III. CVIČENÍ ZE STATISTIKY

Vážení studenti,

úkolem dnešního cvičení je naučit se analyzovat data pomocí chí-kvadrát testu, korelační a regresní analýzy. K tomuto budeme používat program Excel 2007 MS Office, v jehož prostředí jste již pracovali a který je pro Vás snadno dostupný.

Co potřebujete umět? Předpokládám, že umíte pracovat se základními nástroji programu Excel 2007 a že jste se v prvním cvičení ze statistiky naučili vytvářet kontingenční tabulky.

Přeji Vám mnoho úspěchů se studiem této kapitoly.

Cíl dnešního cvičení je naučit se analyzovat data pomocí chí-kvadrát testu, korelační a regresní analýzy. K tomuto budeme používat program Excel 2007 MS Office, v jehož prostředí jste již pracovali a který je pro Vás snadno dostupný.

1. Co je chí-kvadrát test a k čemu jej můžete použít?

V úvodu si jen stručně připomeneme základní fakta z teorie testování hypotéz.

Chí-kvadrát test

chí-kvadrát test je statistická neparametrická metoda, která se používá k zjištění, zda mezi dvěma znaky existuje prokazatelný výrazný vztah.

Znaky mohou být:

- kvalitativní (kategoriální)
- diskrétní kvantitativní
- spojité kvantitativní, ale s hodnotami sloučenými do skupin.

Data uspořádáme do kontingenční tabulky. Kategorie jednoho znaku určují řádky a kategorie druhého znaku sloupce. Jednotlivá pozorování jsou zařazena do příslušné buňky kontingenční tabulky podle hodnot daných dvou znaků. Pokud jeden ze znaků má ***r*** kategorií a druhý znak má ***s*** kategorií, dostáváme kontingenční tabulku typu r x s .

Nejmenší tabulku typu 2 x 2, kterou získáme v případě, že každý znak má pouze dvě kategorie, nazýváme čtyřpolní tabulka.

Kontingenční tabulky umožňují testování různých hypotéz. Tři obvykle testované hypotézy jsou:

- Test homogenity
- Test nezávislosti
- Test dobré shody

Test homogenity – slouží pro porovnání rozložení (distribuce) kvalitativní veličiny ve dvou nebo více populacích.

Test nezávislosti – používá se k posouzení závislosti dvou kvalitativních veličin měřených na prvcích téhož výběru.

Test dobré shody - zjišťuje, zda sledovaná veličina má rozdělení pravděpodobnosti určitého typu.

Základní myšlenka chí-kvadrát testu spočívá v porovnání pozorovaných a očekávaných četností. Pozorované četnosti zjistíme z kontingenční tabulky. Očekávané četnosti je nutné vypočítat. Při výpočtu vycházíme z předpokladu, že platí nulová hypotéza. Tedy provádíme-li test homogenity, předpokládáme, že rozložení hodnot sledované kategoriální veličiny je ve všech populacích shodné. Pokud provádíme test nezávislosti, nulová hypotéza předpokládá, že mezi dvěma kvalitativními veličinami není žádná závislost. V případě testu dobré shody předpokládáme, že sledovaná veličina má rozložení daného typu.

Velikost rozdílů mezi pozorovanými a očekávanými četnostmi posuzujeme pomocí testové statistiky χ^2, její přesný tvar naleznete ve výukových textech. Na základě pravděpodobnostního rozložení chí-kvadrát se vypočítá pravděpodobnost výskytu takovéto nebo ještě extrémnější hodnoty. Tato pravděpodobnost se nazývá dosažená hladina významnosti statistického testu (*p*-hodnota). Pokud je menší než 0,05, nulovou hypotézu zamítáme. Znamená to, že pravděpodobnost, že by pozorované rozdíly či závislosti vznikly pouze náhodou, je menší než 5 %.

2. Jak provést chí-kvadrát test v programu Excel 2007?

V této kapitole si ukážeme postup, který nám umožní testování hypotéz pomocí chí-kvadrát testu.

Abychom mohli k analyzování dat použít výpočetní techniku, je třeba mít data uložená v databázi. Nejběžnější je uložení dat v souboru programu Excel. Data pro naše cvičení jsou uložena na diskové jednotce F: ve složce SOFTWARE. Celá cesta je F:/SOFTWARE/biostatistika/data/analýza dat.xls

Excelovský sešit má 6 listů. První list má název „*chí-kvadrát test*“. Najdete v něm data, která byla zjištěna při preventivní prohlídce 584 zaměstnanců nemocnice. V prvním sloupci (A) **Číslo zaměstnance** je uvedena identifikace zaměstnance. Druhý sloupec (B) **Pohlaví** udává pohlaví zaměstnance (M – muž, Ž – žena), třetí sloupec (C) **Kouření** obsahuje informaci o tom, zda zaměstnanec aktuálně kouří či ne, ve sloupci (D) **ischemie** je zadáno, zda sledovaný jedinec trpí ischemickou chorobou srdeční, ve sloupci (E) **hypertenze** zda trpí zvýšeným krevním tlakem čili hypertenzí, sloupec (F) **BMI** udává hodnocení zaměstnance z hlediska body mass indexu – rozlišujeme zde tři kategorie – norma, nadváha, obezita.

Zadání úkolu

Vaším úkolem bude prověřit závislost mezi pohlavím zaměstnanců a kouřením, výskytem hypertenze resp. výskytem nadváhy či obezity. Jinými slovy se ptáme, zda podíl kuřáků je stejný i mužů i u žen, zda podíl osob s hypertenzí je stejný u obou pohlaví či zda muži i ženy trpí nadváhou a obezitou ve stejné míře.

Stanovíme nulové a alternativní hypotézy:

1. Nulová hypotéza: Podíl kuřáků je stejný u mužů i žen.
   Alternativní hypotéza: Podíl kuřáků u mužů a u žen se liší.

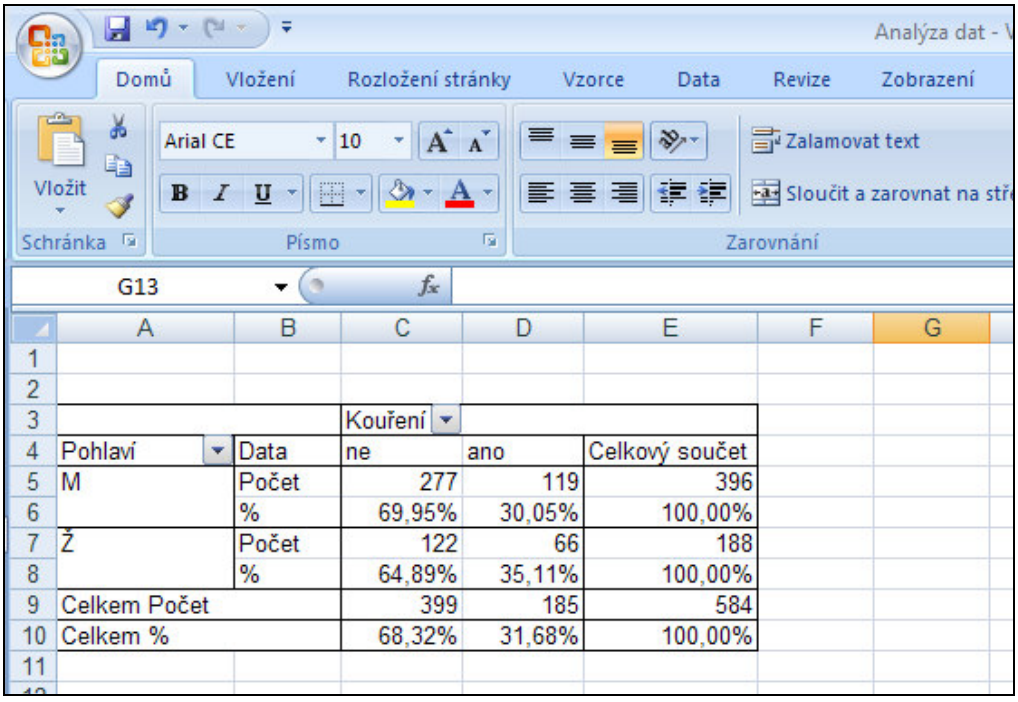
2. Nulová hypotéza: Výskyt hypertenze nezávisí na pohlaví.
   Alternativní hypotéza: Výskyt hypertenze závisí na pohlaví.

3. Nulová hypotéza: Rozdělení zaměstnanců podle BMI je stejné u mužů i žen.
   Alternativní hypotéza: Rozdělení zaměstnanců podle BMI není stejné u mužů i žen.

Postup ověření první hypotézy:

Je zřejmé, že oba znaky (tj. **Kouření**, **Pohlaví**) jsou kvalitativní povahy. Vhodnou metodou pro ověření hypotézy je tedy chí-kvadrát test.

1. Vytvořte kontingenční tabulku. Umístěte ji na nový list. Do řádků tabulky vložte znak **Pohlaví**, do sloupců znak **Kouření**. Použijte postup, který jste se naučili na 1. cvičení ze statistiky.



	A	B	C	D	E	F	G
1							
2							
3			Kouření				
4	Pohlaví	Data	ne	ano	Celkový součet		
5	M	Počet	277	119	396		
6		%	69,95%	30,05%	100,00%		
7	Ž	Počet	122	66	188		
8		%	64,89%	35,11%	100,00%		
9	Celkem Počet		399	185	584		
10	Celkem %		68,32%	31,68%	100,00%		
11							

Z tabulky je možno vyčíst, že v souboru je 396 mužů, z toho 119 kuřáků, což je 30,1 %. Žen je v souboru pouze 188 a z nich je 66 kuřaček, což je 35,1 %.

Vidíme, že podíl kuřáků je o něco vyšší u žen. Zda je tento rozdíl statisticky významný je třeba ověřit chí-kvadrát testem.

Jinak řečeno, budeme zkoumat, zda tento rozdíl je pouze věcí náhody, či zda zde existuje skutečný rozdíl.

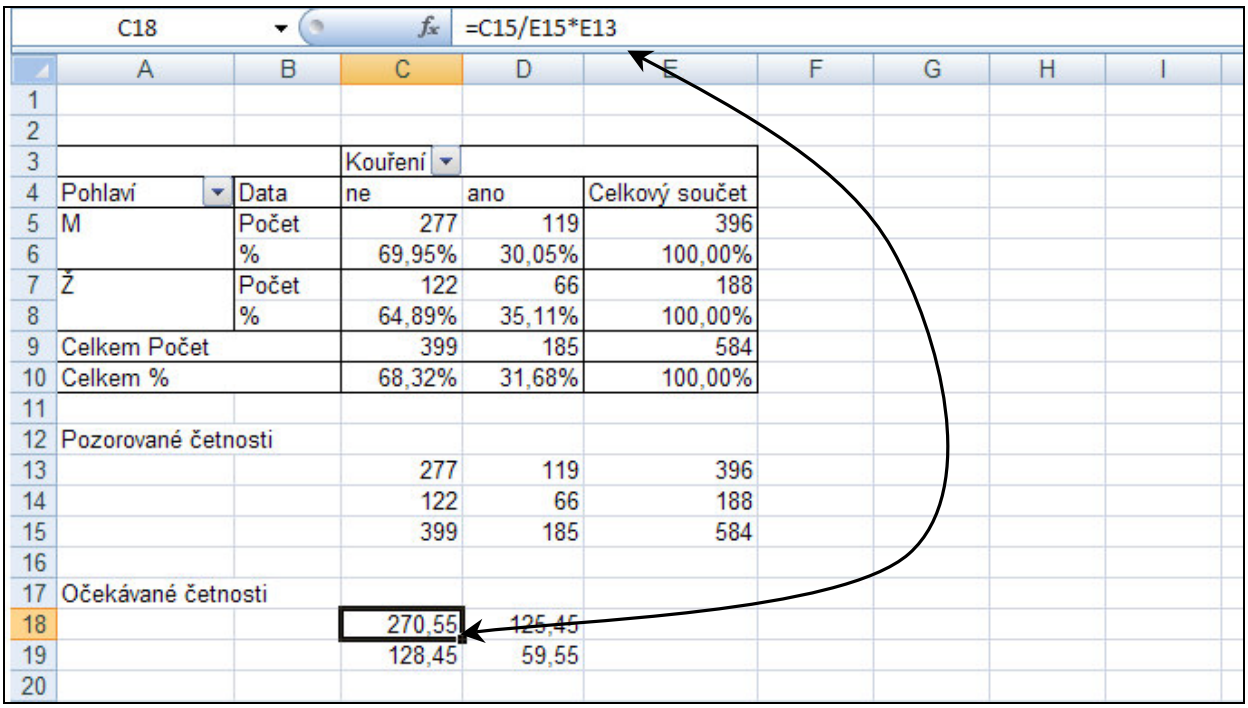
2. Pozorované absolutní četnosti opište pod kontingenční tabulku:

	A	B	C	D	E	F
1						
2						
3			Kouření			
4	Pohlaví	Data	ne	ano	Celkový součet	
5	M	Počet	277	119	396	
6		%	69,95%	30,05%	100,00%	
7	Ž	Počet	122	66	188	
8		%	64,89%	35,11%	100,00%	
9	Celkem Počet		399	185	584	
10	Celkem %		68,32%	31,68%	100,00%	
11						
12	Pozorované četnosti					
13			277	119	396	
14			122	66	188	
15			399	185	584	
16						

3. Vypočítejte očekávané četnosti. Pro výpočet použijte pravidlo:
očekávaná četnost = součet v sloupci / celkový počet * součet v řádku

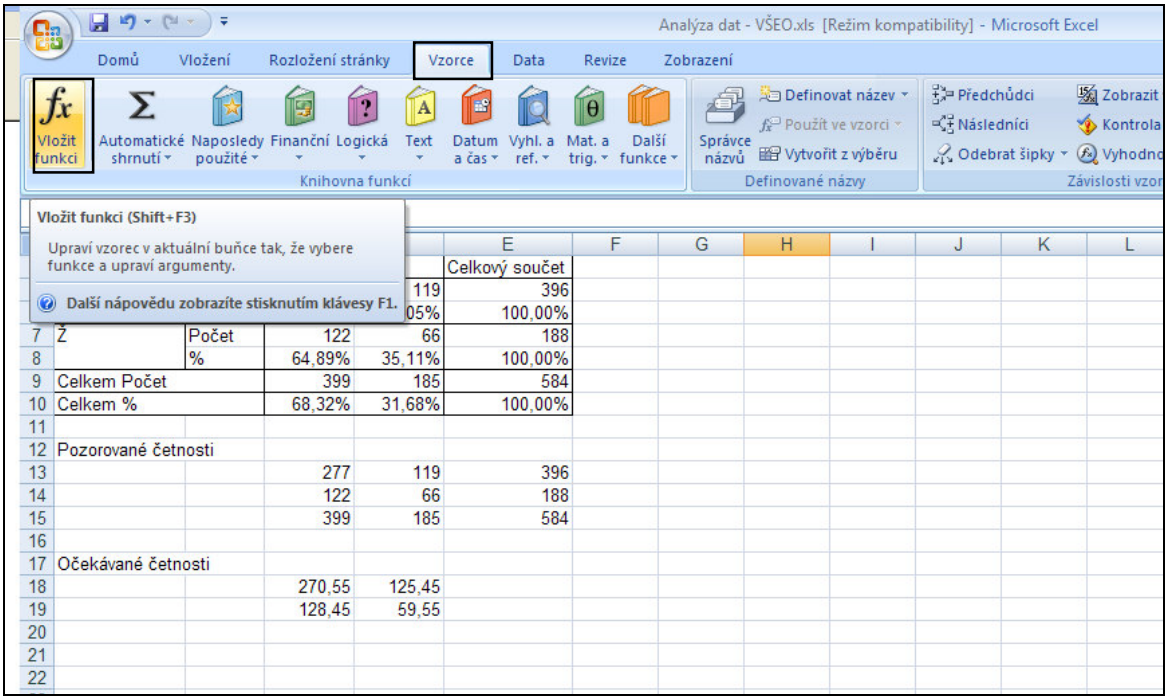
Tedy očekávané četnosti jsou:
= 399/584*396=270,55 = 185/584*396=125,45
= 399/584*188=128,45 = 185/584*188=59,55

Tyto výpočty proveďte pod tabulku Pozorované četností:

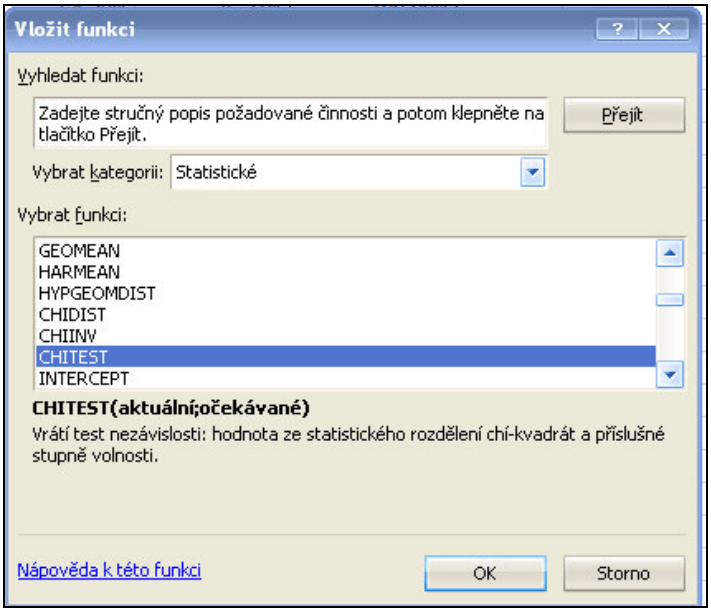


Šipka naznačuje, že do buněk můžete vkládat přímo výpočty. Buňky s příslušnými daty vyberte kliknutím myši.

4. K výpočtu dosažené hladiny statistické významnosti, neboli signifikance (tzv. *p*-hodnoty), použijeme funkci **CHITEST**.
   Klikněte do buňky, kam chcete umístit hodnotu signifikance (např. do buňky E21).
   Z řádkového menu zvolte **Vzorce** a klikněte na ikonu **Vložit funkci**.

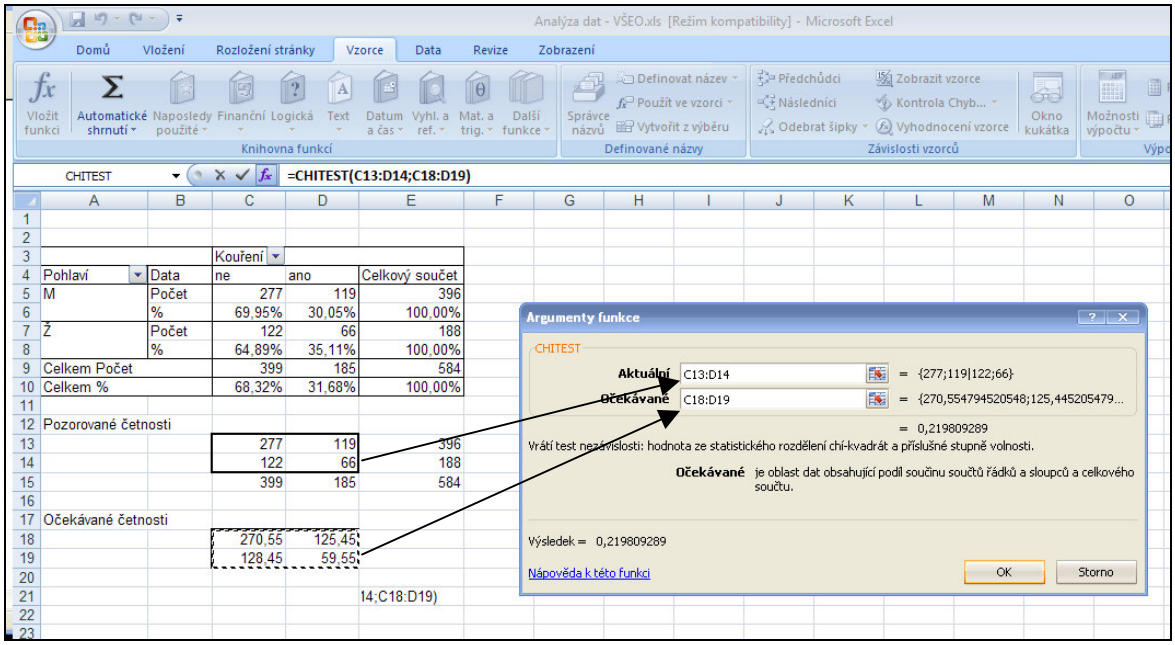


Otevřete dialogové okno **Vložit funkci**. V poli **Vybrat kategorii** vyberte **Statistické**, ze seznamu vyberte funkci **CHITEST**.



Otevřete dialogové okno **Argumenty funkce**. Do pole **Aktuální** zadejte adresu oblasti buněk s pozorovanými četnostmi C13:D14 (pouze čtyři hodnoty!).

Do pole **Očekávané** zadejte adresu oblasti buněk s vypočítanými očekávanými četnostmi C18:D19 (také čtyři hodnoty).



Klikněte na **OK.**

Tabulky s výslednou hodnotou signifikance:

	A	B	C	D	E
3			Kouření		
4	Pohlaví	Data	ne	ano	Celkový součet
5	M	Počet	277	119	396
6		%	69,95%	30,05%	100,00%
7	Ž	Počet	122	66	188
8		%	64,89%	35,11%	100,00%
9	Celkem Počet		399	185	584
10	Celkem %		68,32%	31,68%	100,00%
12	Pozorované četnosti				
13			277	119	396
14			122	66	188
15			399	185	584
17	Očekávané četnosti				
18			270,55	125,45	
19			128,45	59,55	
21	**Signifikance chí-kvadrát testu:**				**0,220**

Před vypočítanou hodnotu (např. do buňky A21) napište text „Signifikance chí-kvadrát testu:“ Hodnotu signifikance zaokrouhlete na 3 desetinná místa.

Funkce chí-kvadrát test v Excelu nezobrazuje hodnotu testového kritéria χ^2, zobrazí pouze *p*-hodnotu.

5. Výsledek, tedy dosaženou hladinu statistické významnosti, porovnáme s hodnotou 0,05. Je-li dosažená hladina statistické významnosti menší než 0,05, nulovou hypotézu zamítáme, v opačném případě nulovou hypotézu zamítnout nemůžeme. V tomto příkladu p = 0,220, nulovou hypotézu tedy zamítnout nemůžeme.
   Závěr testování zní: **Podíl kuřáků je stejný v populaci mužů i žen.**

Postup ověření druhé hypotézy:

Nulová hypotéza: Výskyt hypertenze nezávisí na pohlaví.
Alternativní hypotéza: Výskyt hypertenze závisí na pohlaví.

Postup bude obdobný jako v prvním příkladu:

1. Vytvořte kontingenční tabulku. Do řádků tabulky vložte znak **Pohlaví**, do sloupců znak **Hypertenze**. Tabulku umístěte na nový list.

   Kontingenční tabulka:

	A	B	C	D	E
1					
2					
3			Hypertenze		
4	Pohlaví	Data	ne	ano	Celkový součet
5	M	Počet	361	33	394
6		%	91,62%	8,38%	100,00%
7	Ž	Počet	175	13	188
8		%	93,09%	6,91%	100,00%
9	Celkem Počet		536	46	582
10	Celkem %		92,10%	7,90%	100,00%
11					

   Z tabulky je možno vyčíst, že v souboru je zahrnuto 394 mužů, z nichž 33 (t.j. 8,4 %) trpí hypertenzí, žen je v souboru 188, hypertenzí trpí 13 (t.j.6,9 %) žen. Vidíme, že rozdíl ve výskytu hypertenze u mužů a u žen je malý.

2. Pozorované absolutní četnosti opište pod kontingenční tabulku a spočítejte očekávané četnosti:

	A	B	C	D	E	F
1						
2						
3			Hypertenze			
4	Pohlaví	Data	ne	ano	Celkový součet	
5	M	Počet	361	33	394	
6		%	91,62%	8,38%	100,00%	
7	Ž	Počet	175	13	188	
8		%	93,09%	6,91%	100,00%	
9	Celkem Počet		536	46	582	
10	Celkem %		92,10%	7,90%	100,00%	
11						
12	Pozorované četnosti					
13			361	33	394	
14			175	13	188	
15			536	46	582	
16						
17	Očekávané četnosti					
18			362,86	31,14		
19			173,14	14,86		
20						

K výpočtu dosažené hladiny statistické významnosti opět použijte funkci **CHITEST** (Použijte příkaz **Vzorce** a zvolte **Vložit funkci**.)

3. Pokud jste postupovali správně, dostanete tento výsledek:

	A	B	C	D	E	F	G
1							
2							
3			Hypertenze				
4	Pohlaví	Data	ne	ano	Celkový součet		
5	M	Počet	361	33	394		
6		%	91,62%	8,38%	100,00%		
7	Ž	Počet	175	13	188		
8		%	93,09%	6,91%	100,00%		
9	Celkem Počet		536	46	582		
10	Celkem %		92,10%	7,90%	100,00%		
11							
12	Pozorované četnosti						
13			361	33	394		
14			175	13	188		
15			536	46	582		
16							
17	Očekávané četnosti						
18			362,86	31,14			
19			173,14	14,86			
20							
21	**Signifikance chí-kvadrát testu:**				**0,541**		
22							

4. Dosažená hladina signifikance p = 0,541, nulovou hypotézu tedy zamítnout nemůžeme. ***Závěr testování*** zní: **Výskyt hypertenze nezávisí na pohlaví.**

Postup ověření třetí hypotézy:

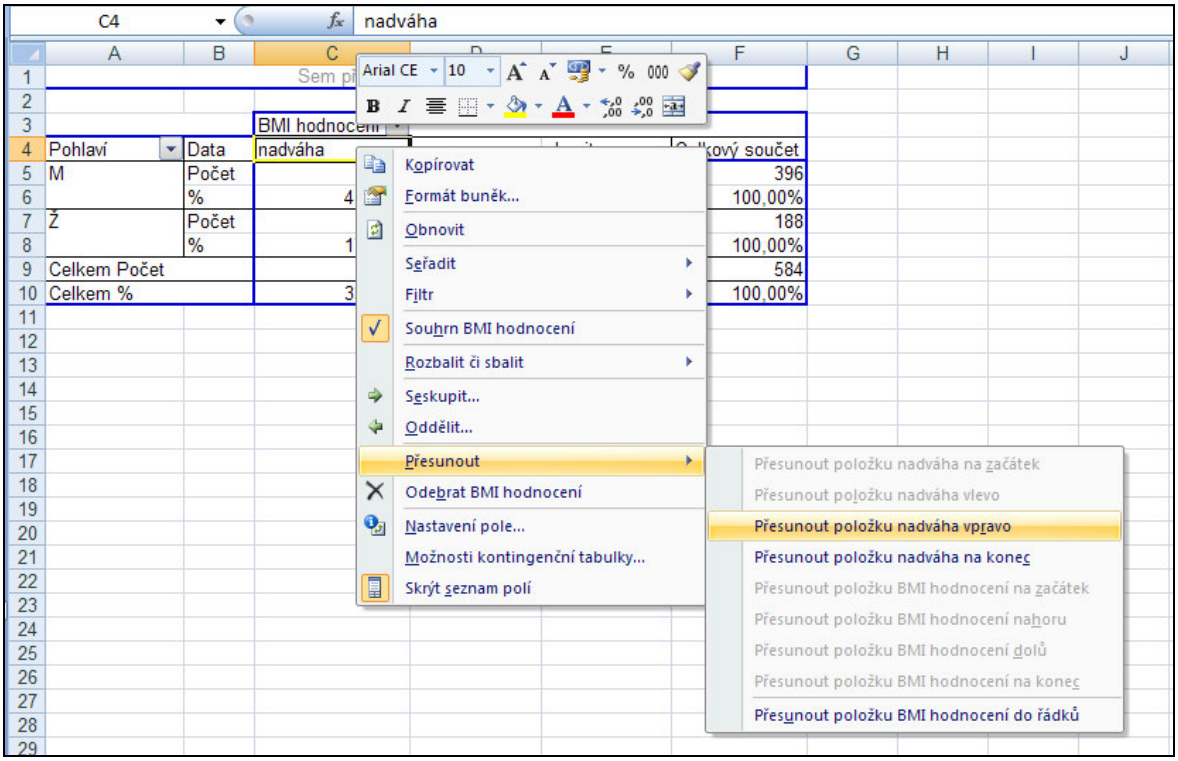
Nulová hypotéza: Rozdělení zaměstnanců podle BMI je stejné u mužů i žen.
Alternativní hypotéza: Rozdělení zaměstnanců podle BMI není stejné u mužů i žen.

Postup:

1. Vytvořte kontingenční tabulku. Do řádků tabulky vložte znak **Pohlaví**, do sloupců znak **BMI hodnocení**. Tabulku umístěte na nový list.
Kontingenční tabulka:

	A	B	C	D	E	F
1						
2						
3			BMI hodnocení			
4	Pohlaví	Data	nadváha	norma	obezita	Celkový součet
5	M	Počet	165	191	40	396
6		%	41,67%	48,23%	10,10%	100,00%
7	Ž	Počet	33	135	20	188
8		%	17,55%	71,81%	10,64%	100,00%
9	Celkem Počet		198	326	60	584
10	Celkem %		33,90%	55,82%	10,27%	100,00%
11						

Dostanete tabulku, která má 2 řádky a 3 sloupce. Kategorie uvedené ve sloupcích jsou uspořádány abecedně: nadváha, norma, obezita. Vzhledem k tomu, že **BMI hodnocení** je ordinální znak, měly by kategorie být logicky správně uspořádány: tedy norma, nadváha, obezita. Uspořádání můžete změnit, vyberte položku "nadváha" a klikněte pravým tlačítkem myši, v místní nabídce vyberte příkaz **Přesunout** a **Přesunout položku nadváha vpravo.**

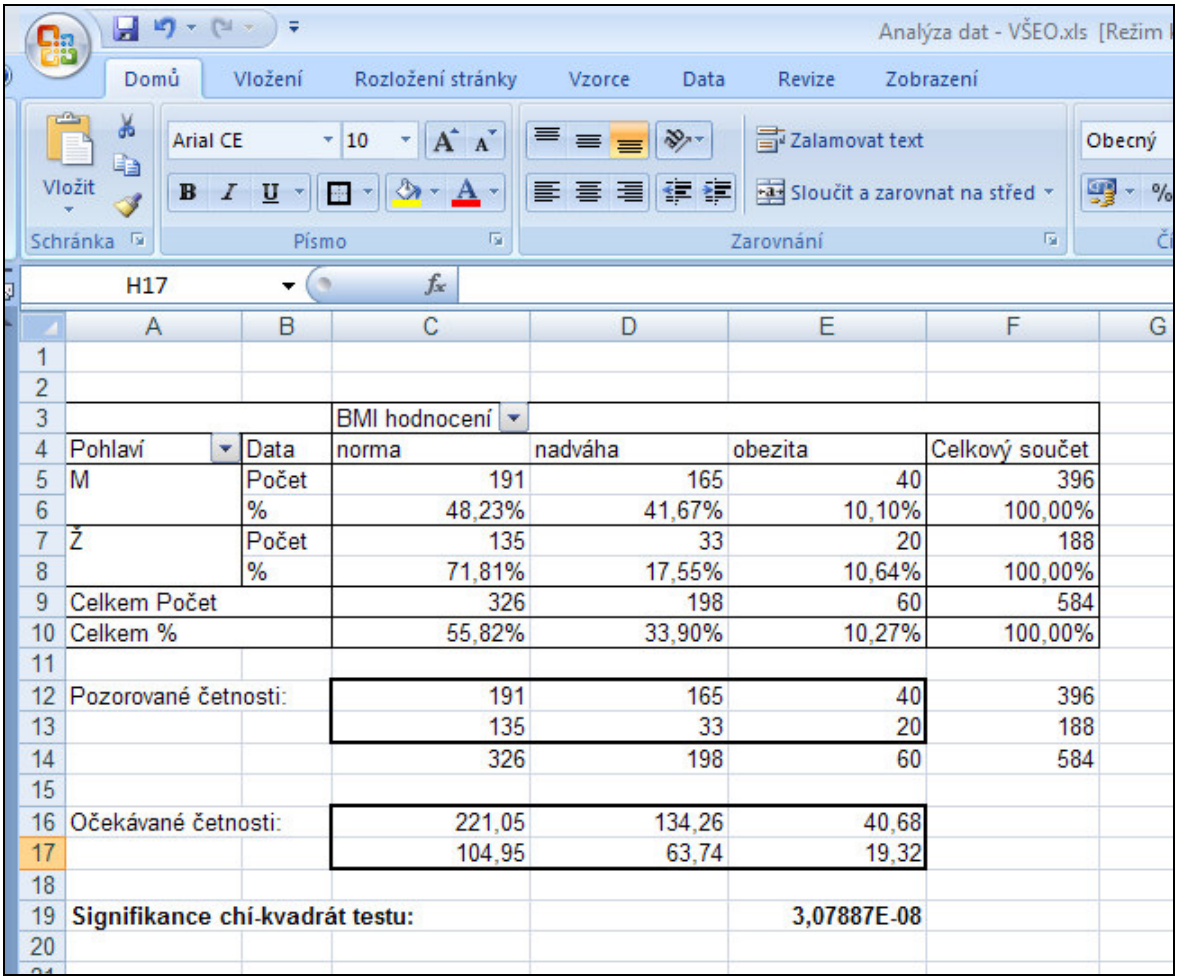


2. Pozorované absolutní četnosti opište pod kontingenční tabulku a spočítejte očekávané četnosti:

	A	B	C	D	E	F	G
1							
2							
3			BMI hodnocení				
4	Pohlaví	Data	norma	nadváha	obezita	Celkový součet	
5	M	Počet	191	165	40	396	
6		%	48,23%	41,67%	10,10%	100,00%	
7	Ž	Počet	135	33	20	188	
8		%	71,81%	17,55%	10,64%	100,00%	
9	Celkem Počet		326	198	60	584	
10	Celkem %		55,82%	33,90%	10,27%	100,00%	
11							
12	Pozorované četnosti:		191	165	40	396	
13			135	33	20	188	
14			326	198	60	584	
15							
16	Očekávané četnosti:		221,05	134,26	40,68		
17			104,95	63,74	19,32		
18							

3. K výpočtu dosažené hladiny statistické významnosti opět použijte funkci **CHITEST** (Použijte příkaz **Vzorce** a zvolte ikonu **Vložit funkci**.)

4. Pokud jste postupovali správně, dostanete tento výsledek:



	A	B	C	D	E	F	G
1							
2							
3			BMI hodnocení				
4	Pohlaví	Data	norma	nadváha	obezita	Celkový součet	
5	M	Počet	191	165	40	396	
6		%	48,23%	41,67%	10,10%	100,00%	
7	Ž	Počet	135	33	20	188	
8		%	71,81%	17,55%	10,64%	100,00%	
9	Celkem Počet		326	198	60	584	
10	Celkem %		55,82%	33,90%	10,27%	100,00%	
11							
12	Pozorované četnosti:		191	165	40	396	
13			135	33	20	188	
14			326	198	60	584	
15							
16	Očekávané četnosti:		221,05	134,26	40,68		
17			104,95	63,74	19,32		
18							
19	**Signifikance chí-kvadrát testu:**				**3,07887E-08**		
20							

Dosažená hladina signifikance $p = 3{,}1*10^{-8}$ je podstatně menší než 0,05, nulovou hypotézu můžeme zamítnout a přijmout její alternativu.

Závěr testování zní: **Rozdělení zaměstnanců podle BMI není stejné u mužů i žen.** 41,7 % muži trpí nadváhou, ženy trpí nadváhou méně často – pouze v 17,6 % případů. Obezitou trpí muži a ženy stejně.

Úkol k samostatnému řešení:

Otevřete list „onkologická léčba“. Zde jsou data pacientů, kteří podstoupili onkologickou léčbu. V sloupci B je uvedena diagnóza pacientů, rozlišujeme dvě diagnózy: rakovinu jazyka a rakovinu spodiny ústní. Ve sloupcích C a D jsou informace o tom, zda pacienti mají polykací potíže při pozření tuhé stravy či zda trpí pocitem pálení v dutině ústní.

1. Ověřte následující hypotézu:

Nulová hypotéza: Výskyt polykacích potíží nezávisí na sledovaných diagnózách.

Alternativní hypotéza: Výskyt polykacích potíží závisí na sledovaných diagnózách.

Návod:

Vytvořte kontingenční tabulku, do řádků vložte znak **Diagnóza**, do sloupců znak **Polykací potíže při pozření tuhé stravy**. Spočítejte očekávané četnosti a použijte funkci **CHITEST**.

2. Ověřte následující hypotézu:

Nulová hypotéza: Výskyt pálení v dutině ústní nezávisí na sledovaných diagnózách.

Alternativní hypotéza: Výskyt pálení v dutině ústní závisí na sledovaných diagnózách.

Návod:

Vytvořte kontingenční tabulku, do řádků vložte znak **Diagnóza**, do sloupců znak **Pocit pálení v dutině ústní při jídle**. Spočítejte očekávané četnosti a k výpočtu signifikance použijte funkci **CHITEST**.

3. Jak můžeme analyzovat závislost mezi kvantitativními znaky?

V kapitole 3 si ukážeme, jakým způsobem analyzujeme závislost mezi daty kvantitativní povahy. Krátce si připomeňme základní fakta ze statistické teorie.

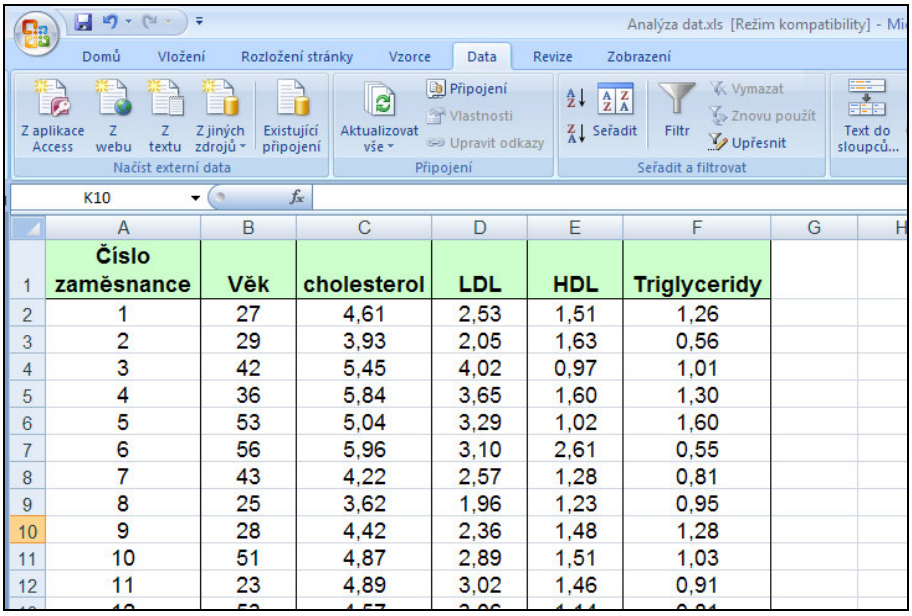
1. Korelační analýza

Posuzuje vzájemné vztahy pomocí různých měr závislosti, většinou pomocí různých korelačních koeficientů. Nejpoužívanější mírou těsnosti vztahu dvou spojitých znaků je Pearsonův korelační koeficient. Je mírou linearity vztahu (jak těsně se body přimykají k přímce). **Pearsonův korelační koeficient** se značí *r* a vzorec pro přesný výpočet najdete ve výukových textech. Pro hodnoty *r* platí: $-1 \leq r \leq 1$. Hodnoty ± 1 nabývá tehdy, když veličiny jsou absolutně závislé, tzn. pokud sestrojíme bodový graf dvojice zkoumaných veličin, všechny body leží na přímce. Pokud r = 0 (nebo nabývá hodnoty blízké nule), veličiny jsou nezávislé. Kladné hodnoty korelačního koeficientu znamenají pozitivní závislost, obě veličiny zároveň rostou nebo klesají. Záporné hodnoty korelačního koeficientu znamenají negativní závislost, jedna veličina roste, zatímco druhá klesá. Míru závislosti podle absolutní hodnoty Pearsonova korelačního koeficientu obvykle interpretujeme:

0,1 – 0,3	korelace slabá
0,4 – 0,6	korelace střední
0,7 – 0,8	korelace silná
nad 0,9	korelace velmi silná.

Data, se kterými budete pracovat, naleznete opět v souboru F://SOFTWARE/biostatistika/data /analýza dat.xls.

Otevřete list „*Korelace*“.



	A	B	C	D	E	F	G	H
1	**Číslo zaměsnance**	**Věk**	**cholesterol**	**LDL**	**HDL**	**Triglyceridy**		
2	1	27	4,61	2,53	1,51	1,26		
3	2	29	3,93	2,05	1,63	0,56		
4	3	42	5,45	4,02	0,97	1,01		
5	4	36	5,84	3,65	1,60	1,30		
6	5	53	5,04	3,29	1,02	1,60		
7	6	56	5,96	3,10	2,61	0,55		
8	7	43	4,22	2,57	1,28	0,81		
9	8	25	3,62	1,96	1,23	0,95		
10	9	28	4,42	2,36	1,48	1,28		
11	10	51	4,87	2,89	1,51	1,03		
12	11	23	4,89	3,02	1,46	0,91		

Na listu „*Korelace*“ jsou data 600 zaměstnanců nemocnice. Ve sloupci A **Číslo zaměstnance** je uvedena identifikace. Druhý sloupec (B) **Věk** poskytuje informaci o věku zaměstnance v letech, sloupce C až F obsahují výsledky testů lipidového profilu v mmol/l (**celkový cholesterol**, **LDL**, **HDL**, **Triglyceridy**).

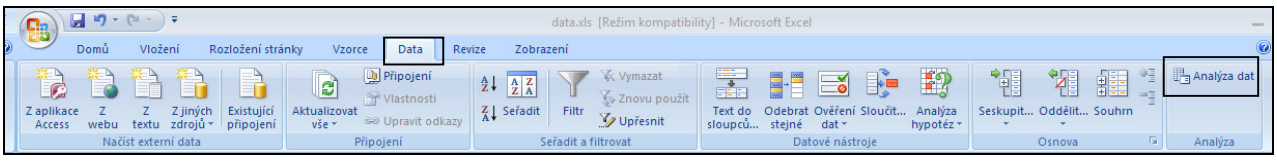
Úkol:
U každého sledovaného znaku určete jeho typ.
Návod: Rozlišujte znaky kvalitativní a kvantitativní.

Zadání úkolu

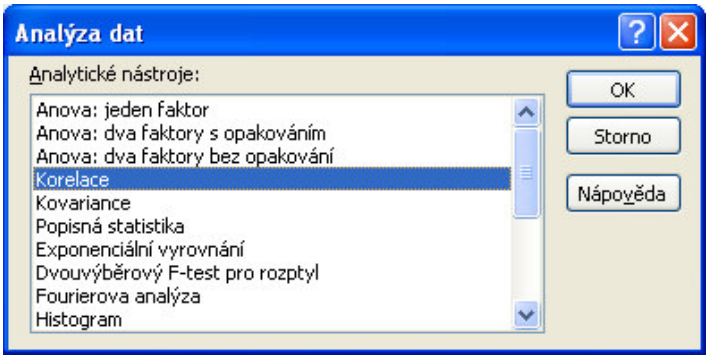
Vaším úkolem bude analyzovat míru závislosti naměřených parametrů.

Postup

K výpočtu Pearsonova korelačního koeficientu použijeme analytický nástroj **Korelace**. Tento nástroj je obsažen v položce **Analýza dat**. (Analýzu dat nastavte stejným způsobem jako při použití nástroje Popisná statistika – klikněte na ikonu  , otevřete **Možnosti aplikace Excel**, vyberte položku **Doplňky**, nastavte **Analytické nástroje** jako **Aktivní doplněk k dispozici** a klikněte na tlačítko **Přejít**. Zaškrtněte **Analytické nástroje** a potvrďte OK. Vyberte položku **Data** a v hlavním menu se Vám objeví nová položka **Analýza dat**:

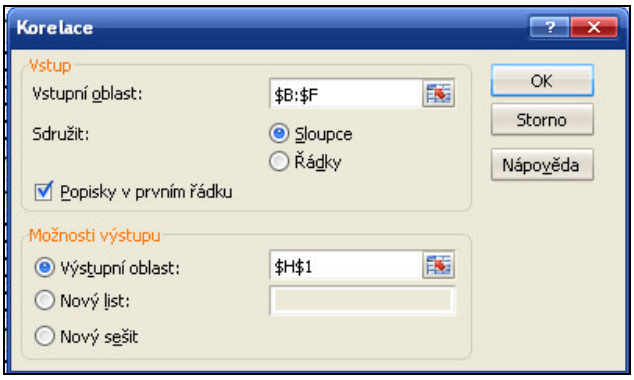


1. Klikněte na **Analýza dat** ze seznamu analytických nástrojů vyberte položku **Korelace**.

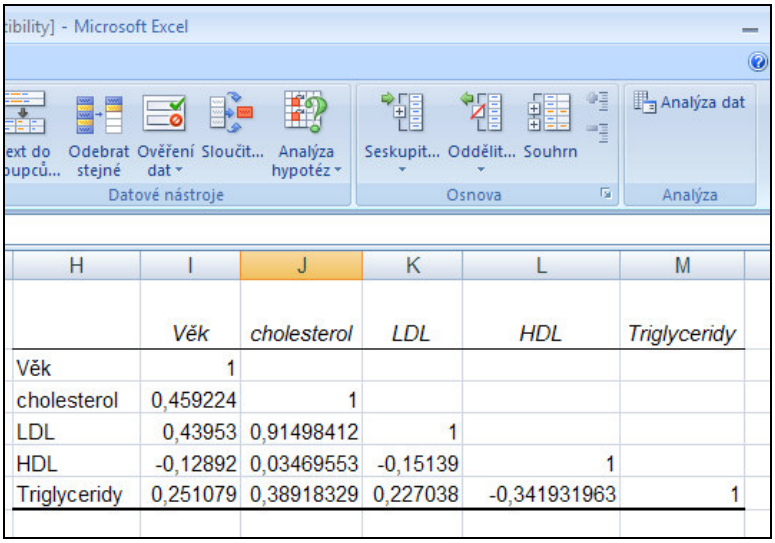


Vyplňte dialogové okno **Korelace**.

2. Do pole **Vstupní oblast** zadejte adresu celých sloupců B až F, které obsahují data týkající se lipidového profilu a věku zaměstnanců. Data jsou sdružena ve sloupcích, zatrhněte položku **Popisky v prvním řádku**.
3. Do pole **Výstupní oblast** zadejte adresu buňky H1. Potvrďte tlačítkem **OK**.



Dostanete korelační matici:



	H	I	J	K	L	M
		Věk	*cholesterol*	*LDL*	*HDL*	*Triglyceridy*
Věk		1				
cholesterol		0,459224	1			
LDL		0,43953	0,91498412	1		
HDL		-0,12892	0,03469553	-0,15139	1	
Triglyceridy		0,251079	0,38918329	0,227038	-0,341931963	1

V řádcích i ve sloupcích jsou uvedeny všechny zkoumané znaky, čísla uvnitř matice jsou hodnoty Pearsonova korelačního koeficientu pro danou dvojici znaků.

Je zřejmé, že nejsilnější pozitivní závislost je mezi celkovým cholesterolem a LDL cholesterolem r = 0,915, naopak téměř nulová korelace, tedy nezávislost byla zjištěna mezi celkovým cholesterolem a HDL cholesterolem r = 0,035. Slabá negativní korelace byla zjištěna mezi triglyceridy a HDL, r = -0,342.

2. Regresní analýza

Metoda regresní analýzy hledá matematické vyjádření vztahu mezi znaky (lineární, kvadratický, exponenciální …) a dává odpověď na otázku, zda lze znak *Y* odhadnout na základě jiného nebo jiných znaků a s jakou chybou.

Postup regresní analýzy lze shrnout do těchto bodů:

1. Sestrojení bodového grafu a jeho posouzení.
2. Volba typu regresní křivky a výpočet jejich koeficientů.
3. Hodnocení kvality nalezeného řešení.

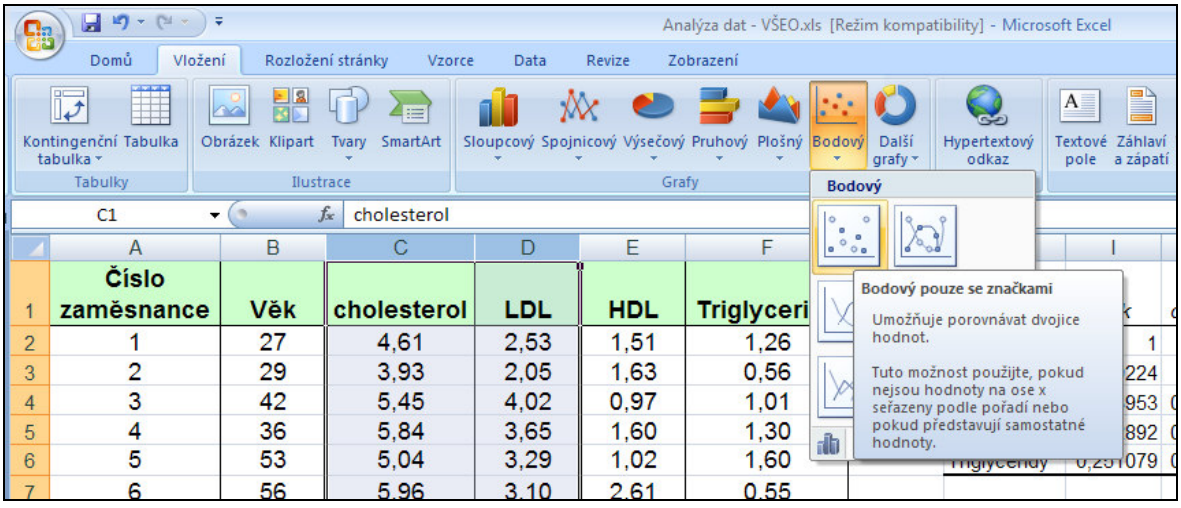
Poznámka: V řadě případů lze vztah popsat přímkou. Nalezením koeficientů této přímky se zabývá tzv. lineární regresní analýza.

Zadání úkolu

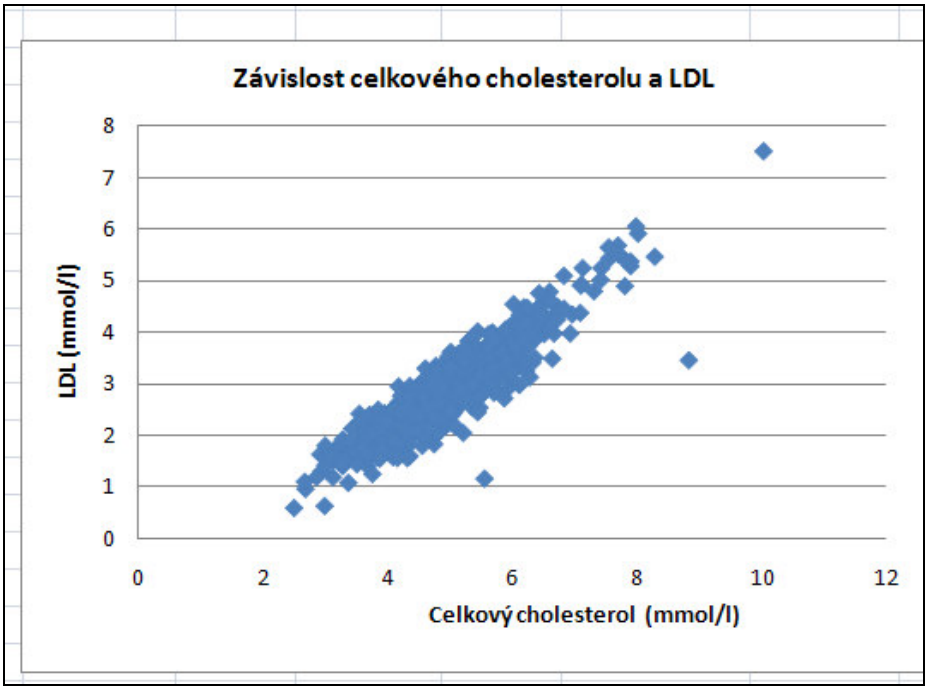
Korelační analýzou bylo zjištěno, že nejsilnější závislost mezi veličinami zkoumanými na listu „*Korelace*“ je mezi **celkovým cholesterolem** a **LDL**. Proveďte regresní analýzu těchto veličin.

Postup

1. Sestrojte bodový graf zkoumaných veličin.
   Pomocí myši vyberte všechny hodnoty sloupců C (**cholesterol**) a D (**LDL**).
2. Klikněte na příkaz **Vložení** a vyberte položku **Bodový** ze skupiny **Grafy**, vyberte první typ z nabízených typů bodových grafů.

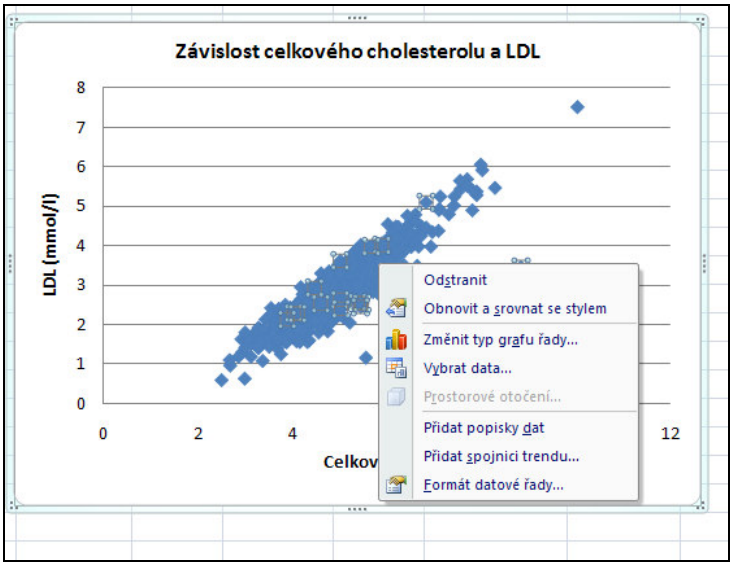


Graf upravte do následující podoby:

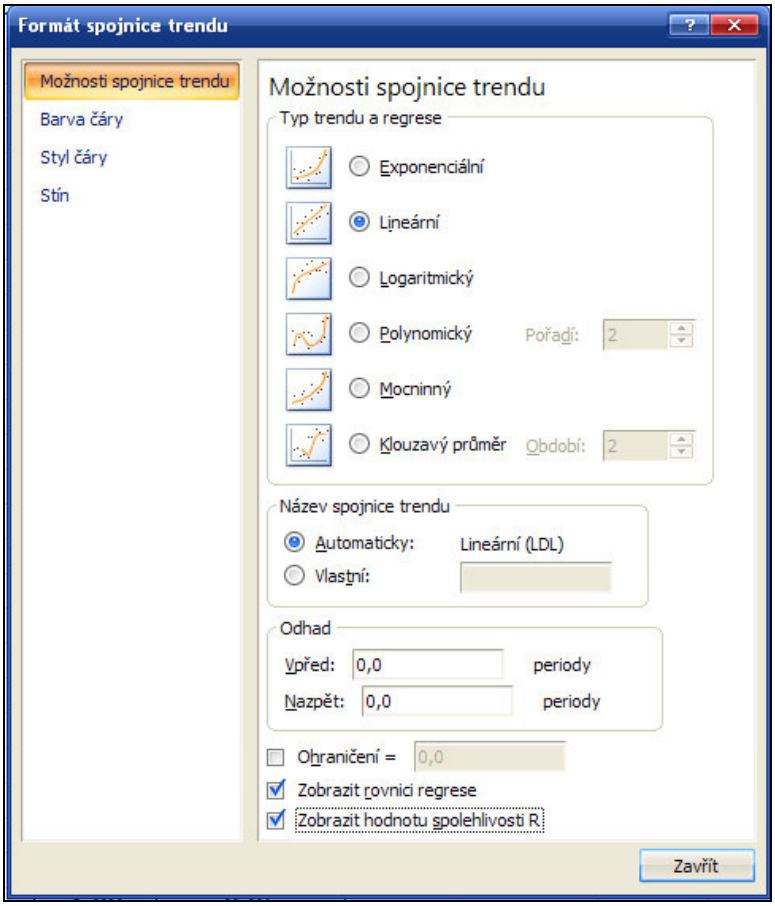


Volba typu závislosti a výpočet koeficientů regresní křivky

3. Klikněte pravým tlačítkem myši na graf mezi modré značky a vyvolejte místní nabídku:



4. Klikněte na položku **Přidat spojnici trendu**, otevře se Vám dialogové okno **Formát spojnice trendu**.
5. Vyberte **Lineární** trend a zatrhněte možnosti **Zobrazit rovnici regrese** a **Zobrazit hodnotu spolehlivosti**.

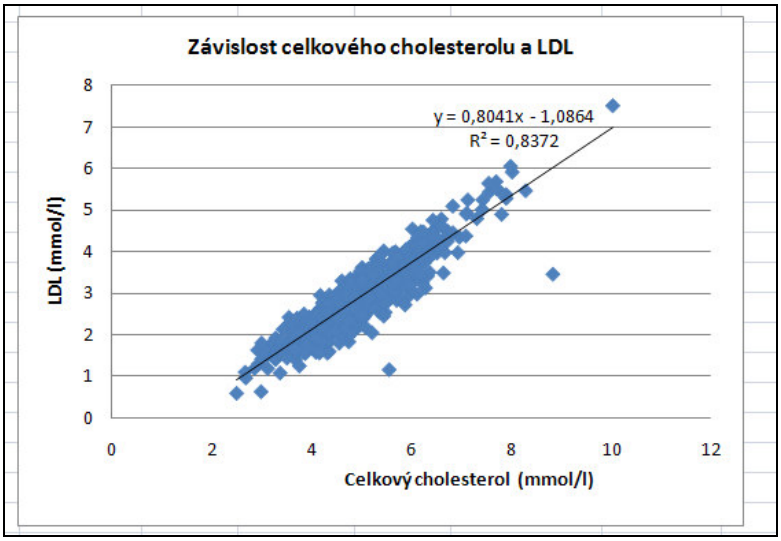


Pokud máte správně vyplněno, zavřete dialogové okno.

Do grafu se vloží regresní rovnice – v našem případě se jedná o rovnici přímky:

LDL = 0,8*Celkový cholesterol – 1,1

Zobrazí se také hodnota spolehlivosti R^2=0, 837.



6. Hodnocení kvality nalezeného řešení.

   Hodnota spolehlivosti, tj. koeficient determinace R^2, udává procento, jakým je rozptyl hodnot závisle proměnné veličiny Y (**LDL**) vysvětlen změnami hodnot nezávisle proměnné veličiny X (**Celkový cholesterol**). Koeficient nabývá hodnot od 0 do 1. Čím je vyšší, tím je nalezený model kvalitnější. V případě lineární regrese je koeficient determinace roven druhé mocnině Pearsonova korelačního koeficientu.

   (Ověřte: $0{,}915^2 = 0{,}837225$)

   V našem případě je hodnota $R^2 = 0{,}837$ poměrně vysoká, lineární model byl vhodně zvolen.

Zadání úkolu k samostatnému řešení

Na listu „*Regresní analýza*“ naleznete data týkající se teploty a dynamické viskozity vody. Metodou regresní analýzy analyzujte závislost viskozity vody na teplotě.

Návod:

1. Sestrojte bodový graf, osa X představuje teplotu, osa Y dynamickou viskozitu.
2. Zvolte nejvhodnější typ regresní křivky a najděte její rovnici.
3. Pomocí koeficientu determinace zhodnoťte kvalitu nalezeného řešení.