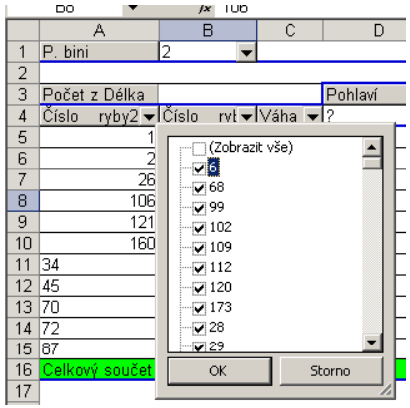
MUNI MED

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.







MUNI MED

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 atd.

Export dat

- Ukládáním ve formátech podporovaných jinými SW, časté jsou textové soubory, dbf soubory nebo starší verze Excelu
- Přímé kopírování přes schránku Windows

Databázová struktura dat v Excelu

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

Řádky tabulky =>
jednotlivé záznamy
(taxon, lokalita,
měření, pacient atd.)

	A	AC	AD	AE	AF	AG	AH	AI	AJ
1	ID	obvod_pasu_po	obvod_boku_po	WHR_po	WHR_riziko_po	syst_tlak_po	diast_tlak_po	hypertenze_po	cholesterol_po
2	1	86,7	98,6	0,88	1	103	68	0	3,82
3	2	70,3	82,9	0,85	1	118	75	0	6,18
4	3	61,2	88,3	0,69	1	114	74	0	3,90
5	4	81,6	87,3	0,93	1	127	73	0	5,06
6	5	89,2	104,2	0,86	3	135	99	1	6,24
7	6	74,2	100,1	0,74	1	111	81	0	3,44
8	7	114,2	108,5	1,05	3	136	80	0	4,17
9	8	65,1	82,5	0,79	1	118	98	1	2,87
10	9	81,5	79,0	1,03	3	116	87	0	4,20
11	10	75,9	124,9	0,61	1	125	82	0	4,12
12	11	89,8	111,6	0,80	2	108	84	0	2,83
13	12	67,8	87,9	0,77	1	120	64	0	4,71
14	13	95,7	92,7	1,03	3	169	112	1	4,67
15	14	86,1	88,5	0,97	2	111	82	0	4,91
16	15	86,4	101,2	0,85	1	108	79	0	4,99
17	16	68,1	86,3	0,79	1	122	81	0	4,83
18	17	65,0	77,9	0,83	1	139	81	0	4,17
19	18	92,5	72,8	1,27	3	143	77	1	5,89
20	19	69,4	81,6	0,85	2	130	90	0	5,90
21	20	93,7	90,8	1,03	3	136	83	0	4,64
22	21	71,4	83,5	0,86	3	123	83	0	4,75
23	22	95,0	95,4	1,00	3	141	91	1	4,[illegible]

Excel neumožňuje pojmenování řádků a sloupců vlastními názvy.

MUNI MED

Typy a triky jak se v datech pohybovat

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

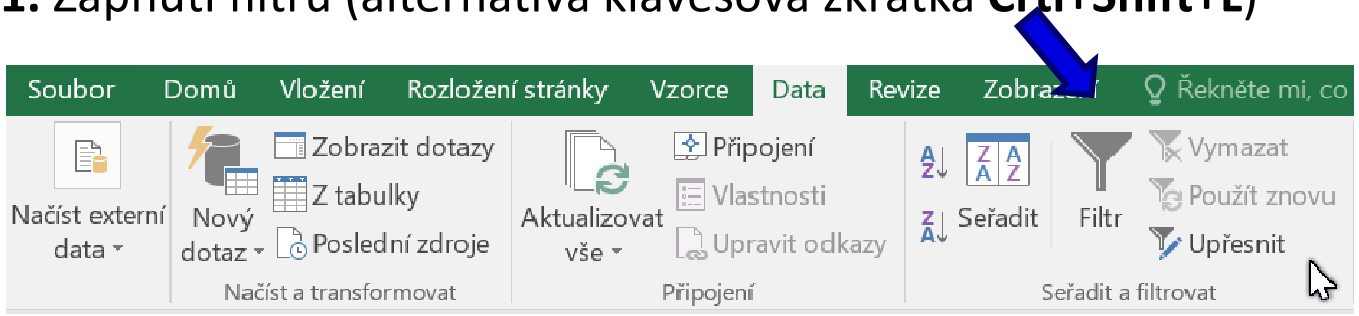
Myš a okraje buňky

- 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)

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
- **Výhodné pro čištění dat (vyhledávání překlepů, kombinace textu a čísel)**

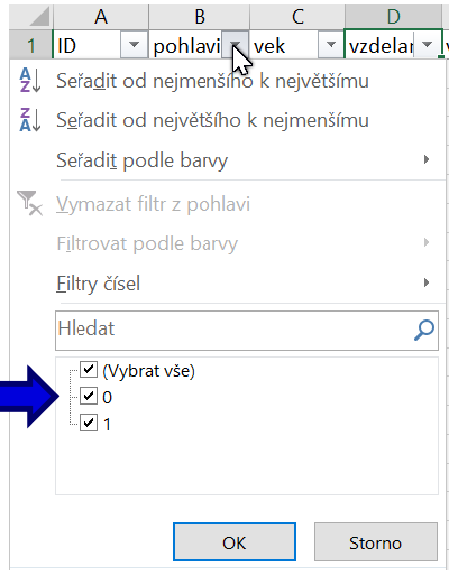
1. Zapnutí filtru (alternativa klávesová zkratka **Crtl+Shift+L**)



2. Objeví se rozbalovací šipka s výčtem všech unikátních hodnot v daném sloupci dat

	A	B	C	D	E
1	ID	pohlavi	vek	vzdelar	vyska
2	1	0	55	1	182
3	2	0	56	2	169
4	3	1	59	3	169

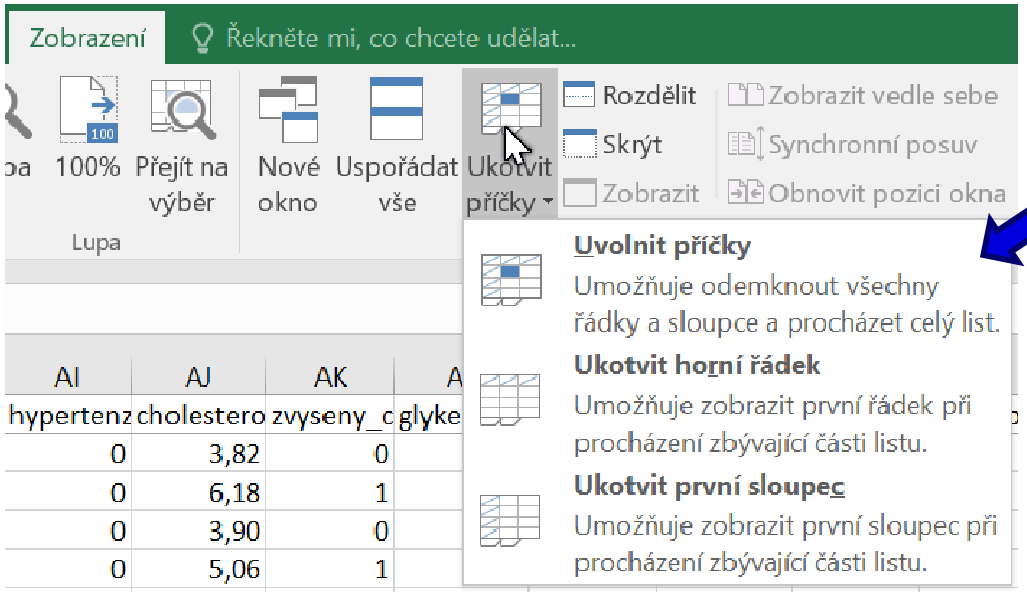
Výběr hodnot pro filtraci



MUNI MED

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“.
- Odstranění ukotvení: Po ukotvení příček se automaticky možnost „Ukotvit příčky “ změní na „Uvolnit příčky“.
- Možnosti:

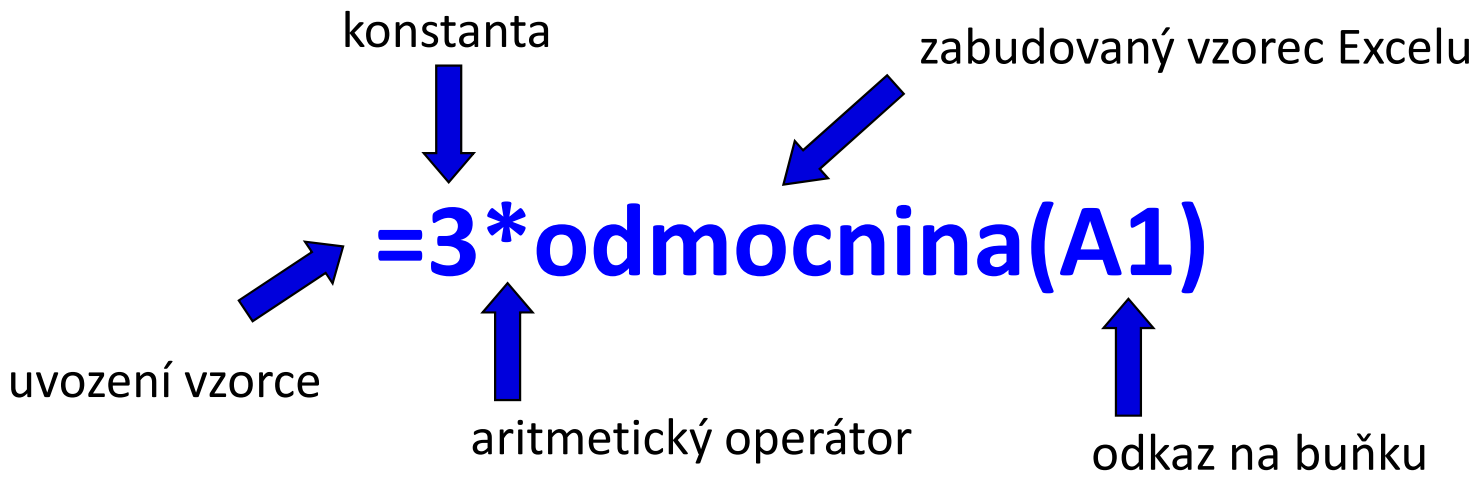


Ukotví řádky nad označenou buňkou a sloupce vlevo od označené buňky

MUNI MED

Vzorce

- vpisují se do buněk sešitu
- vzorce jsou vždy uvozeny **=** (lze též **+ -**)
- aritmetické operátory + zabudované funkce Excelu
- pro „sčítání“ nečíselných položek se používá &
- výpočet je založen buď na číselných konstantách nebo odkazech na buňky



MUNI MED

Vzorce – odkaz na buňku

Relativní odkazy

- **A1** = buňka 1. řádku sloupci A
- **A1:B6** = blok buněk – levý horní roh je v 1. řádku, sloupec A,pravý dolní na řádku 6, sloupec B
- relativní odkaz se při automatickém vyplnění buněk vzorcem posune
- mění se s kopírováním, při vložení a odstranění řádku nebo sloupce

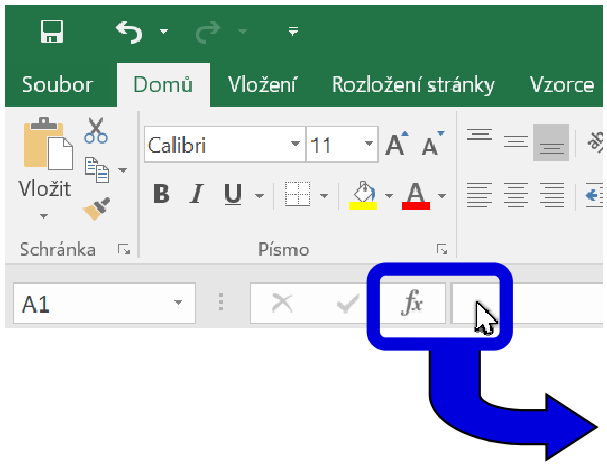
Absolutní odkaz

- odkaz na buňku je pevně dán, při kopírování nebo automatickém vyplnění se nemění
- lze uzamknout jak řádky, tak sloupce samostatně

Pamatuj: Adresu upevníme pomocí znaku **$** (klávesa **F4**)

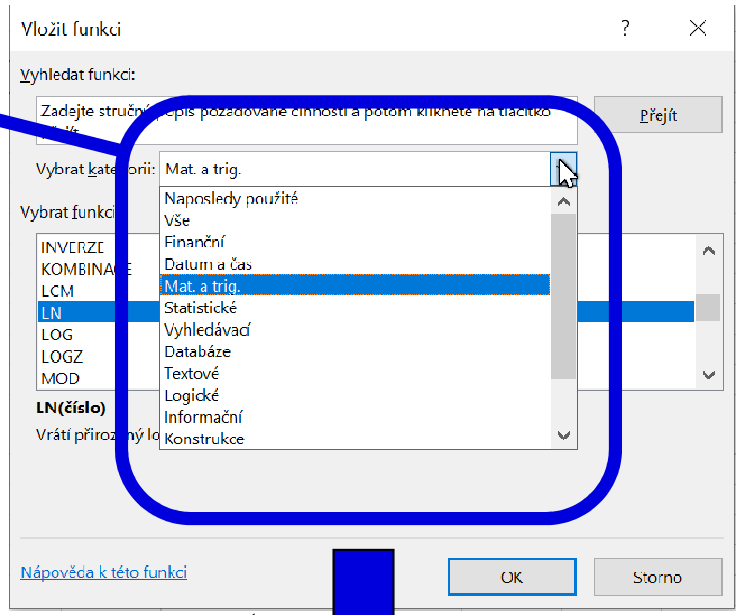
MUNI MED

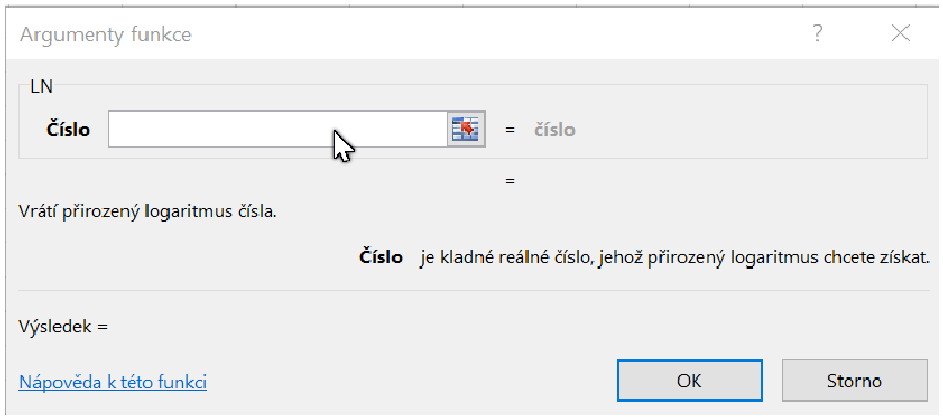
Vzorce – využití seznamu vzorců



Funkce a její stručný popis

Kategorie vzorců





Průvodce funkcí

MUNI MED

Vzorce – užitečné funkce

- **SUMA** – součet číselných hodnot oblasti;
- **SUMIF** – podmíněný součet (podmínky v doplňkové oblasti);
- **PRŮMĚR** – aritmetický průměr číselných hodnot oblasti;
- **GEOMEAN** – geometrický průměr číselných hodnot oblasti;
- **COUNTIF** – počet hodnot oblasti splňujících zadanou podmínku;
- **KDYŽ** – logická podmínka (IF);
- **MAX, MIN** – maximum/minimum číselných hodnot oblasti;
- **MEDIAN** – výpočet mediánu;
- **PERCENTIL** – výpočet percentilů;
- **DATUAM, ROK, MĚSÍC, DEN** – práce s kalendářními daty;
- **ABS** – absolutní hodnota;
- **SVYHLEDAT** – spojování tabulek podle identifikátoru - řádku.

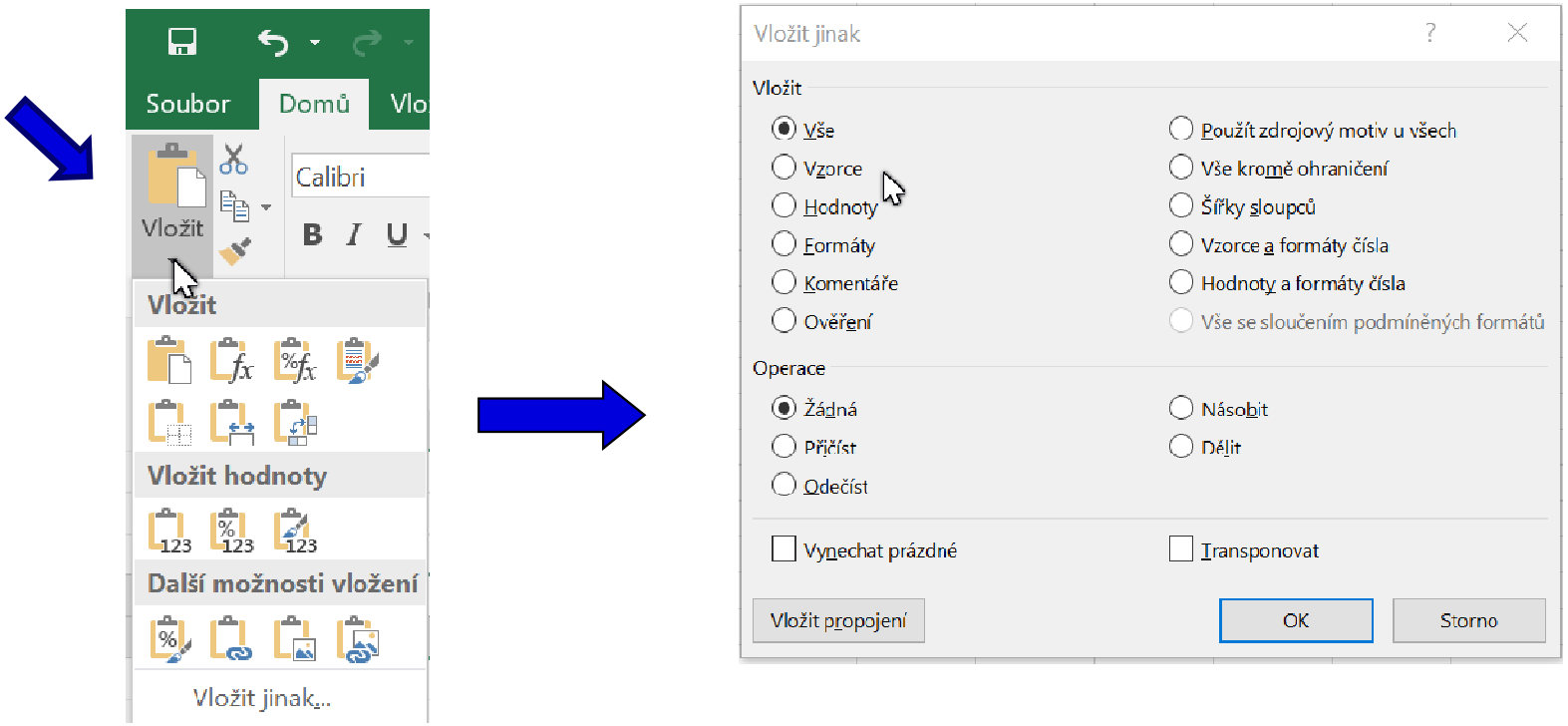
MUNI MED

Statistické funkce v MS Excel

- **CONFIDENCE.NORM** – výpočet intervalu spolehlivosti (při normálním rozdělení);
- **CORREL**, **PEARSON** – výpočet Pearsonova korelačního koeficientu;
- **COVARIANCE.S** – výpočet kovariance dvou množin dat;
- **COUNTIF** – počet hodnot oblasti splňujících zadanou podmínku;
- **DEVSQ** – součet čtverců odchylek od výběrového průměru;
- **F.DIST**, **GAMMA.DIST**, **T.DIST**, **NORM.DIST** aj. – různá rozdělení pravděpodobnosti;
- **PRŮMODCHYLKA** – průměrná hodnota absolutních odchylek;
- **SLOPE** – směrnice lineárního modelu;
- **T.TEST**, **Z.TEST**, **CHISQ.TEST** – statistické testy shodnosti;
- ŘADU DALŠÍCH FUNKCÍ VŠAK EXCEL POSTRÁDÁ A JE TŘEBA VYUŽÍT SILNĚJŠÍHO NÁSTROJE.

Kopírování a vkládání

– Kopírování vzorců, textů, celých sloupců (zkopírování pomocí Ctrl+C); dále „Vložit jinak...“



MUNI MED

MUNI
MED

Praktické cvičení

Datový soubor

Rehabilitace po mozkovém infarktu

	A	B	C	D	E	F	G	H	
1	ID	Pohlavi	Vek	Etiologie	Lokalizace	Terapie	Komorbidity_komplikace	Barthel_index_pred_rehabilitaci	Kategored
2	1	muž	82	okluze nebo stenóza	mozkové tepny	jiná farmakologická terapie	0	25	vysoce
3	2	žena	81	embolie	mozkové tepny	jiná farmakologická terapie	2	20	vysoce
4	3	muž	55	okluze nebo stenóza	mozkové tepny	jiná farmakologická terapie	0	35	vysoce
5	4	žena	46	embolie	mozkové tepny	intravenózní trombolýza rt-PA	0	20	vysoce
6	5	muž	76	okluze nebo stenóza	mozkové tepny	jiná farmakologická terapie	0	45	částečn
7	6	muž	72	okluze nebo stenóza	mozkové tepny	jiná farmakologická terapie	0	25	vysoce
8	7	muž	62	trombóza	mozkové tepny	jiná farmakologická terapie	0	40	vysoce
9	8	muž	64	trombóza	přívodní tepny	jiná farmakologická terapie	0	15	vysoce
10	9	žena	82	okluze nebo stenóza	mozkové tepny	jiná farmakologická terapie	0	10	vysoce
11	10	muž	58	trombóza	mozkové tepny	jiná farmakologická terapie	0	25	vysoce
12	11	muž	84	okluze nebo stenóza	mozkové tepny	jiná farmakologická terapie	0	40	vysoce
13	12	žena	92	okluze nebo stenóza	mozkové tepny	jiná farmakologická terapie	0	30	vysoce
14	13	žena	79	embolie	mozkové tepny	jiná farmakologická terapie	1	40	vysoce
15	14	muž	69	trombóza	mozkové tepny	jiná farmakologická terapie	3	45	částečn
16	15	muž	67	okluze nebo stenóza	mozkové tepny	mechanická trombektomie	0	25	vysoce
17	16	žena	70	trombóza	přívodní tepny	mechanická trombektomie	0	40	vysoce
18	17	žena	59	trombóza	mozkové tepny	jiná farmakologická terapie	0	25	vysoce
19	18	žena	63	okluze nebo stenóza	přívodní tepny	jiná farmakologická terapie	0	40	vysoce

MUNI MED

Rehabilitace po mozkovém infarktu

- Cvičný datový soubor obsahuje záznamy o **celkem 407 pacientech hospitalizovaných pro mozkový infarkt** na neurologickém oddělení akutní péče, kde jim byla poskytnuta terapie pro obnovu krevního oběhu v postižené části mozku.
- Po zvládnutí akutní fáze byl u pacientů vyhodnocen stupeň soběstačnosti v základních denních aktivitách (ADL) pomocí tzv. **indexu Barthelové** (BI) a byli přeloženi na **rehabilitační oddělení**.
- Po dvou týdnech byl opět dle BI vyhodnocen stupeň soběstačnosti a pacienti byli buď propuštěni do ambulantní péče, nebo přeloženi na oddělení následné péče.

Rehabilitace po mozkovém infarktu

Sbírané informace:

- základní demografické údaje (**pohlaví** a **věk**),
- informace o samotné diagnóze mozkové příhody (**etiologie** a **lokalizace uzávěru cévy**),
- informace o léčbě (typ indikované **terapie** a **výskyt komplikací**)
- informace o **způsobu ukončení rehabilitace**.
- Stupeň soběstačnosti před rehabilitací byl dodatečně zjištěn z neurologie a na konci rehabilitace byl vyplněn nový dotazník pro určení výsledného **indexu Barthelové**.

Úkol č. 1 – kontrola a příprava dat

1. Do všech řádků tabulky vyplňte do sloupce *Barthel_index_reference* hodnotu 64,4.
2. **Ukotvěte** ID pacientů a názvy proměnných ve sloupcích. (**nápověda**: vyber buňku pro levý horní roh → karta „Zobrazení“→ funkce Ukotvit příčky).
3. Zapněte **automatický filtr** nad celou datovou tabulkou a **zkontrolujte přítomnost chybných hodnot** ve sloupcích Pohlavi, Vek, Etiologie, Lokalizace, Terapie. Chybné hodnoty opravte. (**nápověda**: označ všechny sloupce → karta „Data“→ funkce Filtr).

MUNI MED

Úkol č. 1 – kontrola a příprava dat

4. Pomocí **podmíněného formátování** nalezněte **duplicitní záznamy** ID pacientů. Jsou všechny Vámi označené záznamy skutečně duplicitní? Duplicitní údaj smažte. (**nápověda**: označ sloupec → karta „Domů“→ podmíněné formátování → zvýraznit pravidla buněk → duplicitní hodnoty → filtrovat podle barvy).
5. Spočítejte hodnoty ve sloupci *Barthel_index_po_rehabilitaci* jako celkový **součet** dosažených bodů v jednotlivých otázkách Barthelové testu po rehabilitaci. (**nápověda**: prostý součet jednotlivých buněk nebo funkce SUMA(…)).

MUNI MED

Úkol č. 1 – kontrola a příprava dat

6. Spočítejte hodnoty ve sloupci *Barthel_index_zmena* jako **rozdíl** Barthelové indexu před a po rehabilitaci (**nápověda**: prostý vzorec pro rozdíl).
7. Sloupce *Barthel_index_pred_rehabilitaci* a *Barthel_index_po_rehabilitaci* **překódujte** do sloupců *Kategorie_zavislosti_pred_rehabilitaci* a *Kategorie_zavislosti_po_rehabilitaci* následovně: 0 až 40 = vysoce závislý, 45 až 100 = částečně soběstačný. (**nápověda**: pomocí funkce KDYŽ(…)).