ZÁKLADY BIOSTATISTIKY PRO STUDENTY VŠEOBECNÉHO LÉKAŘSTVÍ

URČENO PRO VZDĚLÁVÁNÍ V AKREDITOVANÝCH STUDIJNÍCH PROGRAMECH

HANA TOMÁŠKOVÁ

ČÍSLO OPERAČNÍHO PROGRAMU: CZ.1.07
NÁZEV OPERAČNÍHO PROGRAMU:
VZDĚLÁVÁNÍ PRO KONKURENCESCHOPNOST
OPATŘENÍ: 7.2
ČÍSLO OBLASTI PODPORY: 7.2.2

INOVACE VÝUKY INFORMATICKÝCH PŘEDMĚTŮ VE STUDIJNÍCH PROGRAMECH OSTRAVSKÉ UNIVERZITY

REGISTRAČNÍ ČÍSLO PROJEKTU: CZ.1.07/2.2.00/28.0245

OSTRAVA 2012

Tento projekt je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky

Recenzent: Mgr. Petr Bujok

Název: Základy biostatistiky pro studenty všeobecného lékařství
Autor: Ing. Hana Tomášková, Ph.D.
Vydání: první, 2012
Počet stran: 70

Jazyková korektura nebyla provedena, za jazykovou stránku odpovídá autor.

OBSAH

Úvod

Výuková opora *Základy biostatistiky pro studenty všeobecného lékařství* je určena pro studenty všeobecného lékařství a pro studenty ostatních oborů na lékařské fakultě, kteří začínají využívat statistické funkce MS Excelu pro zpracování dat.
V rámci předmětu *Lékařská biofyzika, výpočetní technika I* jsou studenti seznámeni se základy biostatistiky v šesti přednáškách a navazujících cvičeních. Předmětem této opory není podrobně vysvětlit teorii k probírané problematice, ale na praktických příkladech předvést a vysvětlit použití vybraných statistických funkcí MS Excelu verze 2007 a programu Open Epi (2). Teorii k probíraným tématům je obsahem řady učebnic (1, 3-10). Program MS Excel umožňuje provést dané výpočty různými způsoby, v rámci této opory není možné podat vyčerpávající výčet všech postupů. Budou zde předvedeny jen vybrané funkce, ale pokud je student zvyklý používat jiné postupy, které povedou ke stejnému výsledku, může je využívat. V rámci praktických cvičení je problém v různé úrovni základních znalostí práce s programem MS Excel u studentů prvních ročníků. Proto jsou v úvodních kapitolách vysvětleny i některé základní funkce, které již studenti mohou znát ze střední školy.
Uváděné příklady a úkoly budou řešeny na cvičeních Lékařská biofyzika, výpočetní technika I. Znalosti získané v jednotlivých kapitolách na sebe navazují.

Literatura

1. Anděl J. *Statistické metody*. 4. vydání. Praha: Matfyzpress, 2007.
2. Dean AG, Sullivan KM, Soe MM. OpenEpi: *Open Source Epidemiologic Statistics for Public Health*, Version 2.3.1. www.OpenEpi.com, updated 2011/23/06.
3. Hendl J. *Přehled statistických metod*. Portál, 2012.
4. HRACH, Karel. *Sbírka úloh ze statistiky*. 1. vydání. Ústí nad Labem: FSE UJEP, 2006.
5. Kasal P., Svačina Š. a kol. *Lékařská informatika*. Praha: Karolinum,1998.
6. Procházka B. *Biostatistika pro lékaře. Principy základních metod a jejich interpelace*. Praha: Karolinum, 1999.
7. Rusnák M., Rusnáková V., Majdan M. *Bioštatistika pre študentov verejného zdravotníctva*, Trnava, 2010.
8. Tomášková H. *Základy biostatistiky*. 2. vydání. Ostrava: Ostravská univerzita, 2010.
9. Tvrdík J. *Základy matematické statistiky*, učební text pro kombinované a distanční studium. Ostrava: Ostravská universita, 2008.
10. Zvárová J. *Biomedicínská statistika I.* - Základy statistiky pro biomedicínské obory. 1.vydání. Praha: Karolinum., 2001.

1 Biostatistika a pravděpodobnost

V této kapitole se dozvíte:

- Co je předmětem biostatistiky.
- Co je to pravděpodobnost.
- Jaká jsou základní pravidla pravděpodobnostního počtu.
- Co je to úplná pravděpodobnost.
- Co je senzitivita, specificita, prediktivní hodnota.

Po jejím prostudování byste měli být schopni:

- Vysvětlit pojem biostatistika a pravděpodobnost.
- Zvládnout výpočty pomocí základních pravidel počítání s pravděpodobností.
- Použít v praxi výpočty založené na úplné pravděpodobnosti.
- Vypočítat senzitivitu, specificitu a prediktivní hodnotu diagnostických testů.

Klíčová slova této kapitoly:

Biostatistika, pravděpodobnost, podmíněná pravděpodobnost, věta o úplné pravděpodobnosti, diagnostický test, senzitivita, specificita, prediktivní hodnota.

Doba potřebná ke studiu a zpracování úkolů: **5 hodin**

Průvodce studiem

V této úvodní kapitole jsou studenti seznámeni se základními funkcemi a výpočty v MS Excelu. Vzhledem k tomu, že se studenti s pojmem pravděpodobnost setkali na střední škole, patří tato kapitola jen k méně obtížným.
Teorie k uvedeným tématům je probírána ve druhé a třetí přednášce Lékařská biofyzika, výpočetní technika I – Biostatistika.
Pro pochopení použití biostatistiky v medicíně mají studenti za úkol stručně interpretovat výsledky odborné publikace, která se zabývá hodnocením zdravotnických dat (korespondenční úkol 1).

1.1 Biostatistika

Biostatistika neboli biometrie představuje soubor statistických metod a postupů, které se používají v biomedicínckých oborech při studiu a pochopení biologických jevů.
Každý výzkum vychází z výzkumné hypotézy. Hypotéza obsahuje předpoklad o vlastnostech nebo chování daného jevu nebo předmětu. Předpoklad vzniká většinou na základě nashromáždění množství faktů o daném problému.
Hypotéza je tedy tvrzení, jehož objektivní platnost se předpokládá, ale je ji nutné empiricky prověřit, verifikovat.

K ověření hypotézy je možné použít ***indukci.*** Je to proces, který se snaží dospět od dílčích pozorování k zákonitosti, která by platila pro všechny objekty daného druhu.
Induktivní úsudek není vlastně nikdy zcela pravdivý, dokud nebyla prověřena jeho platnost ve všech případech, což je většinou nemožné. Závěry induktivních myšlenkových pochodů jsou ovlivněny subjektivními postoji a mají pouze omezenou platnost. Existuje však ***induktivní statistika***, která zahrnuje metody, jak poznatky získané v procesu indukce využít. Umožňuje z pozorovaných dat vytvářet obecné závěry s udáním ***stupně jejich spolehlivosti***. Výpočet stupně spolehlivosti je objektivní, neboť je založen na poznatcích ***teorie pravděpodobnosti*** a nezávisí na subjektivním názoru hodnotitele.

1.2 Pravděpodobnost

Veličiny sledované u objektů souboru mají povahu ***náhodných jevů***. Výskyt nebo hodnoty těchto veličin jsou ovlivněny řadou nekontrolovatelných faktorů. Nelze je předvídat s jistotou, ale pouze s určitou spolehlivostí, která je číselně vyjádřena ***pravděpodobností*** výskytu jevu.
Náhodný jev může být např. hodnota krevního tlaku, tělesná výška, barva očí, výsledek léčby apod.
Pro jevy se používá značení velkými písmeny *A, B,….* a pro jejich pravděpodobnosti *P(A), P(B)*, atd.

Četnostní definice pravděpodobnosti vychází z takzvaného zákona velkých čísel, podle kterého při velkém počtu nezávislých pozorování (*n*) kolísá relativní četnost výskytu jevu *A* kolem čísla označeného jako pravděpodobnost:
$P(A) \approx m/n$ pro *n* velké,
n – počet navzájem nezávislých pozorování
m – počet pozorování, ve kterých nastal jev *A*

1.2.1 Základní pravidla počítání s pravděpodobností

1. Platí: $0 \leq P(A) \leq 1$

2. Pro libovolné dva jevy A, B platí:
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

3. Pro ***jev doplňkový*** (komplementární = *nonA*) jevu *A* platí:
$$P(nonA) = 1 - P(A)$$

4. Jevy ***A, B jsou nezávislé***, jestliže výskyt jednoho jevu není ovlivněn výskytem druhého jevu. Jevy jsou nezávislé tehdy a jen tehdy, jestliže platí:
$$P(A \cap B) = P(A) \,.\, P(B)$$

5. Nejsou-li jevy ***A, B nezávislé***, pak platí:
$$P(A \cap B) = P(A|B).\, P(B),$$
kde $P(A|B)$ označuje ***podmíněnou pravděpodobnost*** tzn. pravděpodobnost výskytu jevu *A* je podmíněná výskytem jevu *B*.

Příklad 1.1

Při odběru krve za určité období bylo zjištěno následující zastoupení krevních skupin:

Krevní skupina	A	B	AB	0
Počet osob	348	153	76	272

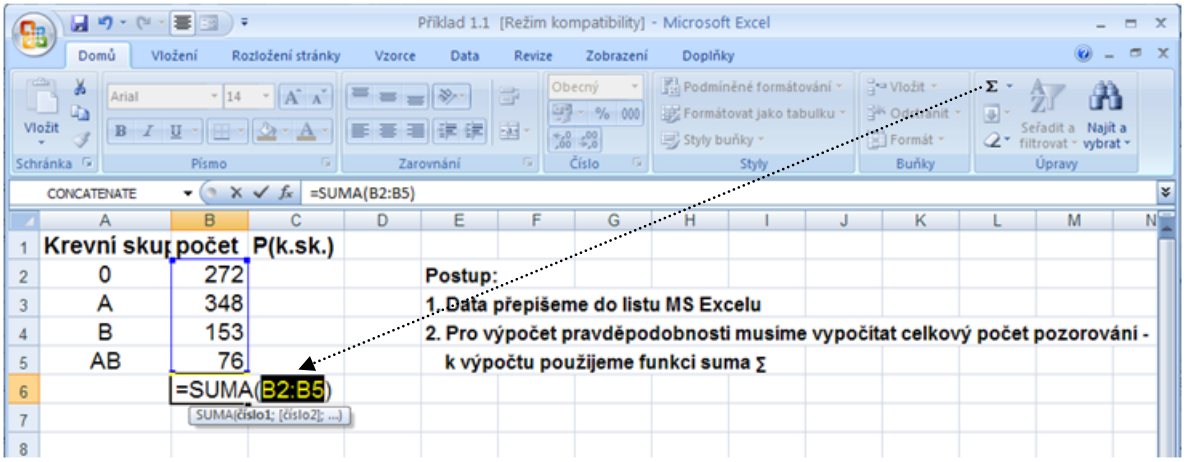
Jestliže bude proveden náhodný výběr z této skupiny, jaká je pravděpodobnost, že osoba bude mít krevní skupinu:

a) A b) B c) AB d) 0

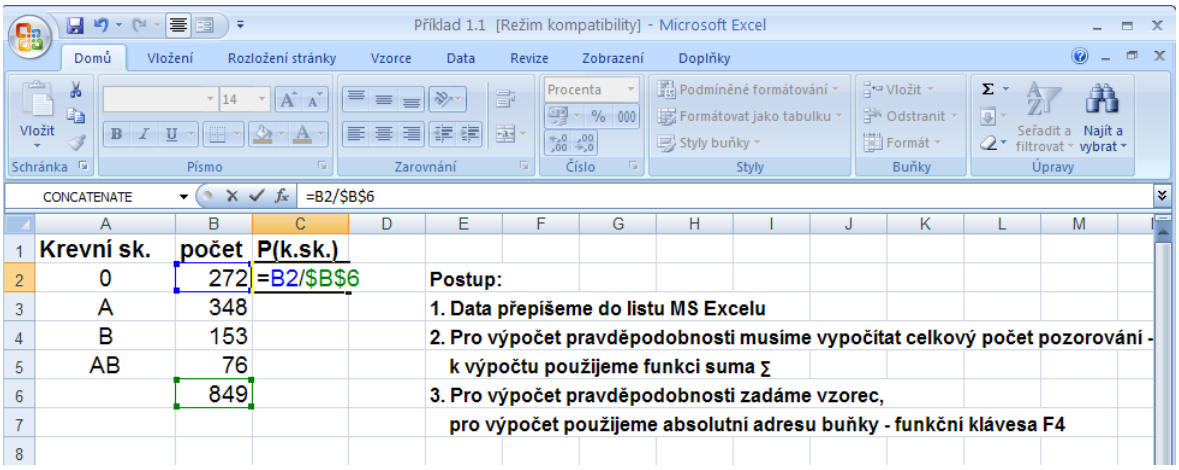
e) Jaká je pravděpodobnost, že náhodně vybraná osoba nebude mít krevní skupinu 0.

f) Jaká je pravděpodobnost, že krevní skupina náhodně vybrané osoby bude A nebo B?

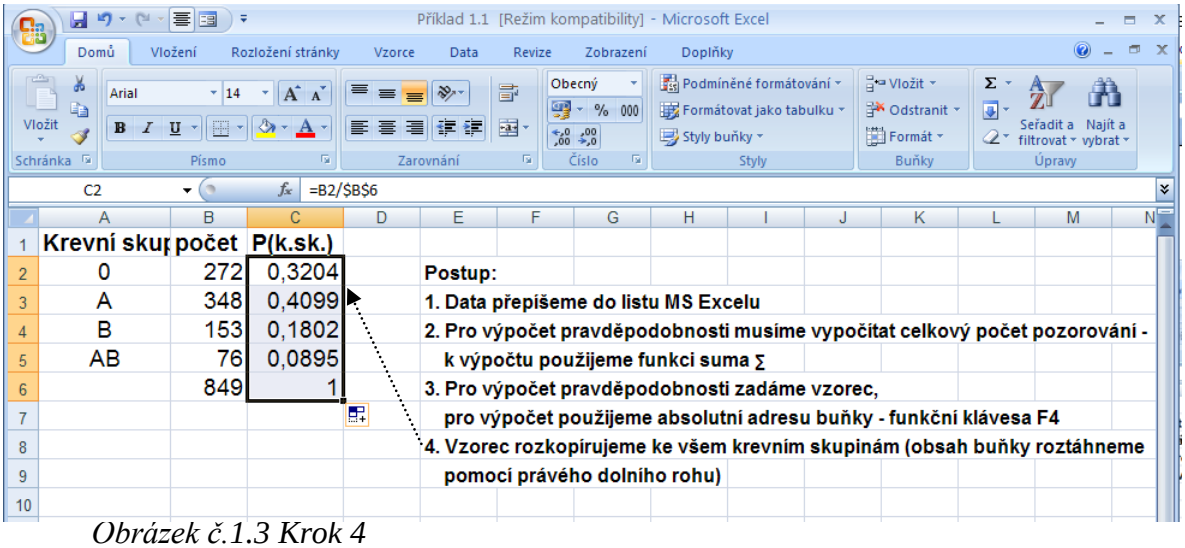
Řešení (postup je uveden přímo v obrázku):



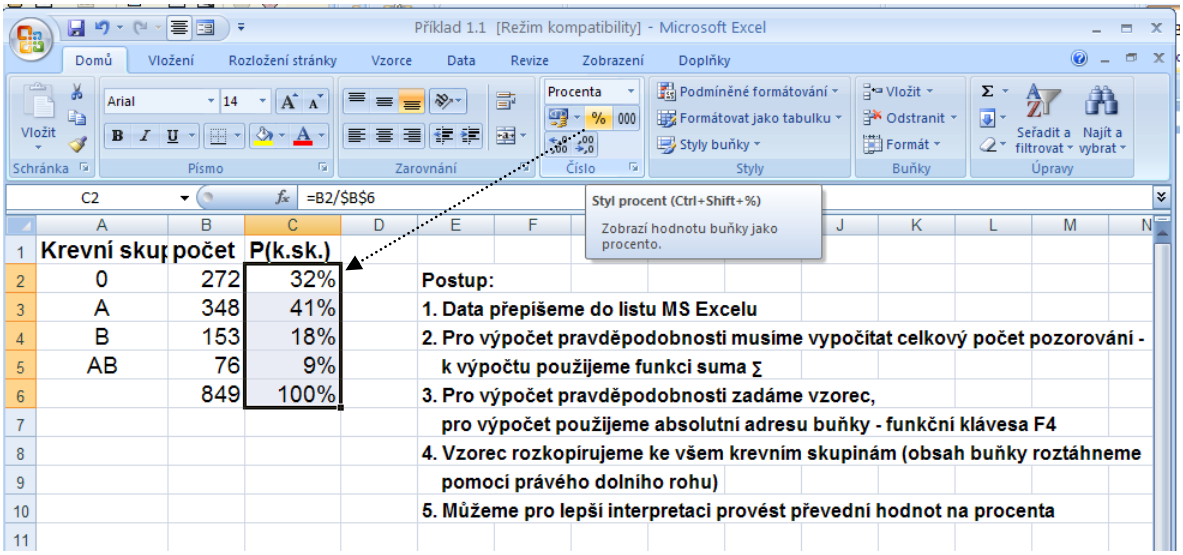
Obrázek č.1.1 Krok 1 – 2



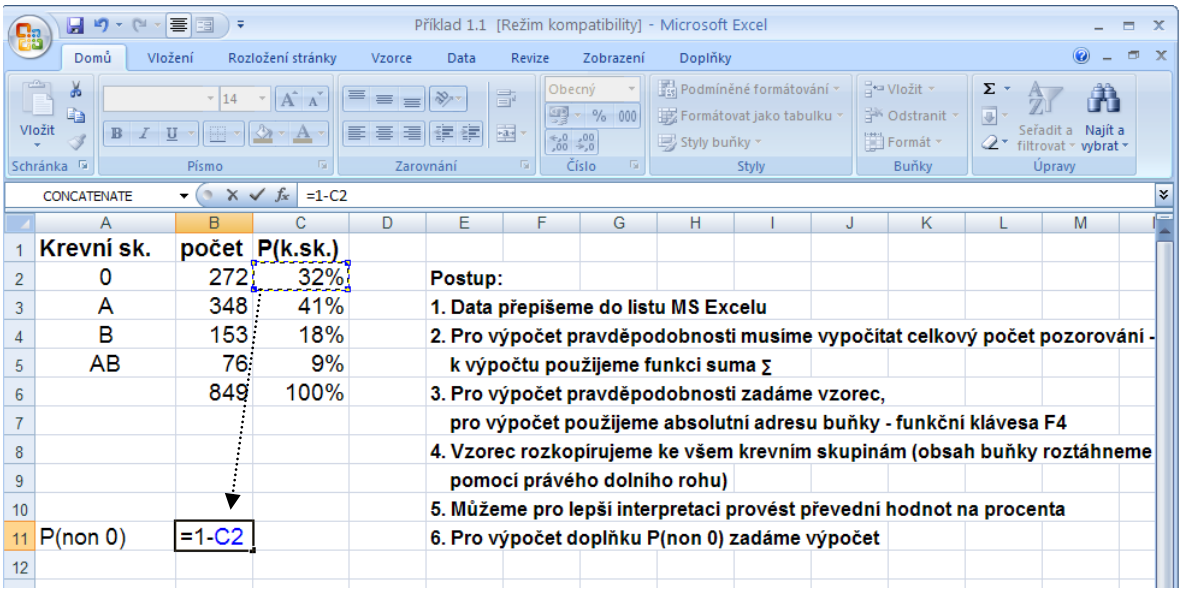
Obrázek č.1.3 Krok 3



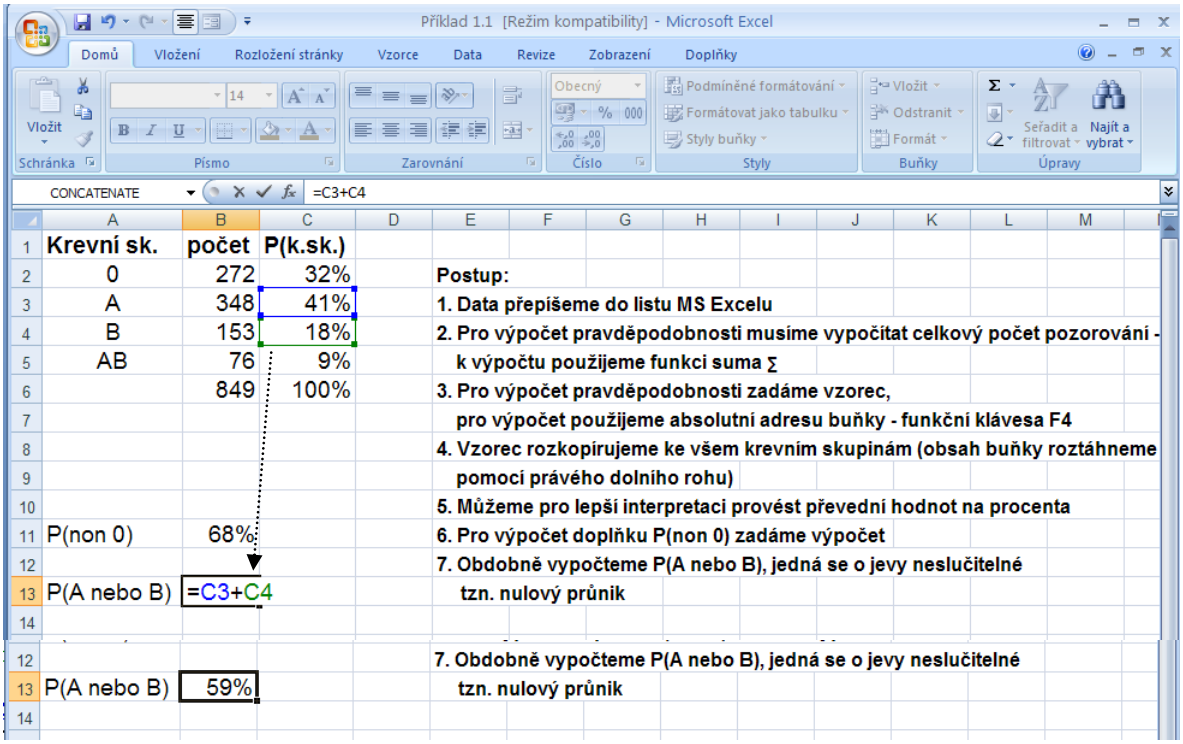
Obrázek č.1.3 Krok 4



Obrázek č.1.4 Krok 5(výsledek a) – d))



Obrázek č.1.5 Krok 6



Obrázek č. 1.6 Krok 7(výsledek e), f))

1.2.2 Úplná pravděpodobnost

Podle tohoto pravidla lze vypočítat pravděpodobnost jevu ze známých podmíněných pravděpodobností.
Jestliže jevy (B_1, B_2, ..., B_k) tvoří úplnou množinu jevů vzájemně neslučitelných a A je libovolný jev, pak platí:

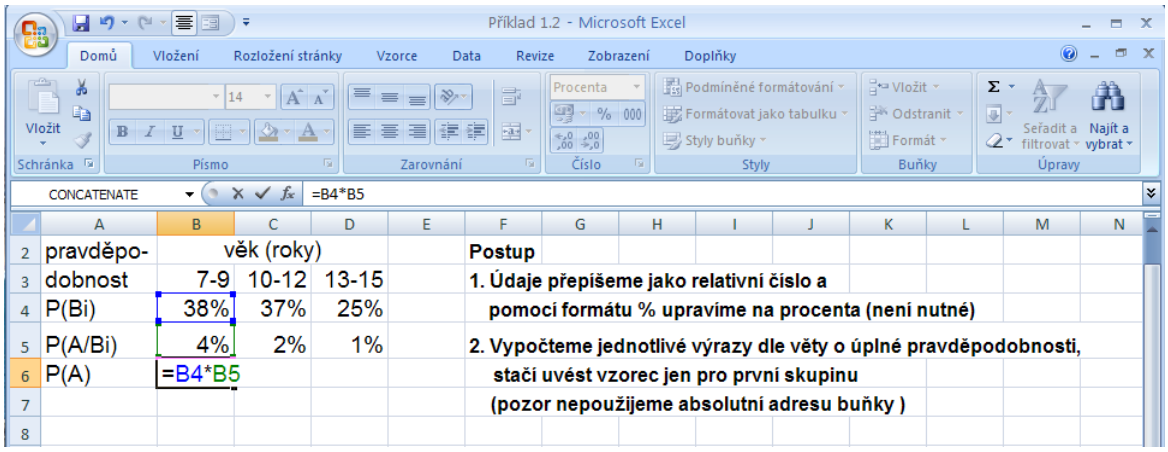
$$P(A) = P(A|B_1).P(B_1) + P(A|B_2).P(B_2) + \ldots + P(A|B_k).P(B_k)$$

Příklad 1.2
Na dětském táboře je 500 dětí, z nichž je 38 % ve věku 7 - 9 let, 37 % ve věku 10 – 12 let, 25 % ve věku 13 - 15 let.
Pravděpodobnost úrazu během jednoho tábora je (podle dlouhodobého sledování) v jednotlivých věkových kategorií postupně: 0,04 0,02 0,01.

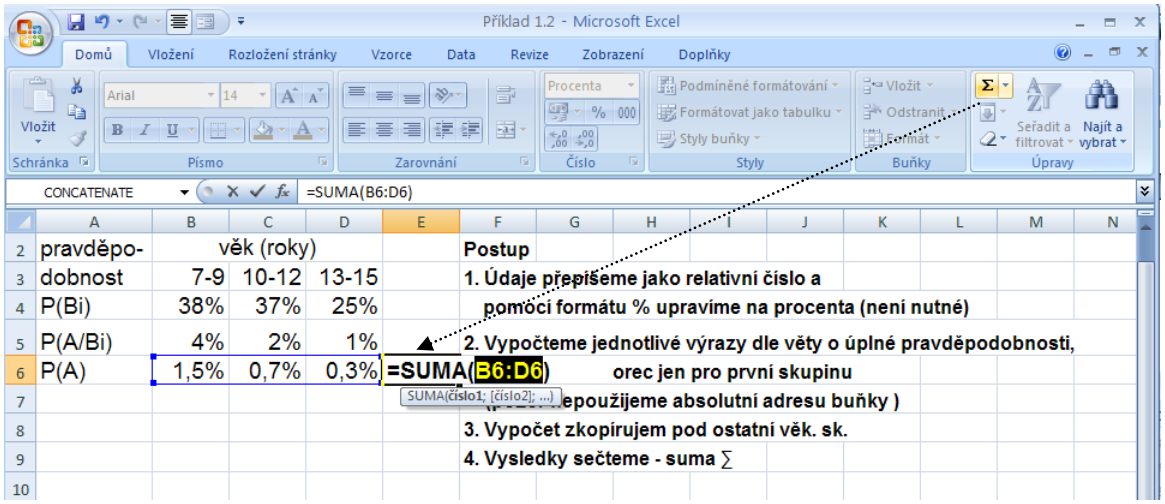
a) Jaká je pravděpodobnost vzniku úrazu na dětském táboře?
b) Kolik úrazů lze na táboře předpokládat?
c) Jak se změní předpokládaný počet úrazů, pokud se změní věkové složení - 65 %, 20 % a 15 %.

Označme: A - jev, že dojde k úrazu
B_1 věk 7 – 9 let, B_2 věk 10 – 12 let, B_3 věk 12 – 15 let
Jevy B_1, B_2, B_3 tvoří úplnou množinu jevů vzájemně neslučitelných (jeden z nich nastane, nastane právě jen jeden). Pravděpodobnosti těchto jevů jsou známé, jedná se o věkové složení souboru. Pravděpodobnosti úrazu v jednotlivých věkových skupinách (tj. pravděpodobnosti úrazu podmíněné věkem) jsou také známé a jsou:
$P(A|B_1) = 0{,}04$ $P(A|B_2) = 0{,}02$ $P(A/B_3) = 0{,}01$

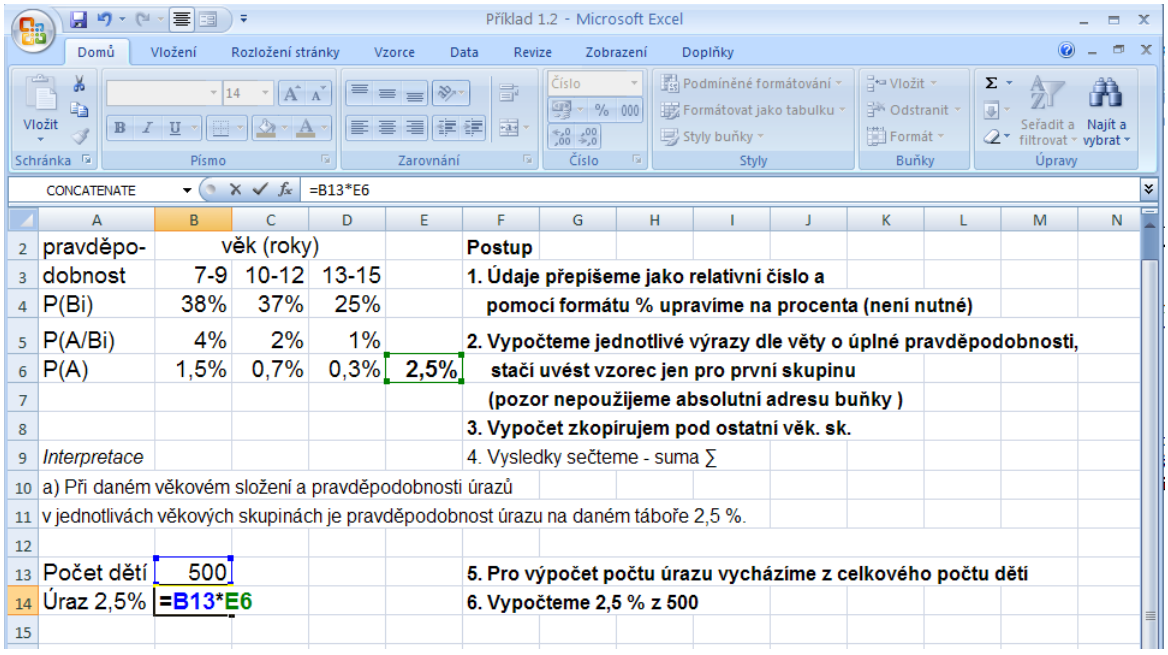
Řešení:



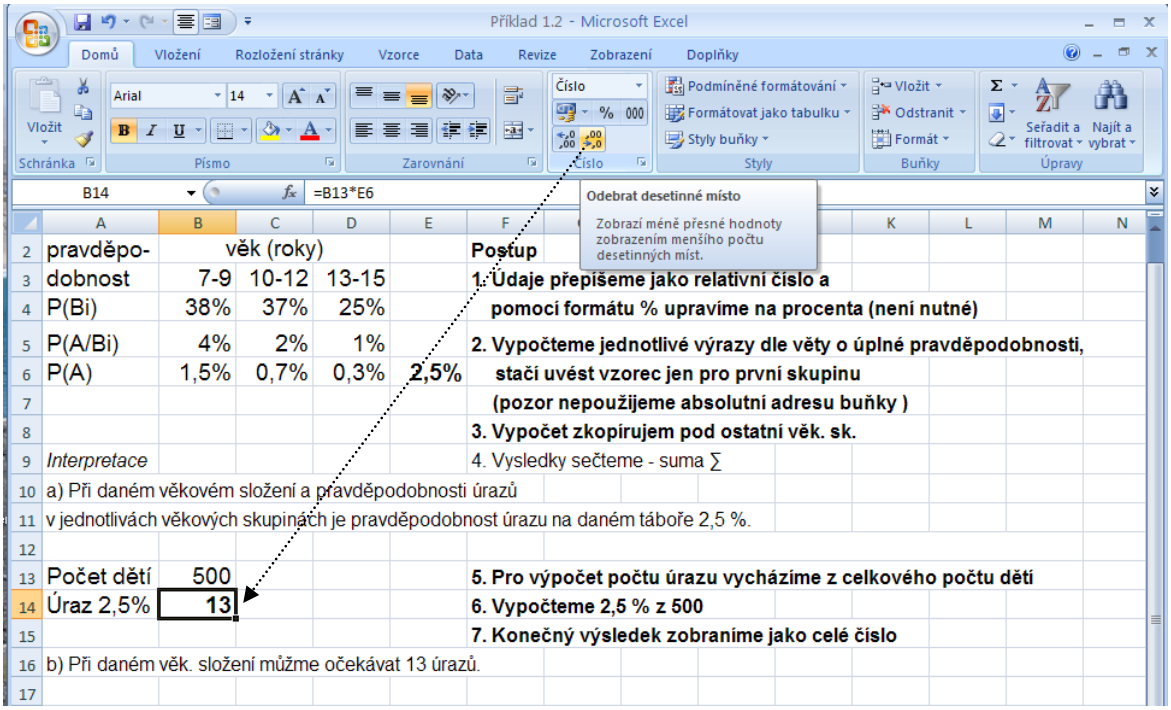
Obrázek č. 1.7 Krok 1 - 2



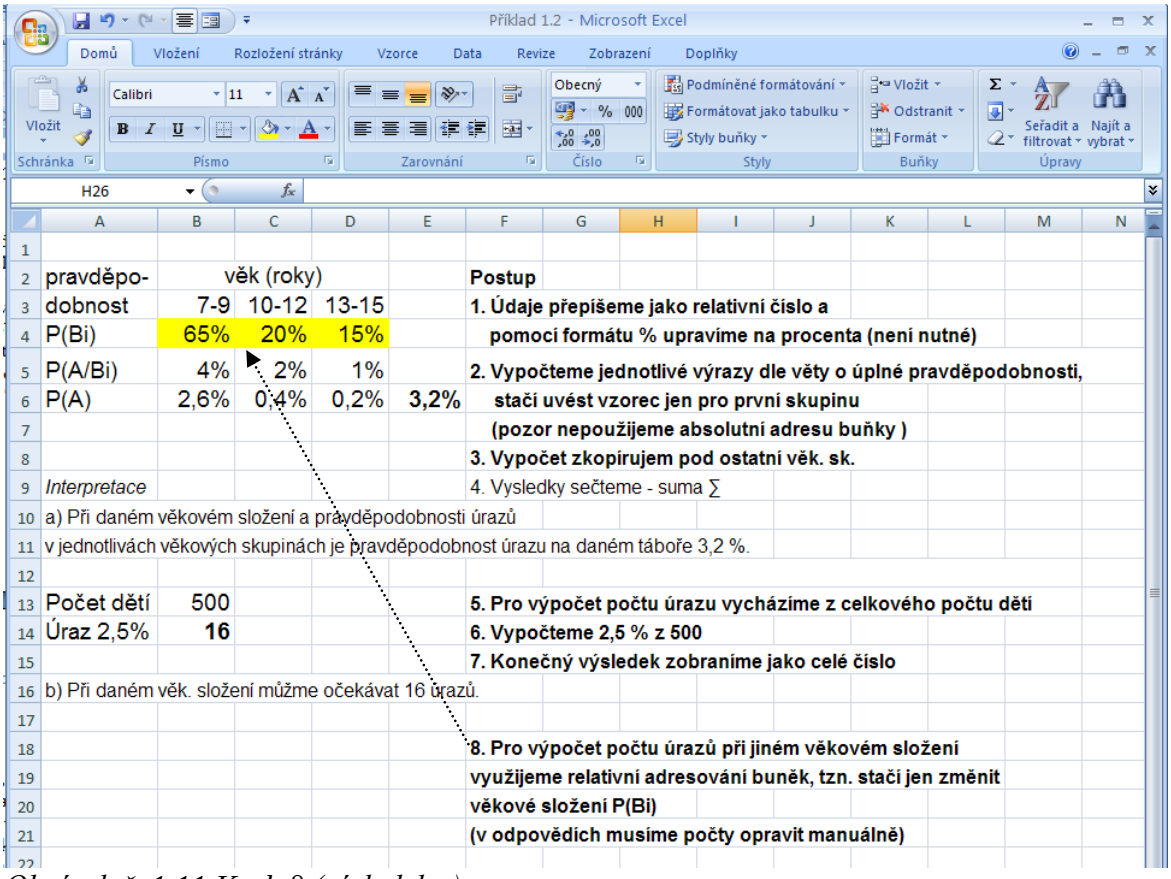
Obrázek č. 1.8 Krok 3 - 4



Obrázek č. 1.9 Krok 5 – 6 (výsledek a))



Obrázek č. 1.10 Krok 7(výsledek b))



Obrázek č. 1.11 Krok 8 (výsledek c)

1.2.3 Senzitivita, specificita, prediktivní hodnota

Tab. Označení výsledků diagnostického testu ve vztahu k diagnóze:

	Skutečnost	
Výsledek testu	Osoby s nemocí (N+)	Osoby bez nemoci (N-)
Pozitivní (T+)	SP	FP
Negativní (T-)	FN	SN

SP správně pozitivní, FP falešně poz., FN falešně negativní, SN správně neg.

Senzitivita – pravděpodobnost pozitivního výsledku testu u osob s nemocí

$$P(T+|N+) = SP/(SP + FN)$$

Specificita – pravděpodobnost negativního výsledku testu u osob bez nemocí

$$P(T-|N-) = SN/(SN + FP)$$

Pokud je v testovacím souboru stejné zastoupení nemocných osob, jaký je jejich podíl v celé populaci, potom platí pro výpočet prediktivních hodnot následující vztahy:

Pozitivní prediktivní hodnota – pravděpodobnost, že osoba s pozitivním výsledkem testu má nemoc:

$$P(N+|T+) = SP/(SP + FP)$$

Negativní prediktivní hodnota – pravděpodobnost, že osoba s negativním výsledkem testu je bez nemocí:

$$P(N-|T-) = SN/(SN + FN)$$

Pokud v testovacím souboru je jiný podíl nemocných osob, než je jejich zastoupení v populaci, pak se pro prediktivní hodnotu musí použít ***Bayesův vzorec*** se skutečnou prevalencí nemoci.

Bayesův vzorec pro pozitivní prediktivní hodnotu:

$$P(N^+ \mid T^+) = \frac{P(T^+ iN^+)}{P(T^+)} = \frac{P(T^+ \mid N^+).P(N^+)}{P(T^+ \mid N^+).P(N^+) + P(T^+ \mid N^-).P(N^-)}$$

Bayesův vzorec pro negativní prediktivní hodnotu:

$$P(N^- \mid T^-) = \frac{P(T^- iN^-)}{P(T^-)} = \frac{P(T^- \mid N^-).P(N^-)}{P(T^- \mid N^-).P(N^-) + P(T^- \mid N^+).P(N^+)}$$

$P(N^+)$ je pravděpodobnost výskytu nemoci v populaci, ze které je vybrán pozorovaný objekt (osoba). Nazývá se ***prevalence*** nemoci (podíl osob s nemocí ve zkoumané populaci). Jev N^- je jevem doplňkovým k jevu N^+ a jeho pravděpodobnost se vypočte

$P(N^-) = 1 - P(N^+)$.

$P(N^-)$ je vyjadřuje pravděpodobnost výskytu osob bez nemoci ve sledované populaci.

Příklad 1.3

Pro diagnostiku nemoci byl vyvinut nový test. Jeho kvalita se ověřovala na souboru osob, u kterých se po předchozím podrobném vyšetření vědělo, zda jsou nemocní nebo zdraví. Ze 135 nemocných byl test pozitivní u 108, ze 195 osob bez nemoci mělo 25 osob výsledek testu pozitivní.

Vypočítejte:

a) Senzitivitu testu

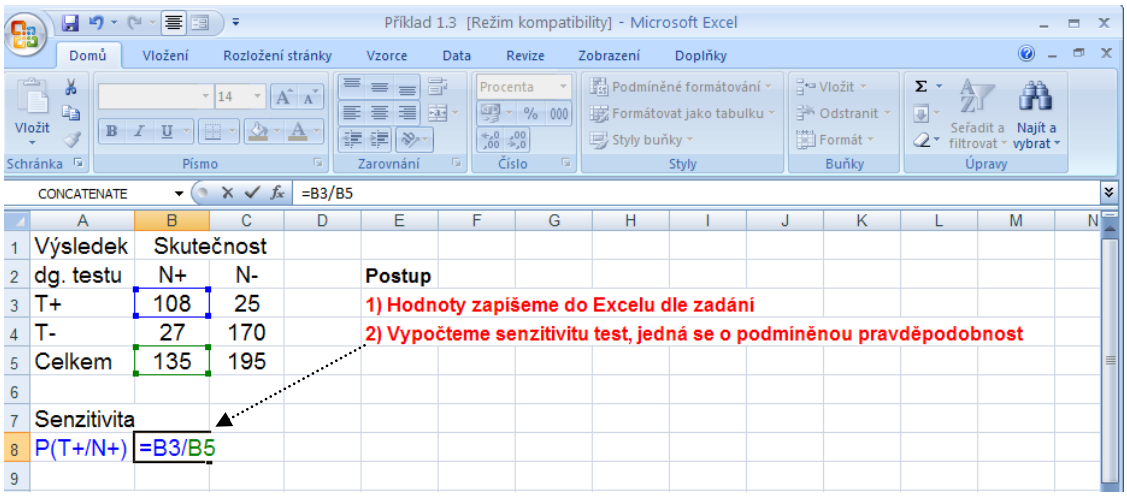
b) Specificitu testu

c) Pravděpodobnost nemoci u pacienta, který měl pozitivní výsledek testu a byl vybrán z populace, ve které je prevalence nemoci 0,03.

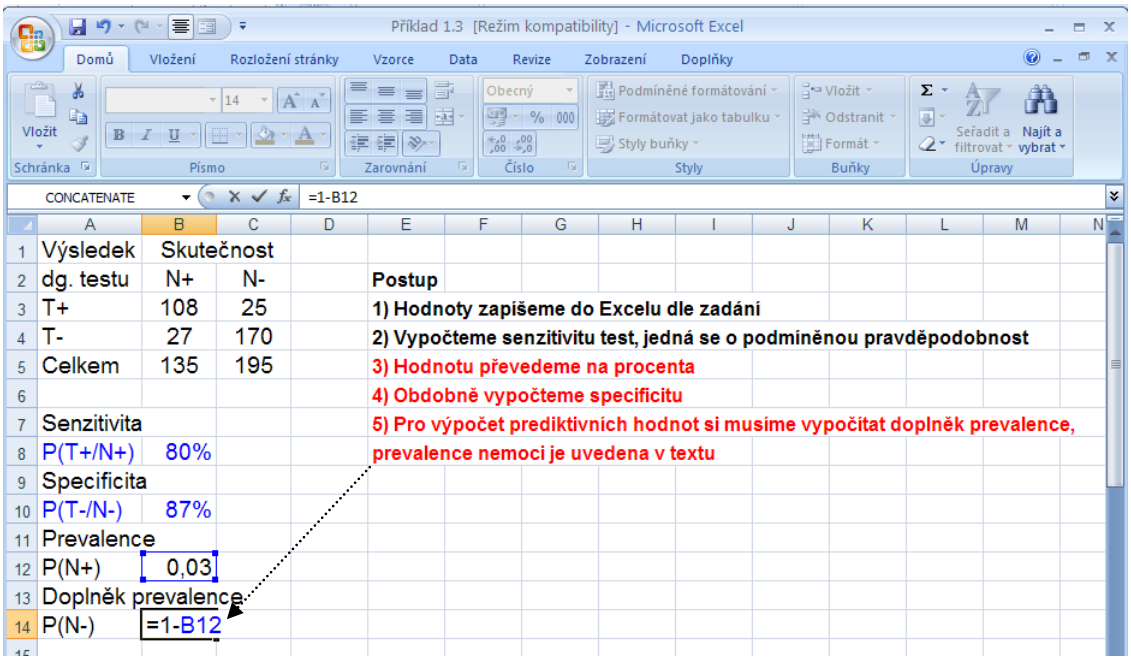
d) Pravděpodobnost, že osoba s negativním výsledkem testu nemá sledovanou nemoc, když prevalence v populaci, ze které byla osoba vybrána, je 0,03.

e) Jak se změní prediktivní hodnoty, pokud prevalence nemoci v populaci bude 0,2?

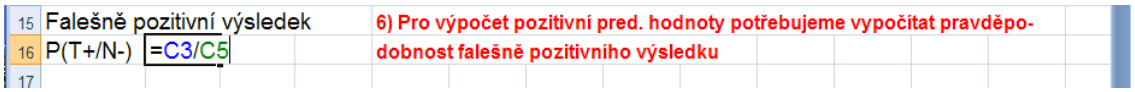
Řešení:



Obrázek č. 1.12 Krok 1 – 2

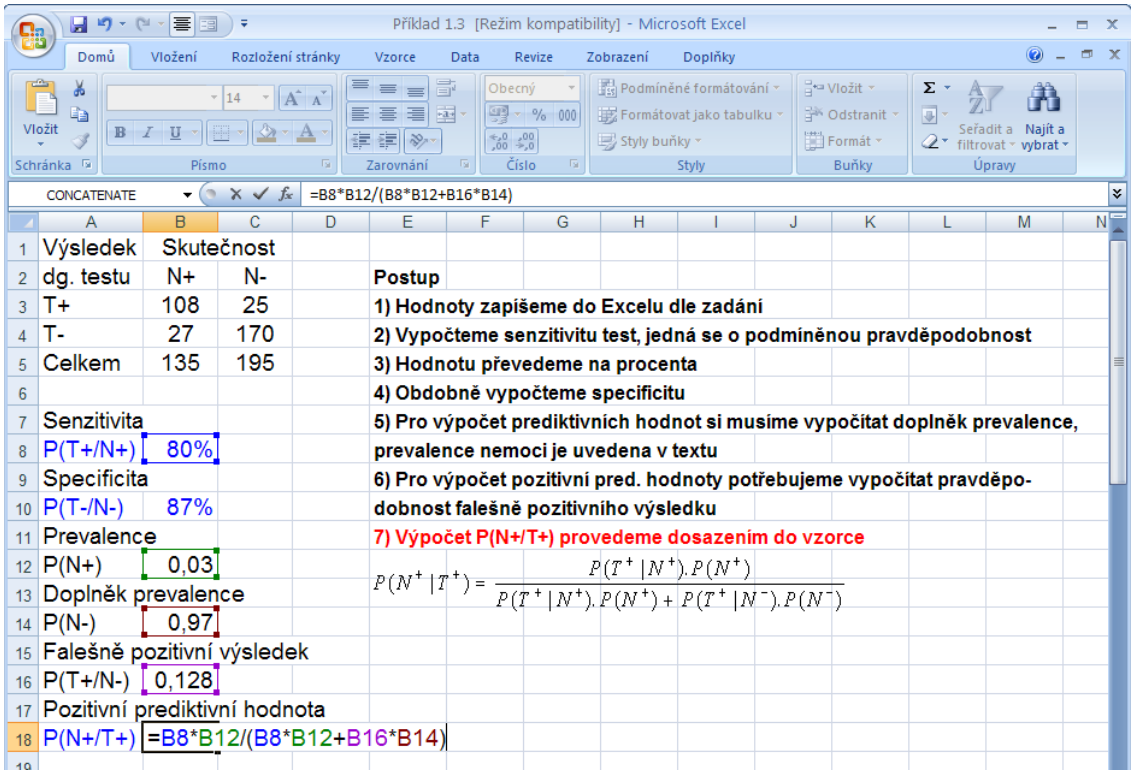


Obrázek č. 1.13 Krok 3 – 5(výsledek a), b))



15	Falešně pozitivní výsledek	6) Pro výpočet pozitivní pred. hodnoty potřebujeme vypočítat pravděpo-
16	P(T+/N-) =C3/C5	dobnost falešně pozitivního výsledku
17		

Obrázek č. 1.14 Krok 6



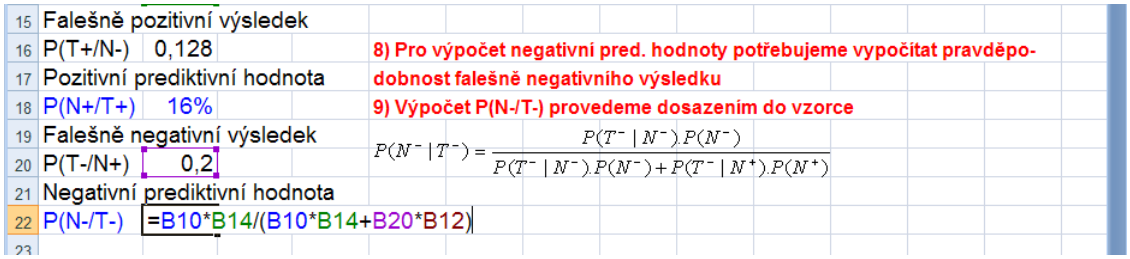
Příklad 1.3 [Režim kompatibility] - Microsoft Excel

Domů | Vložení | Rozložení stránky | Vzorce | Data | Revize | Zobrazení | Doplňky

CONCATENATE =B8*B12/(B8*B12+B16*B14)

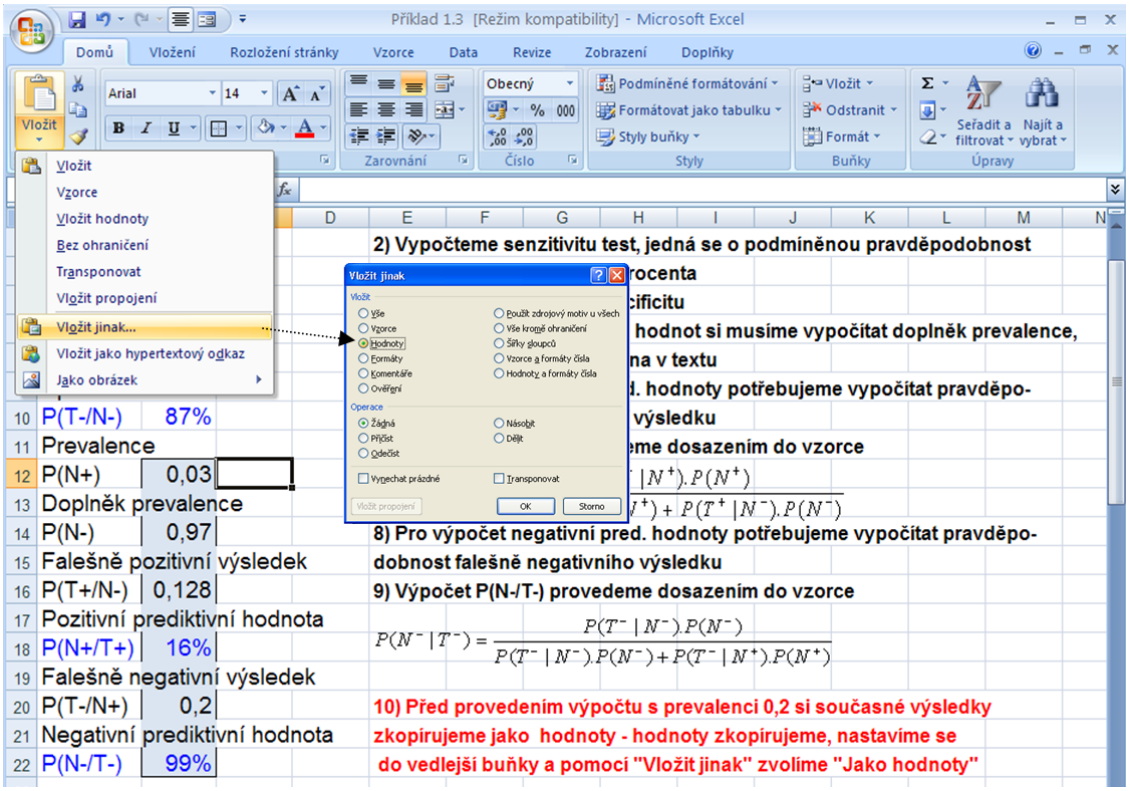
	A	B	C	E
1	Výsledek	Skutečnost		
2	dg. testu	N+	N-	**Postup**
3	T+	108	25	**1) Hodnoty zapíšeme do Excelu dle zadání**
4	T-	27	170	**2) Vypočteme senzitivitu test, jedná se o podmíněnou pravděpodobnost**
5	Celkem	135	195	**3) Hodnotu převedeme na procenta**
6				**4) Obdobně vypočteme specificitu**
7	Senzitivita			**5) Pro výpočet prediktivních hodnot si musíme vypočítat doplněk prevalence,**
8	P(T+/N+)	80%		**prevalence nemoci je uvedena v textu**
9	Specificita			**6) Pro výpočet pozitivní pred. hodnoty potřebujeme vypočítat pravděpo-**
10	P(T-/N-)	87%		**dobnost falešně pozitivního výsledku**
11	Prevalence			**7) Výpočet P(N+/T+) provedeme dosazením do vzorce**
12	P(N+)	0,03		$P(N^+ \mid T^+) = \frac{P(T^+ \mid N^+).P(N^+)}{P(T^+ \mid N^+).P(N^+) + P(T^+ \mid N^-).P(N^-)}$
13	Doplněk prevalence			
14	P(N-)	0,97		
15	Falešně pozitivní výsledek			
16	P(T+/N-)	0,128		
17	Pozitivní prediktivní hodnota			
18	P(N+/T+)	=B8*B12/(B8*B12+B16*B14)		

Obrázek č. 1.15 Krok 7

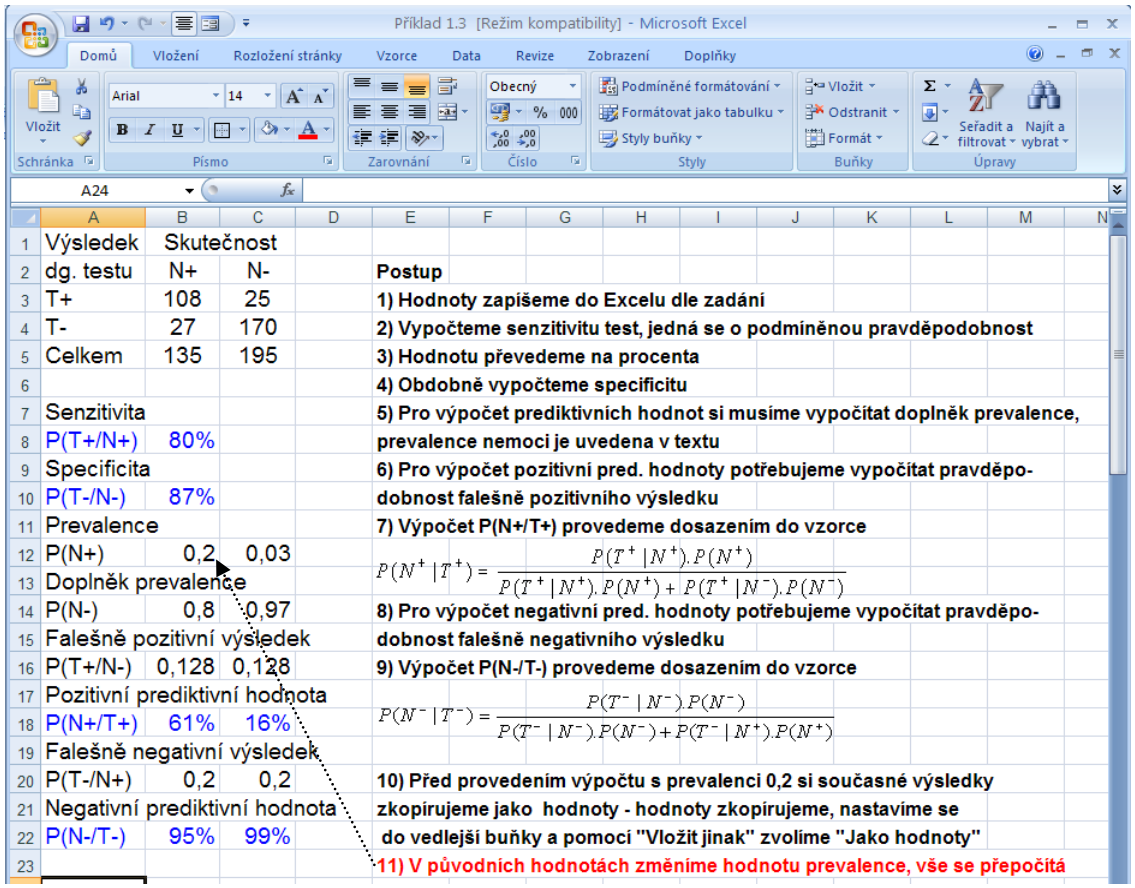


15	Falešně pozitivní výsledek	
16	P(T+/N-) 0,128	**8) Pro výpočet negativní pred. hodnoty potřebujeme vypočítat pravděpo-**
17	Pozitivní prediktivní hodnota	**dobnost falešně negativního výsledku**
18	P(N+/T+) 16%	**9) Výpočet P(N-/T-) provedeme dosazením do vzorce**
19	Falešně negativní výsledek	$P(N^- \mid T^-) = \frac{P(T^- \mid N^-).P(N^-)}{P(T^- \mid N^-).P(N^-) + P(T^- \mid N^+).P(N^+)}$
20	P(T-/N+) 0,2	
21	Negativní prediktivní hodnota	
22	P(N-/T-) =B10*B14/(B10*B14+B20*B12)	

Obrázek č. 1.16 Krok 8 -9 (výsledek c))



Obrázek č. 1.17 Krok 10 (výsledek d))



Obrázek č. 1.18 Krok 11 (výsledek e))

Při změně (zvýšení) prevalence došlo ke zvýšení pozitivní prediktivní hodnoty a naopak došlo ke snížení negativní prediktivní hodnoty.

Shrnutí obsahu kapitoly

Biostatistika používá statistické metody pro vyhodnocení biomedicínských dat. Pro pochopení statistiky je nutné znát pojem pravděpodobnost. Ze základních pravidel počítání s pravděpodobností vychází výpočty založené na větě o úplné pravděpodobnosti, výpočet senzitivity, specificity a prediktivních hodnot u diagnostických testů.

MS Excel - funkce suma, absolutní a relativní adresa buňky, zadání vzorce, kopírování, vložení jinak – hodnoty, formát %, nastavení počtu desetinných míst.

Kontrolní otázky:

1. Čím se zabývá biostatistika?
2. Co je to pravděpodobnost?
3. Kdy můžeme využít výpočet založený na větě o úplné pravděpodobnosti?
4. Co je senzitivita a specificita diagnostického testu?

Korespondenční úkol:

1. V odborném časopise se vyhledejte článek se zdravotnickou problematiku, který zpracovává data (epidemiologická studie). Připravte v PowerPointu stručný popis studie – práce ve dvojících (max. 5-8 min.).

   Možné zdroje:
   časopis Hygiena http://www1.szu.cz/svi/hygiena/index.php
   Časopisy ČLS JEP http://www.cls.cz/casopisy-clsjep
   CEJPH http://www1.szu.cz/svi/cejph/index.php
   Mezinárodní časopisy http://www.ncbi.nlm.nih.gov/pubmed/

Úkoly

1.1 Jaká bude koncentrace výsledného roztoku, pokud smícháme obsah 5 nádob s roztokem různé koncentrace. Údaje jsou uvedeny v tabulce.

Nádoba	1	2	3	4	5
Koncentrace	20 %	50 %	25 %	35 %	46 %
Objem (l)	3	2	7	5	3

(pozn. objem se musí převést na relativní údaj)

1.2 Jaká je pravděpodobnost vítězství jisté politické strany, když by ji z voličů do 60 let volilo 65 % osob a z voličů ve věku 60 a více let jen 28 %. Voliči do 60 let tvoří 76 % populace s a osoby starší 24 % populace (pozn. voliči jsou osoby ve věku 18 a více let).

1.3 Ověřování kvality nového testu pro diagnostiku poruchy sluchu, bylo provedeno na osobách, u nichž byl stav sluchu vyšetřen dříve podrobnými klinickými postupy. Soubor byl vytvořen uměle, proto prevalence poruch sluchu v tomto souboru nemusí odpovídat skutečné prevalenci poruch v populaci. Pozitivní výsledek testu by zjištěn u 78 osob z 90 osob s poruchou sluchu a u 15 osob z 300 osob bez poruchy sluchu.

a) Vypočtěte senzitivitu a specificitu testu.

b) Vypočtěte prediktivní hodnoty v případě, že prevalence poruch sluchu v populaci je 0,15.

2 Databáze, kontingenční tabulky, grafy

V této kapitole se dozvíte:

- Co je databáze, typy veličin.
- Jaké jsou nejčastější metody sběru dat.
- Základní zásady pro vytváření dotazníku.
- Jak se vytvářejí kontingenční tabulky, co obsahují.
- Sestavní vybraných typů grafů.

Po jejím prostudování byste měli být schopni:

- Vytvořit databázi v MS Excelu.
- Použít funkce KDYŽ().
- Vytvořit kontingenční tabulku z jedné a ze dvou veličin (1 a 2-rozměrné kontingenční tabulky).
- Sestavit a upravit graf z jedné a ze dvou veličin.

Klíčová slova této kapitoly:

Databáze, veličina, frekvenční (kontingenční) tabulka, graf.

Doba potřebná ke studiu a zpracování úkolů: 4 - 5 hodin

Průvodce studiem

Druhá kapitola je více zaměřena na praktické zpracování dat. Předpokládá zvládnutí základních operací v MS Excelu, které byly probírány v kapitole 1. Náročnost probírané problematiky je individuální v závislosti na znalostech MS Excelu získaných na střední škole.
Teorie k uvedeným tématům je probírána v první přednášce Lékařská biofyzika, výpočetní technika I – Biostatistika.

2.1 Databáze

Objektem studia statistiky je soubor (kolektiv, skupina, populace). Jevy, kterými se statistika zabývá, jsou vždy nějakým způsobem vázány na určitý soubor, označují se proto jako ***jevy hromadné***. Na všech objektech souboru probíhá měření, které zjišťuje vlastnosti objektů – definují se ***veličiny*** *(znaky)*. Tyto údaje se označují jako ***data*** a vytvářejí ***databázi***. Data můžeme sbírat různými metodami.

2.1.1 Metody sběru dat

1. *Observační metody* – jedná se o přímé měření nebo metody vyžadující speciální znalosti a techniku, např. klinické vyšetření, biochemické vyšetření, mikrobiologické vyšetření, psychologické testy nebo genetické vyšetření.
2. *Rozhovor a dotazník* – shromažďují údaje prostřednictvím záměrně cílených otázek. Získané informace mohou však být zkresleny

nepochopením otázek, špatným záznamem odpovědí a při rozhovoru také vlivem sociální interakce.

3. Použití *dokumentace* – je poměrně jednoduchý způsob sběru dat, který poskytuje objektivní informace i o minulosti. Jedná se o původní zdravotní dokumentace, jako jsou záznamy o zdraví a nemoci, hlášení o narození dítěte nebo list o prohlídce mrtvého a údaje rutinní zdravotnické statistiky, statistik jiných odvětví apod.
4. Další způsob sběru dat je *modelování*, např. modely znečištění ovzduší, hlukové modely, které jsou schopny na základě vstupních údajů provést predikci, ale také informovat o stavu v minulosti.

Jeden z nejčastějších nástrojů získávání dat je ***dotazník.*** Při sestavení dotazníků je třeba dodržovat jisté zásady, které ulehčí převod dat do elektronické podoby a tím i redukují procento chyb, které by vznikly vkládáním dat.

2.1.2 Zásady při sestavování dotazníků

- Zavést označení dotazníků - pořadové číslo, rodné číslo, ….
- Zavést jednotné kódování např. ano – 1, ne – 0
- Pokud se jedná o odpověď s více možnostmi, možnosti označit číslem
  např. Jak často sportujete během týdne : 1 vůbec
  2 příležitostně
  3 pravidelně
- Při kódování používat číslice ne písmena
- V případě více odpovědí, tuto možnost uvést u otázky
  např. Jakými zdravotními potížemi trpíte? (*Můžete vybrat více odpovědí.*)
  1) bolestmi hlavy
  2) bolestmi zad
  3) bolestmi dolních končetin
  4) bolestmi v oblasti dýchacích cest
  5) zažívacími obtížemi
  6) gynekologickými obtížemi
  7) jiné – specifikujte ……..
- Předcházet otevřeným otázkám tzn. u odpovědí nabídnout max. možný výčet odpovědí.
- V případě, že otázky jsou podmíněné předchozí odpovědí, upozornit na přechod na další otázky i graficky (odsazení).
- Dotazník vytvářet přehledně jak pro dotazovaného, tak pro pracovníka, který bude data vkládat.

2.1.3 Vytváření databáze v MS Excelu

1. V prvním řádku zavedeme označení položek, jeden sloupec = jedna veličina - jedna otázka (položka) v dotazníku. Pro označení používáme krátké výstižné označení, případně použijeme označení např. O1 – první otázka apod.
2. V prvním sloupci by mělo být umístěno identifikační označení respondenta (objektu).

3. V dalších řádcích jsou již data u jednotlivých respondentů, jeden řádek = jeden respondent (1 objekt).
4. V databázi nesmí být sloupce bez označení.
5. V případě chybějících údajů se políčko nevyplňuje, tzn. neuvedeme „x“ nebo podobný znak.

Příklad 2.1

Při dotazníkovém šetření byly zjišťovány demografické údaje a informaci o kouření. Dotazník k tomuto šetření je na obrázku 2.1 a databáze vytvořená v MS Excelu na obrázku 2.2.

Dotazníkové šetření

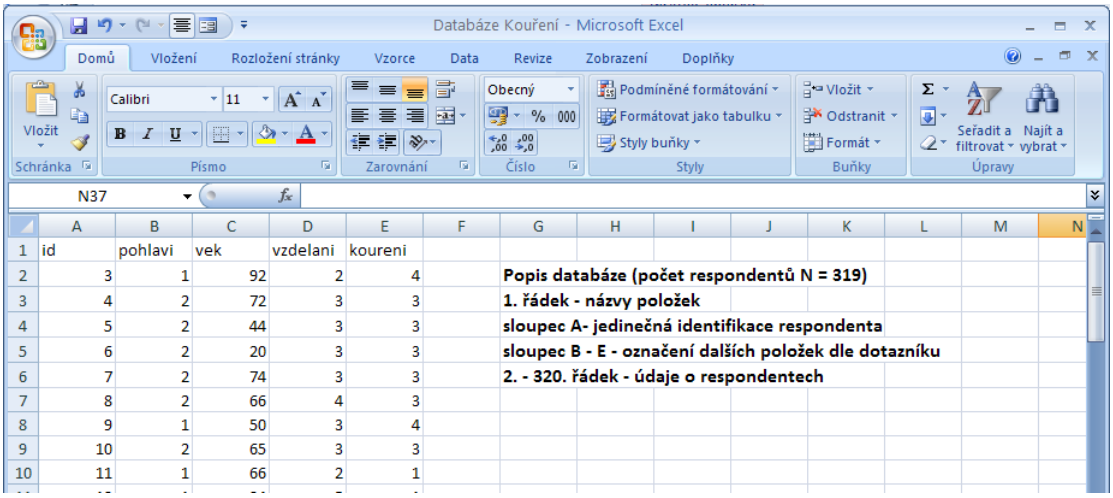
ID: ……..

1. Pohlaví: 1 muž 2. Kolik je Vám let?
2 žena *Uveďte v celých letech* …….

3. Jaké je Vaše nejvyšší dokončené vzdělání?
1 Základní
2 Střední odborné bez maturity
3 Střední odborné s maturitou
4 Střední všeobecné s maturitou
5 Vyšší odborné /pomaturitní/
6 Vysokoškolské
7 Jiné: ……

4.Kouříte v současnosti?
1 Ano, pravidelně
2 Ano, příležitostně
3 Ne, nikdy jsem nekouřil/a
4 Ne, přestal/a jsem kouřit

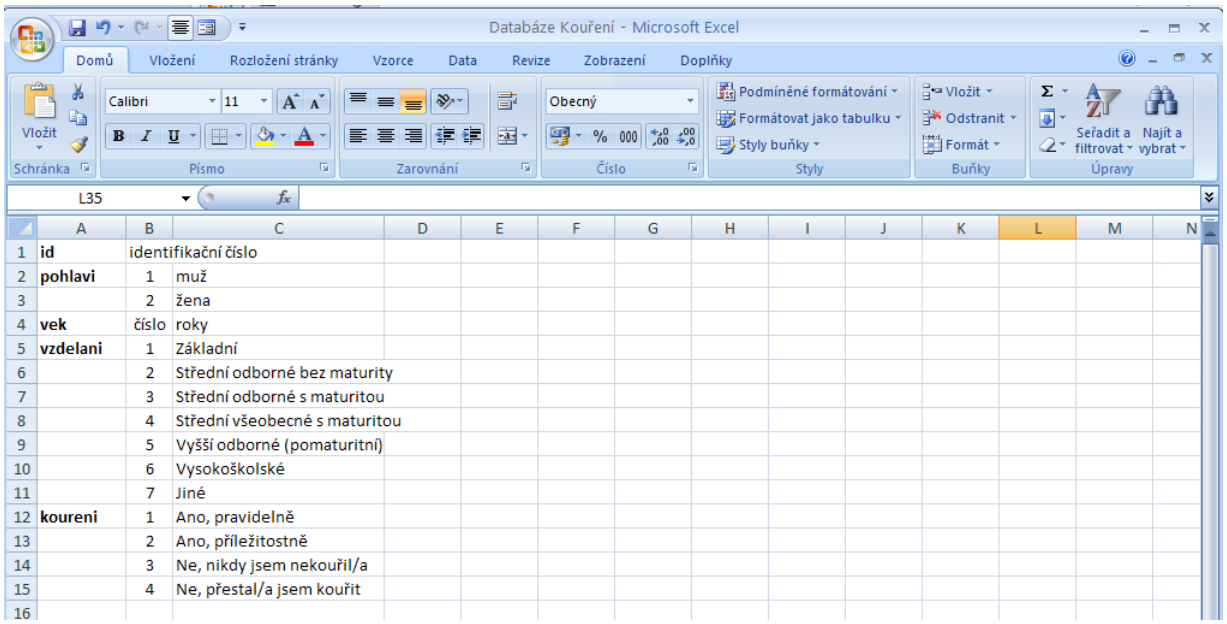
Obr. 2.1 Ukázka dotazníku



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	id	pohlavi	vek	vzdelani	koureni									
2	3	1	92	2	4		**Popis databáze (počet respondentů N = 319)**							
3	4	2	72	3	3		**1. řádek - názvy položek**							
4	5	2	44	3	3		**sloupec A- jedinečná identifikace respondenta**							
5	6	2	20	3	3		**sloupec B - E - označení dalších položek dle dotazníku**							
6	7	2	74	3	3		**2. - 320. řádek - údaje o respondentech**							
7	8	2	66	4	3									
8	9	1	50	3	4									
9	10	2	65	3	3									
10	11	1	66	2	1									

Obr. 2.2 Ukázka databáze v MS Excelu

Při vytvoření databáze je nutné provést popis jednotlivých veličin (sledovaných znaků) (obr. 2.3) .



	A	B	C
1	**id**	identifikační číslo	
2	**pohlavi**	1	muž
3		2	žena
4	**vek**	číslo	roky
5	**vzdelani**	1	Základní
6		2	Střední odborné bez maturity
7		3	Střední odborné s maturitou
8		4	Střední všeobecné s maturitou
9		5	Vyšší odborné (pomaturitní)
10		6	Vysokoškolské
11		7	Jiné
12	**koureni**	1	Ano, pravidelně
13		2	Ano, příležitostně
14		3	Ne, nikdy jsem nekouřil/a
15		4	Ne, přestal/a jsem kouřit
16			

Obr. 2.3 Popis veličin v databázi

2.2 Kontingenční tabulky, grafy

Veličiny (znaky), které popisují dané objekty se dělí na kvalitativní a kvantitativní (tab. 2.1).

Tab. 2.1 Typy veličin

Veličiny	**Stupnice**	**Příklady**
Kvalitativní (Kategoriální)	Nominální	Krevní skupiny, pohlaví, národnost, druh onemocnění
	Ordinální	Závažnost poranění, stupeň invalidity, míra spokojenosti
Kvantitativní (Metrické)	Intervalová	Teplotní stupnice
	Poměrová	BMI, krevní tlak, hodnota cholesterolu

Základní operací při zpracován dat je ***třídění*** podle určitého kvalitativního znaku např. pohlaví – muži, ženy. Výsledkem třídění jsou ***četnostní (frekvenční) tabulky***, které udávají počty prvků souboru patřících do určité kategorie sledovaného znak, jedná se o ***absolutní četnosti***. ***Relativní četnosti*** udávají tyto počty v procentech, podíl objektů z celkového počtu. Rozdělení četností sledovaného znaku můžeme zobrazit i graficky. V MS Excelu se četnostní (frekvenční) tabulky nazývají ***kontingenční tabulky***.

V některých případech jsou data kvantitativní převedena na data kvalitativní (ordinální), provádí se tzv. degradace. Např. hodnoty BMI jsou rozděleny do kategorií - normální hodnoty, nadváha, obezita. Při vytváření kategorií musí být přesně definovány hraniční hodnoty pro jednotlivé kategorie. V MS Excelu

je pro vytváření kategorií možné použít funkci KDYŽ(). Platí, že počet funkcí KDYŽ je o jeden méně něž počet kategorií.

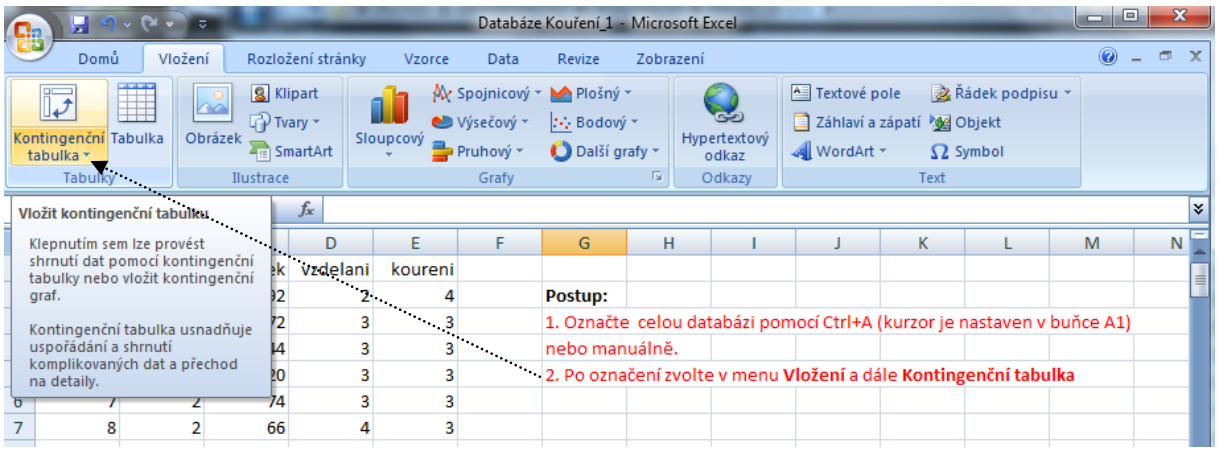
Příklad 2.2

Z databáze z příkladu 2.1 zjistěte:

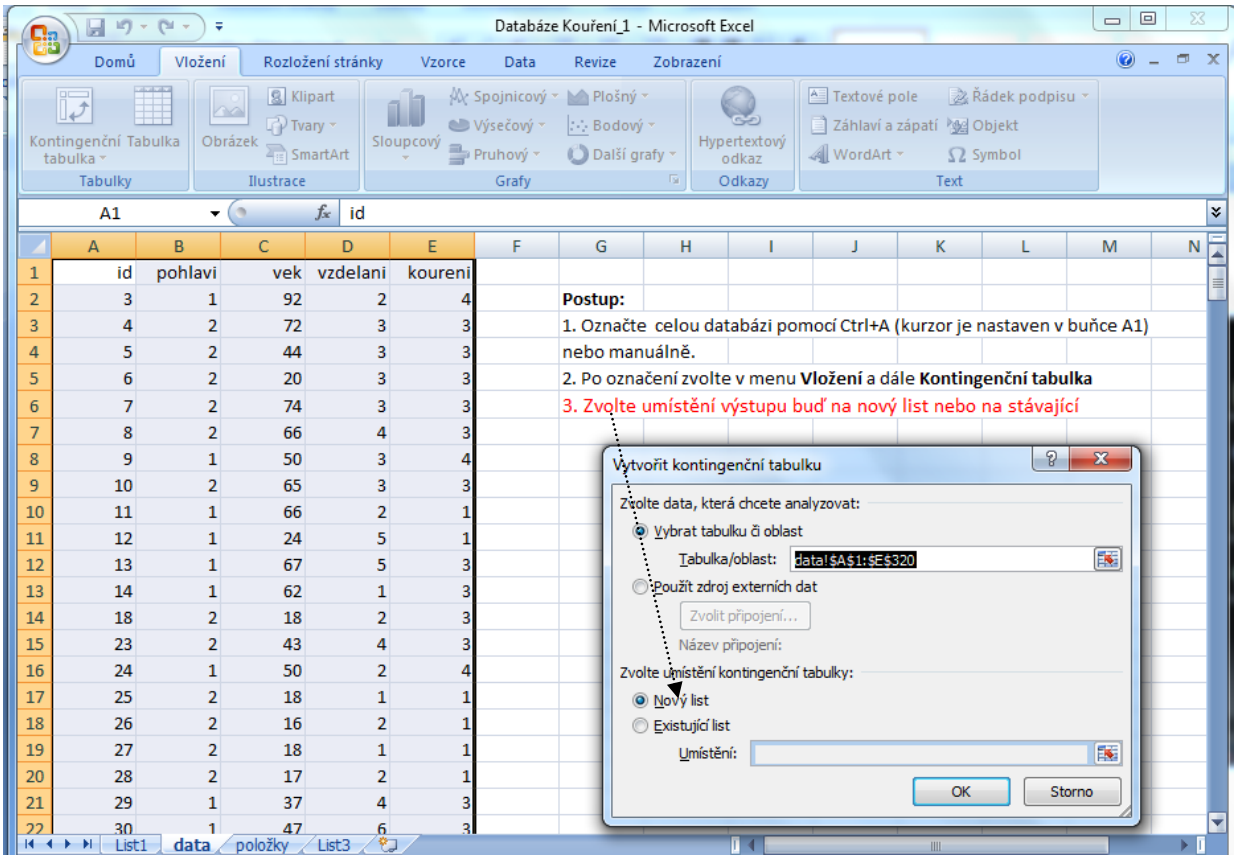
a) Kolik je v databázi mužů a žen (sestavte kontingenční tabulku, vyjádřete v absolutních i relativních číslech), sestavte graf.
b) Rozdělte respondenty do 6 věkových skupiny po 10 letech (1. skupina <=30 let, 2. skupina <=40, ……., 6. skupina - více než 70 let).
c) Zjistěte kolik respondentů patří do vytvořených věkových skupin (absolutní i relativní počty) a sestavte graf.
d) Vyjádřete věkové zastoupení mužů a žen (sestavte kontingenční tabulku se 2 znaky), věkové zastoupení mužů a žen znázorněte graficky.

Řešení

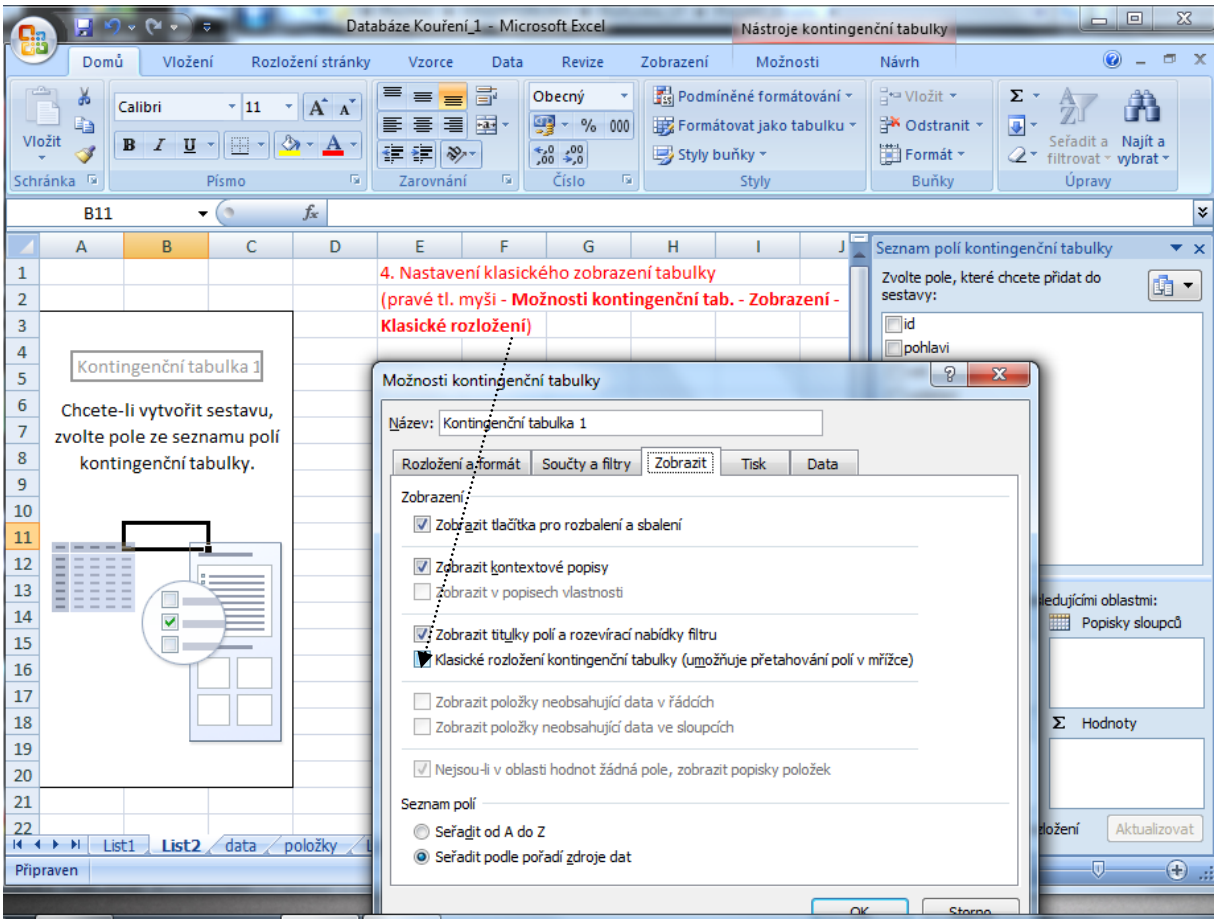
a) Na obrázcích 2.4 až 2.11 je popsán postup pro sestavení kontingenční tabulky pro jeden znak. Tabulka se muže zobrazit ve formátu sestava, pak je nutné ji přepnout do klasického zobrazení (obr. č. 2.6). V tabulce jsou po úpravě zobrazeny absolutní počty, kopírování, převod na hodnoty a výpočet % (relativních počtů) je uveden v kapitole 1.



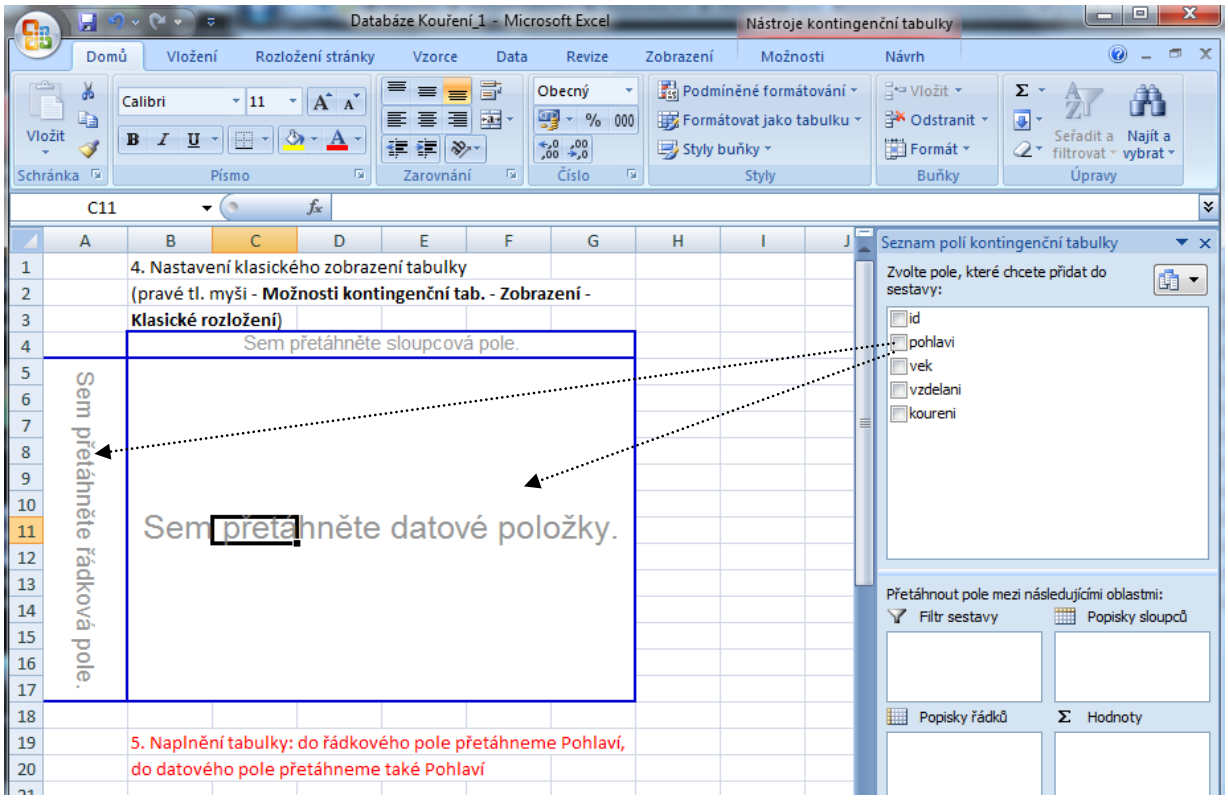
Obr. 2.4 Krok 1a) a 2a)



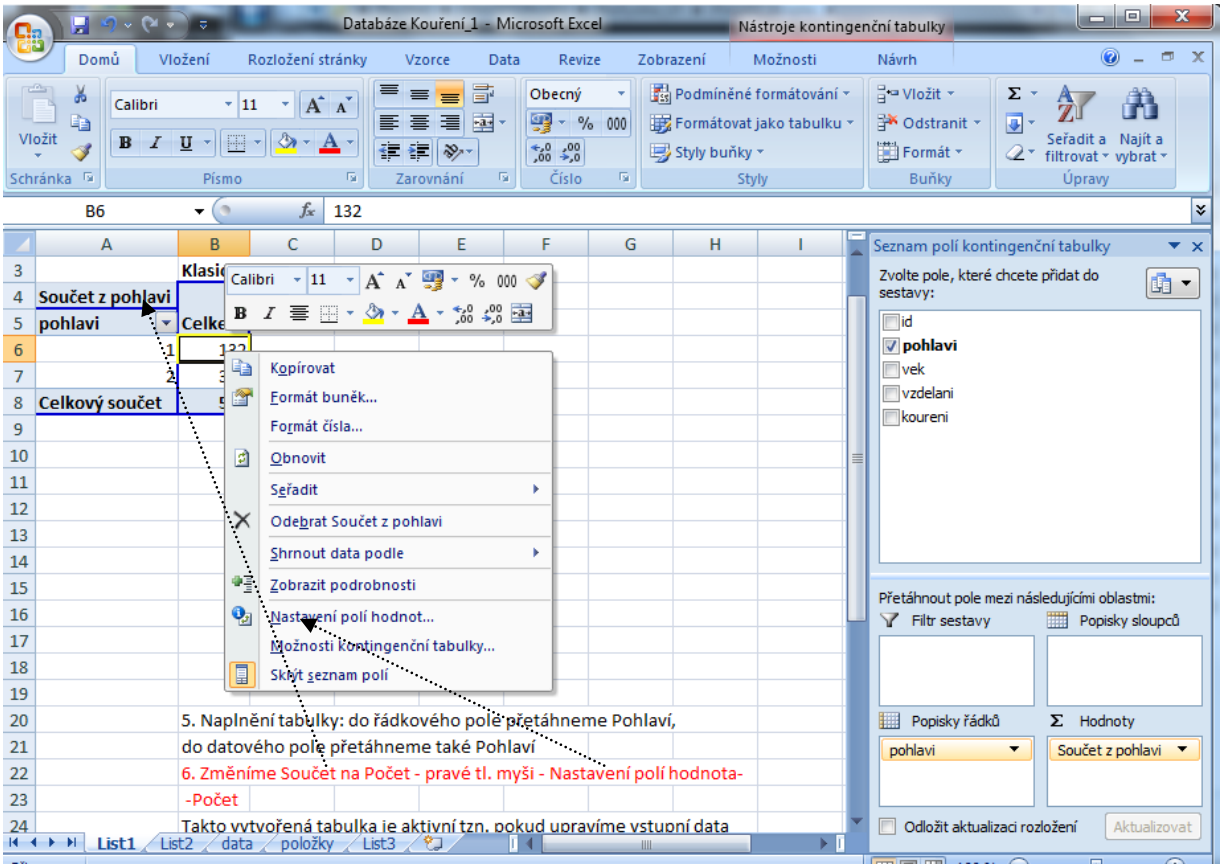
Obr. 2.5 Krok 3a) Označení dat a umístění tabulky



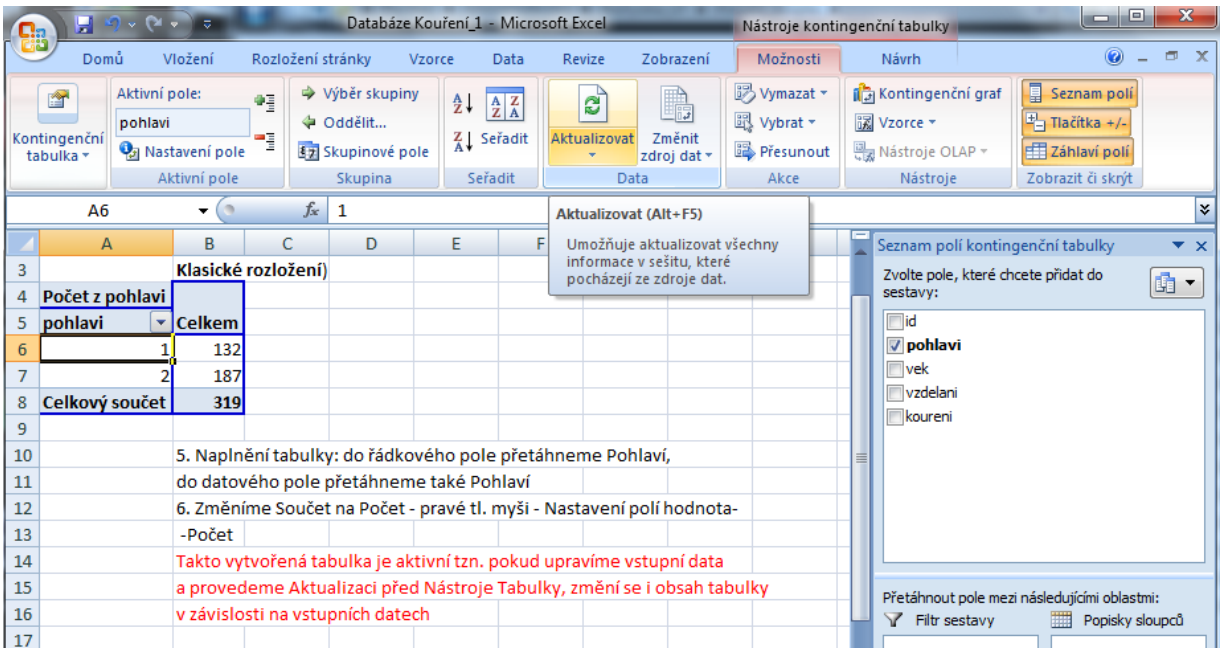
Obr. 2.6 Krok 4a) Klasické zobrazení tabulky



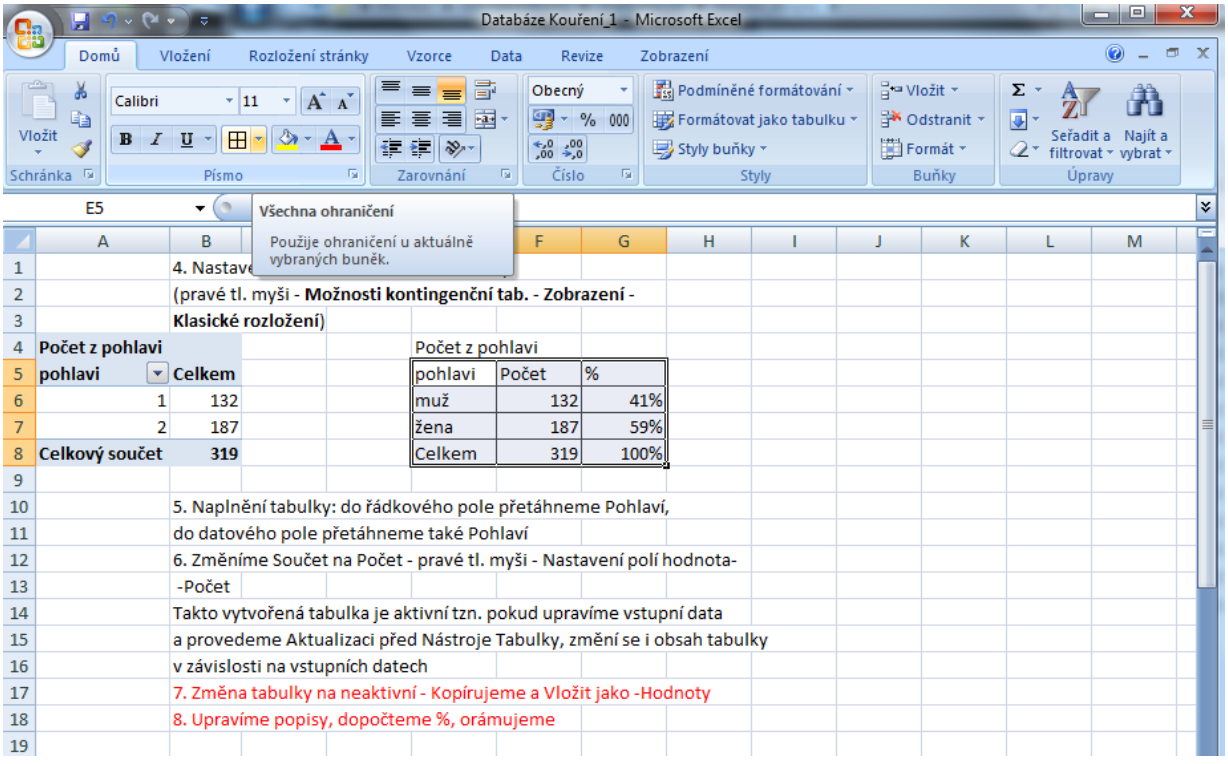
Obr. 2.7 Krok 5a) Naplnění tabulky



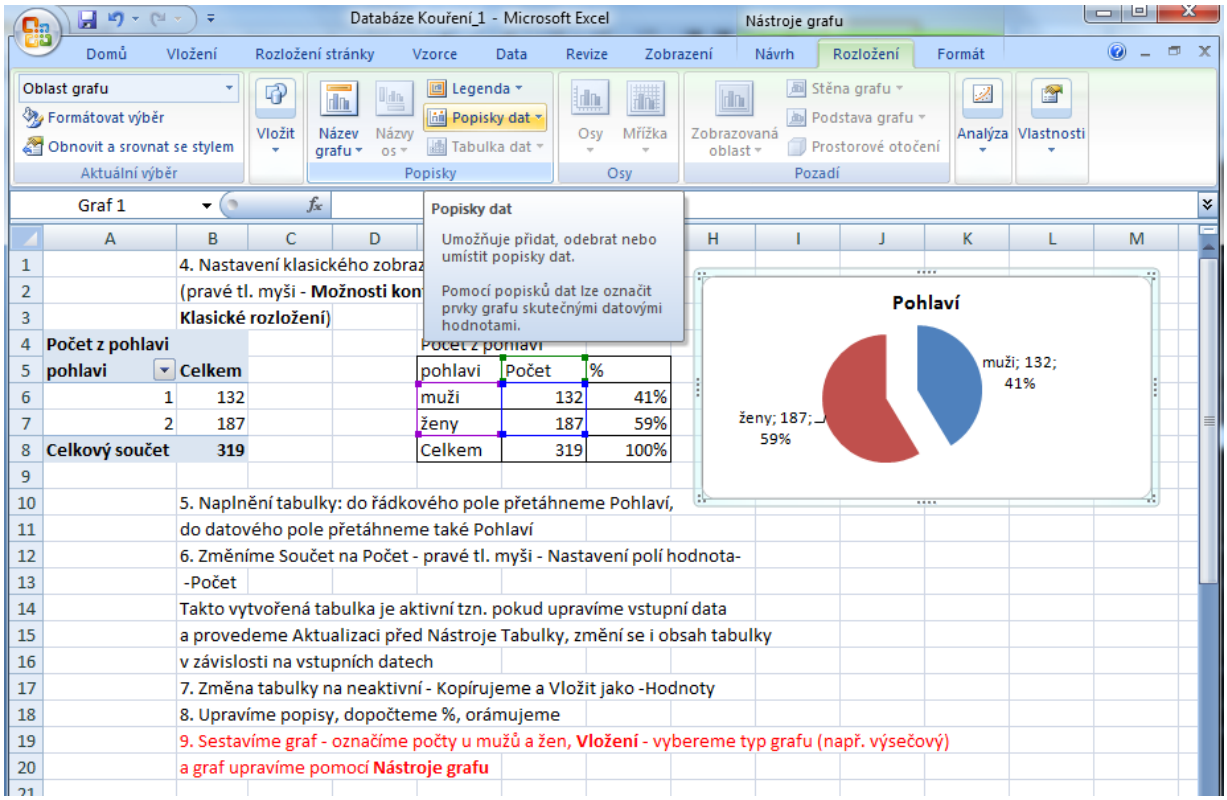
Obr. 2.8 Krok 6a) Změna Součtu na Počet



Obr. 2.9 Krok 6a) Možnost aktualizovat data v tabulce

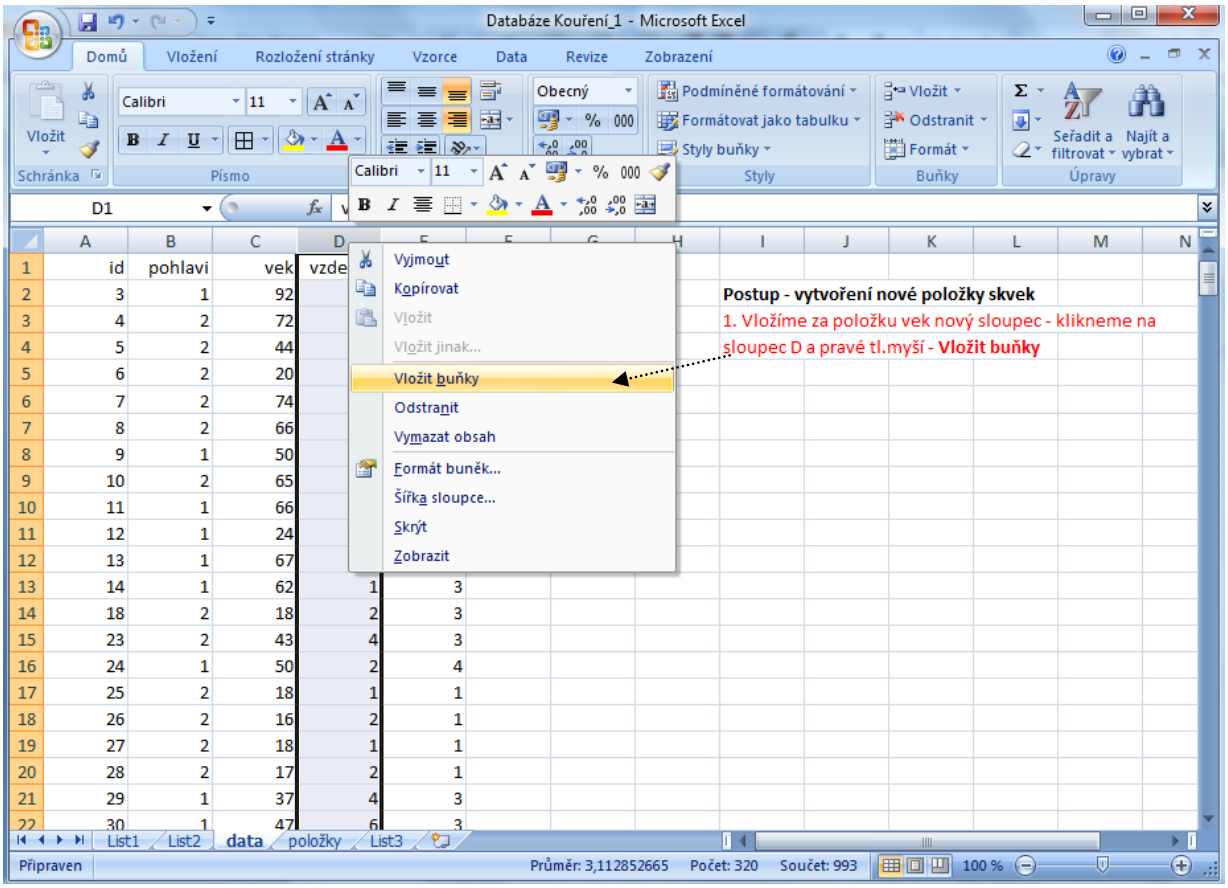


Obr. 2.10 Krok 7a) - 8a) Vytvoření tabulky s absolutními a relativními počty (řešení a))

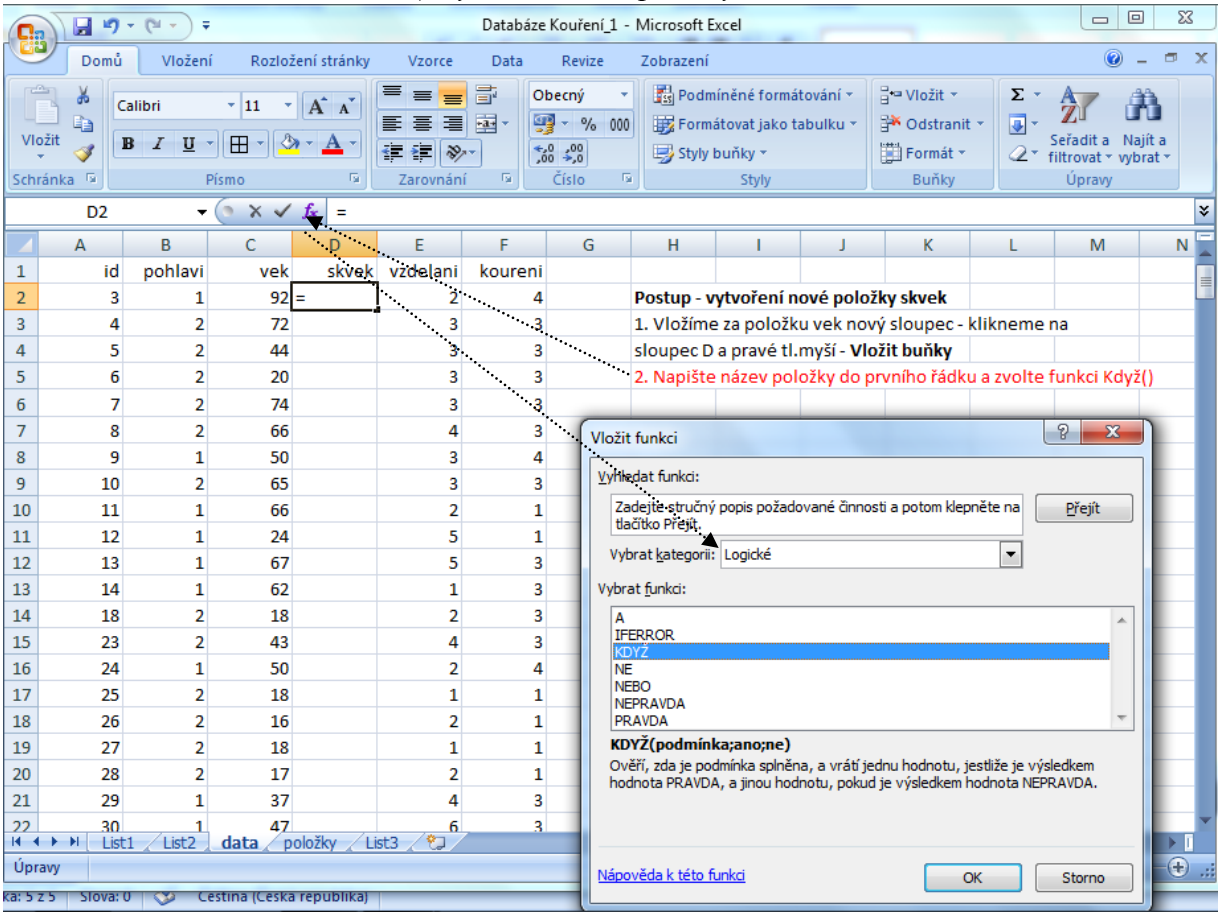


Obr. 2.11 Krok 9a) Vytvoření grafu (dokončení úkolu a))

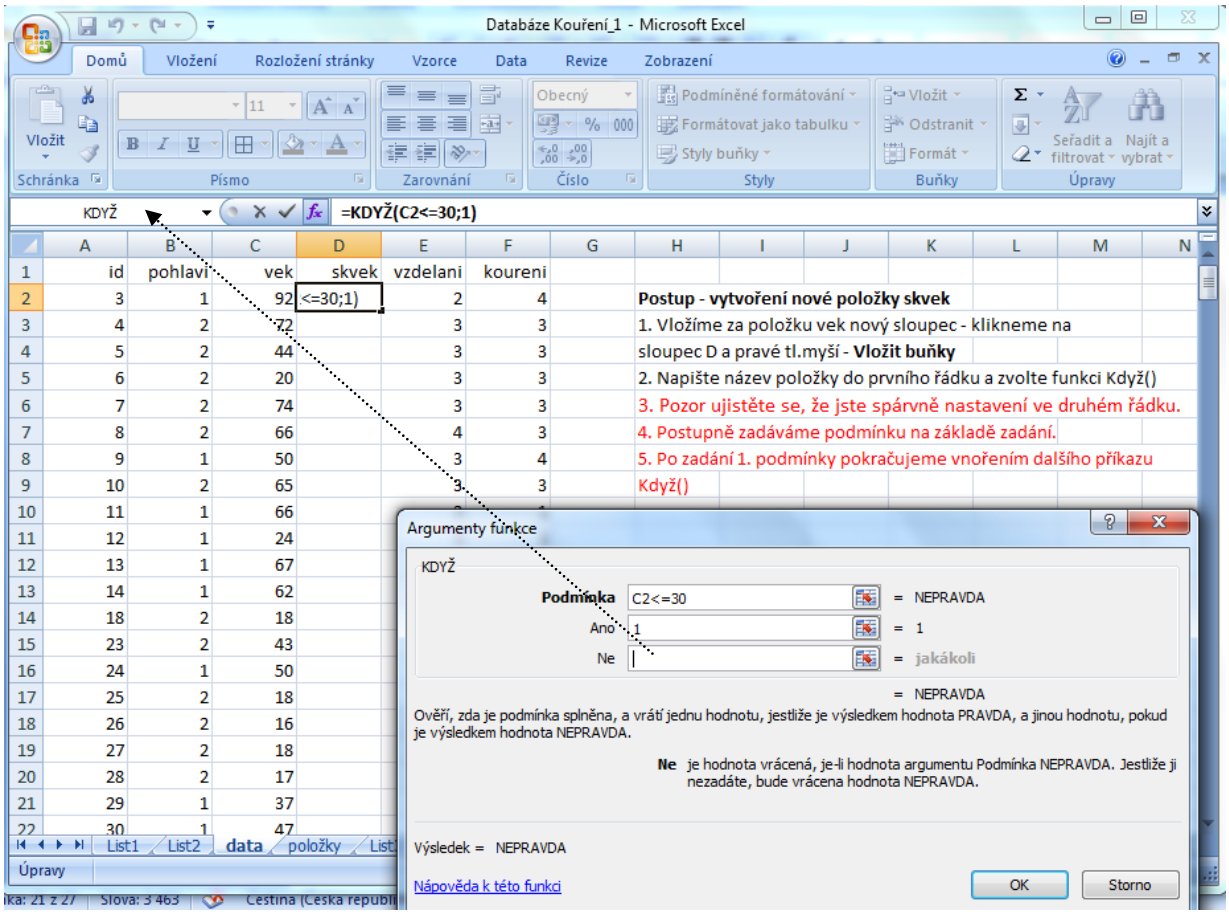
b) Vytvoření věkových skupin je popsáno na obr. 2.12 až 2.15. Novou položku si označíme např. skvek. Tuto novou položku musíme vytvořit v základní databázi (list Data). Pro vytvoření použijeme funkci Když(). Ve funkci Když() zadáme hlavní podmínky a v případě splnění podmínky se do buňky vloží „1“, označující první věkovou skupinu (obr. 2.14). Pokračujeme variantou, kdy podmínka není splněna. Vnoříme další funkci Když() a pokračujeme až do předposlední kategorie. U předposlední kategorie bude v případě nesplnění podmínky uvedena kategorie poslední (obr. 2.15). Ale to platí jen v případě, že se postupovalo od nejnižší kategorie po nejvyšší.



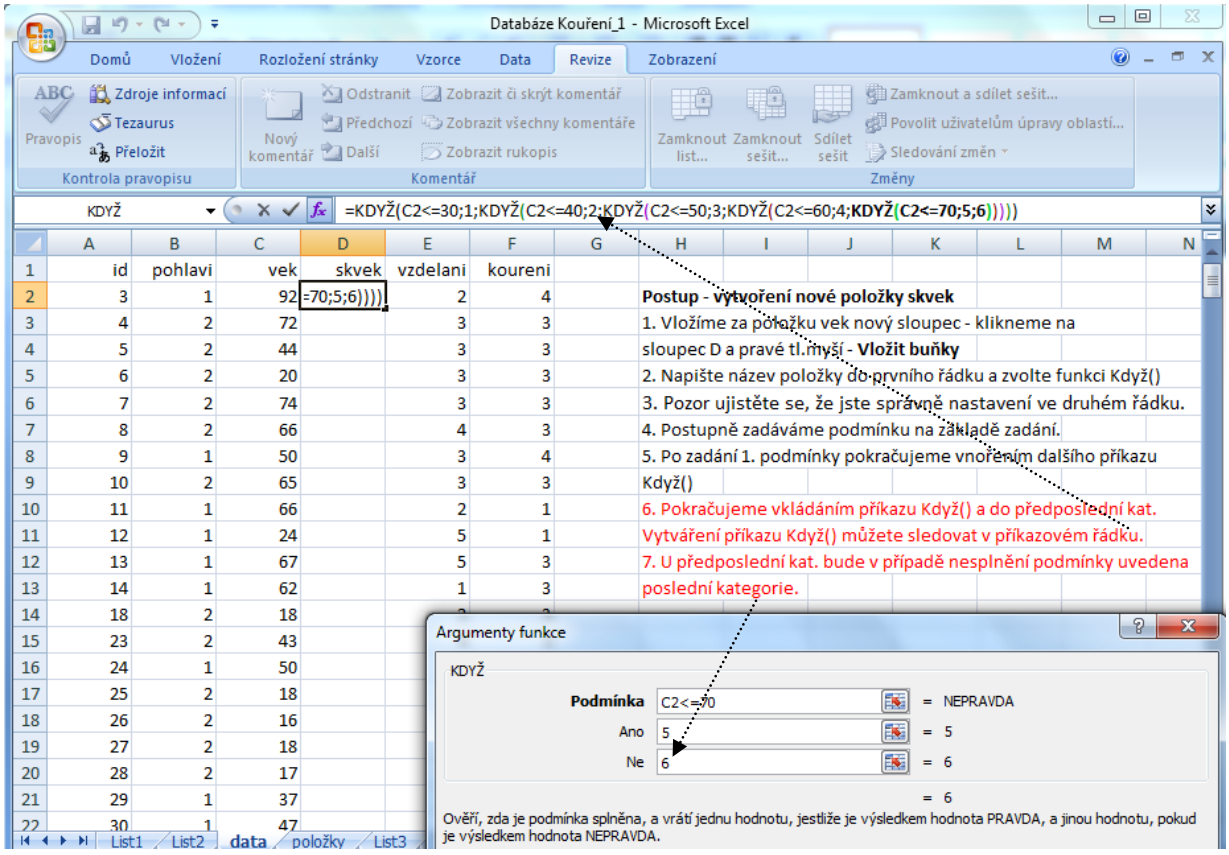
Obr. 2.12 Krok 1b) Vytvoření nové položky na listu data



Obr. 2.13 Krok 2b) Zadání funkce Když()

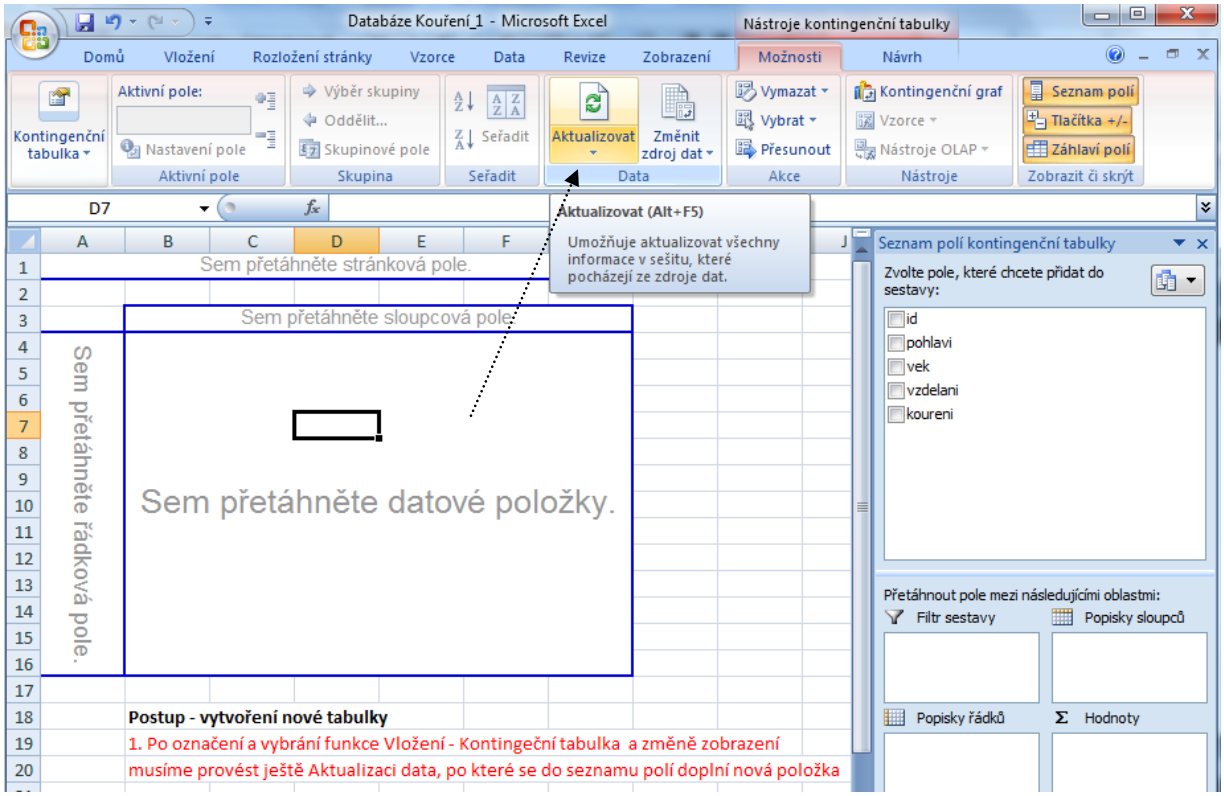


Obr. 2.14 Krok 3b)- 5b) Zadání funkce Když() – vnořeného

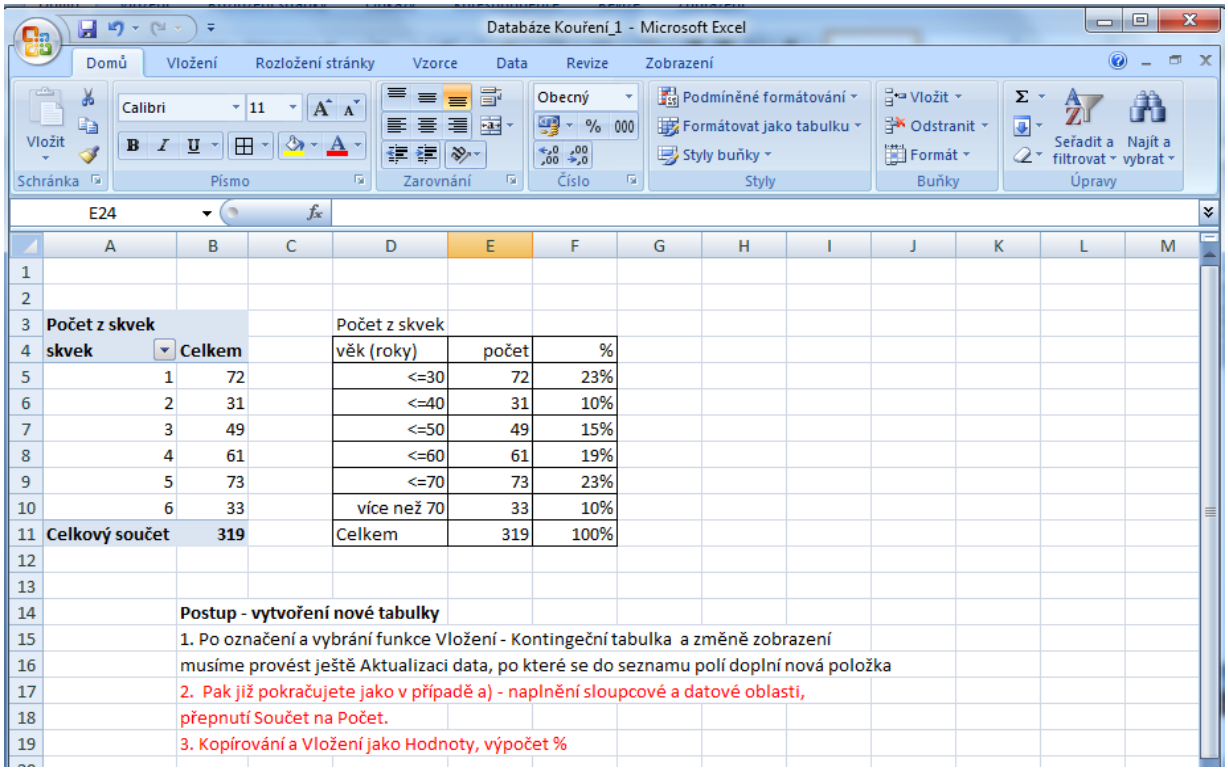


Obr. 2.15 Krok 6b)- 7b) Zadání funkce Když() u předposlední kategorie

c) Postup při vytvoření tabulky je obdobný jako v případě a), ale musí se provést aktualizace, která zahrne do zpracování i novou položku skvek (obr. 2.16 – 2.18).



Obr. 2.16 Krok 1c) Vytvoření tabulky - aktualizace dat



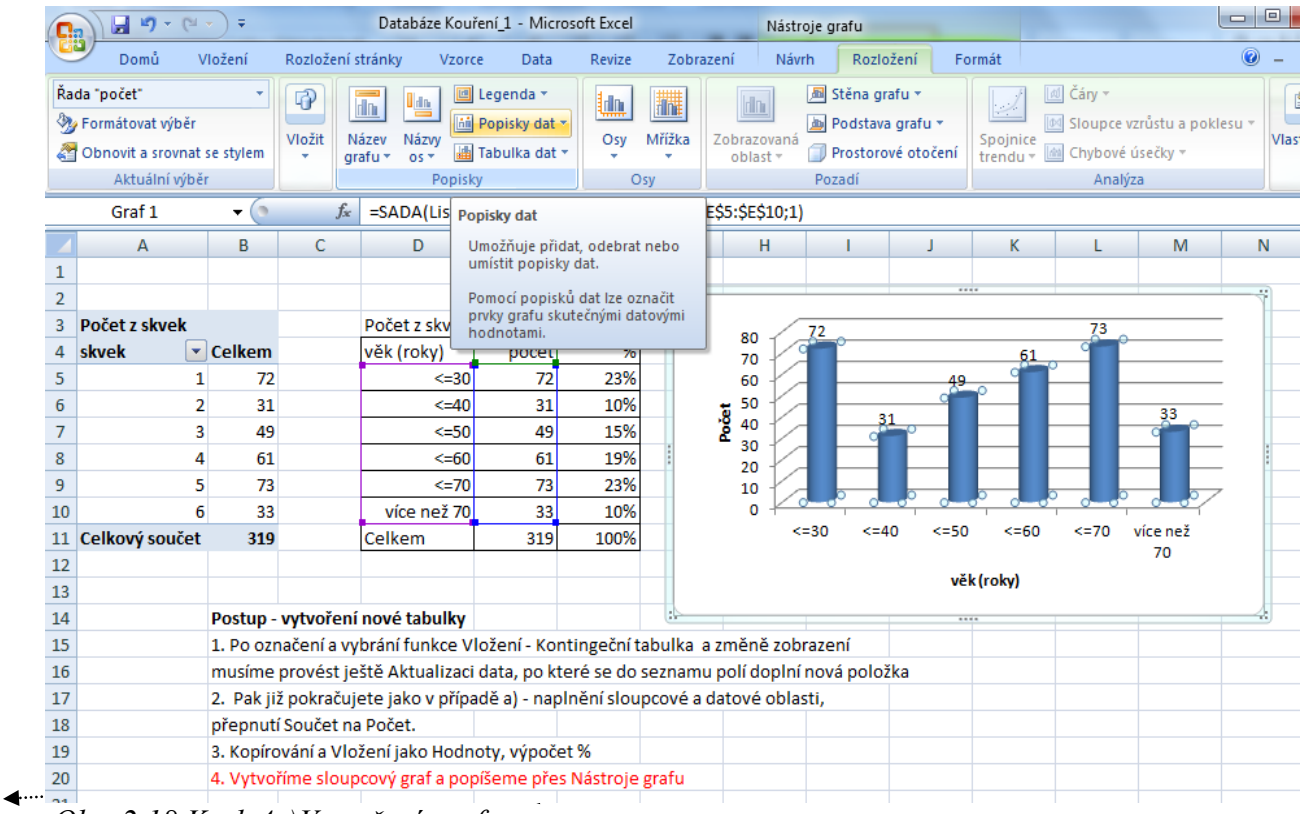
Počet z skvek	
skvek	Celkem
1	72
2	31
3	49
4	61
5	73
6	33
Celkový součet	319

Počet z skvek		
věk (roky)	počet	%
<=30	72	23%
<=40	31	10%
<=50	49	15%
<=60	61	19%
<=70	73	23%
více než 70	33	10%
Celkem	319	100%

Postup - vytvoření nové tabulky

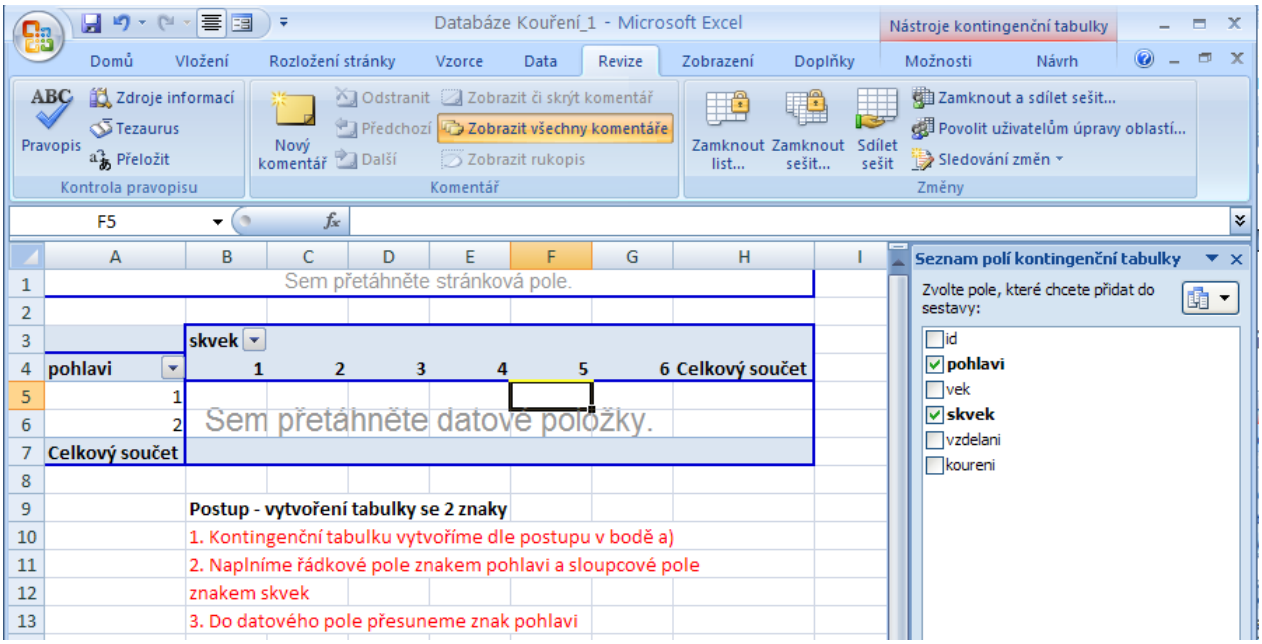
1. Po označení a vybrání funkce Vložení - Kontingeční tabulka a změně zobrazení musíme provést ještě Aktualizaci data, po které se do seznamu polí doplní nová položka
2. Pak již pokračujete jako v případě a) - naplnění sloupcové a datové oblasti, přepnutí Součet na Počet.
3. Kopírování a Vložení jako Hodnoty, výpočet %

Obr. 2.17 Krok 2c) – 3c) Výpočet procent a úprava tabulky

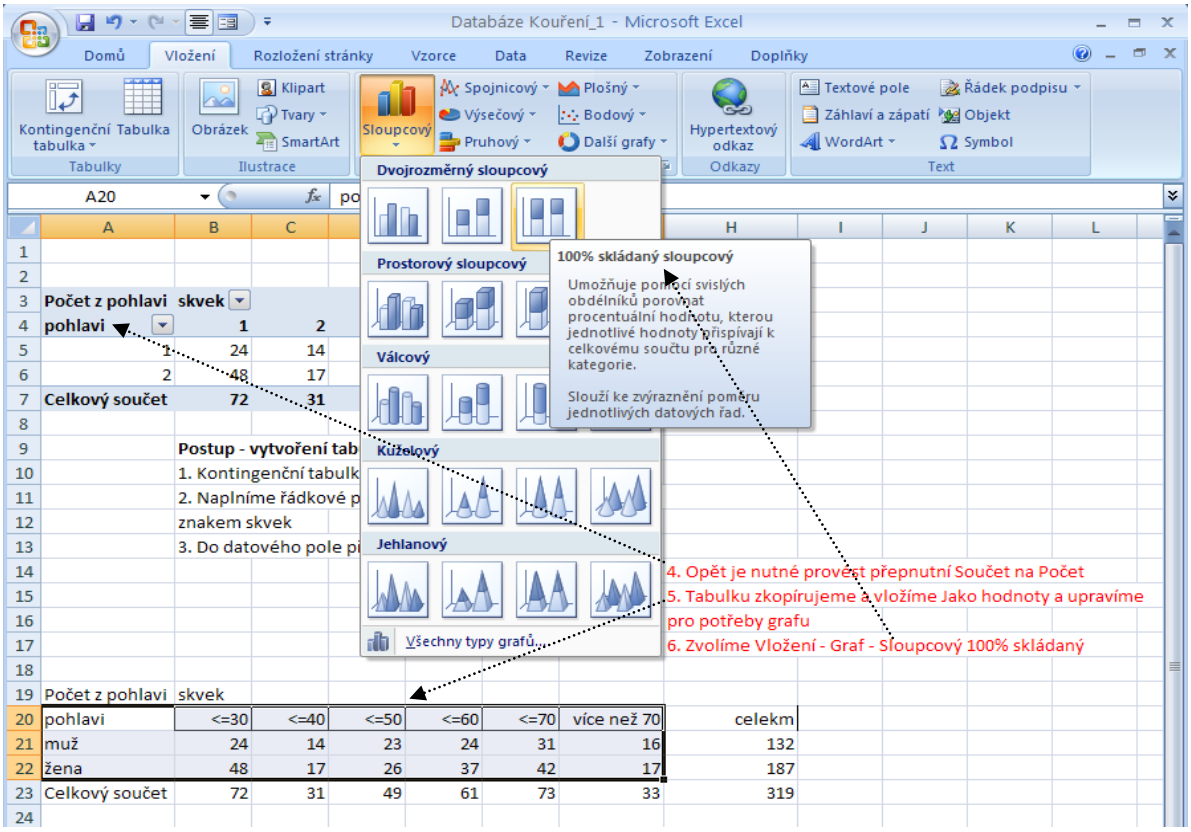


Obr. 2.18 Krok 4c) Vytvoření grafu

d) Při vytvoření tabulky s veličinami pohlaví a skvek budeme postupovat obdobně je v případě tabulky pro jednu veličinu, ale bude se naplňovat i pole řádkové (obr. 2.19). Při vytvoření grafu zvolíme graf, který bude vycházet ze skupin muži, ženy jako z celky vzhledem k tomu, že celkový počet osob ve skupinách není shodný. Pro srovnání musíme použít relativní počty (obr. 2.20).

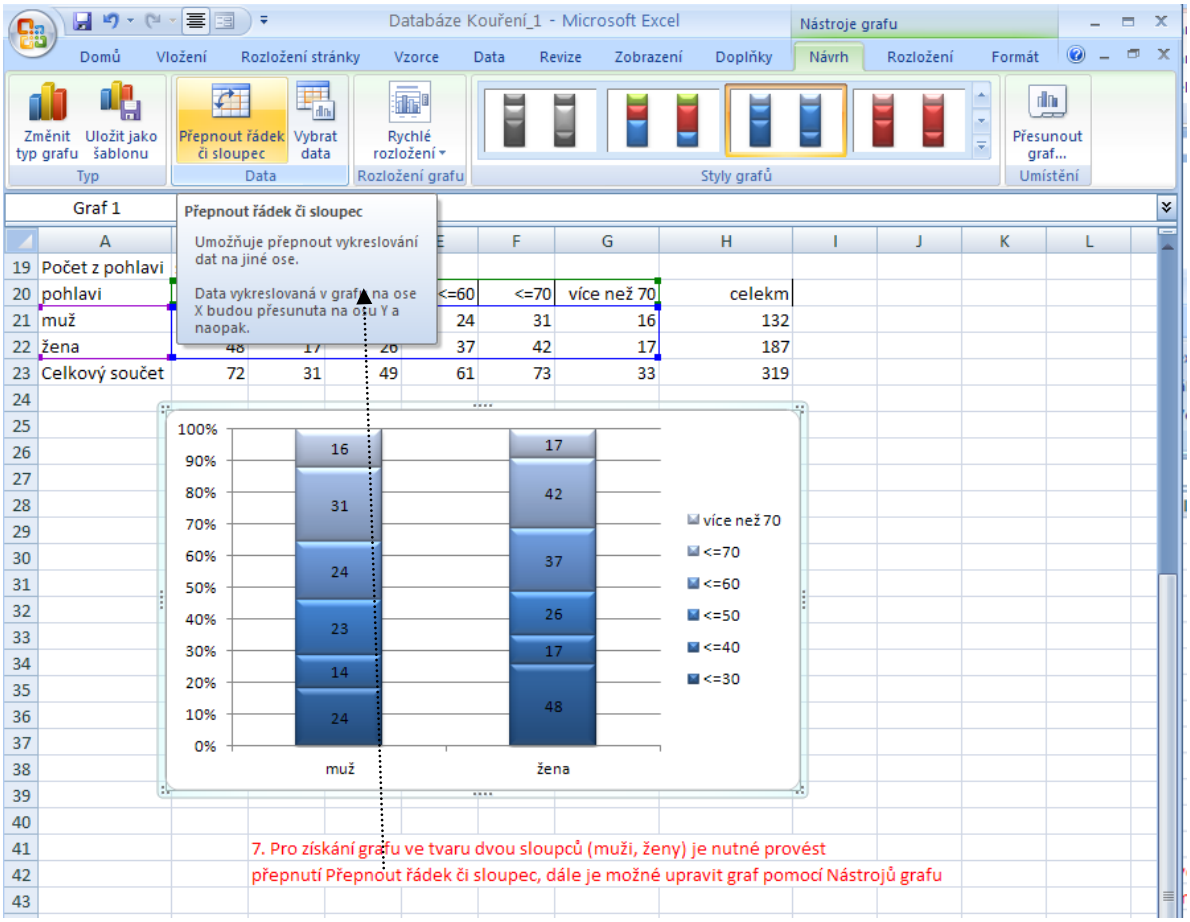


Obr. 2.19 Krok 1d) -3d) Vytvoření takulky – dvě veličiny



Počet z pohlavi	skvek						
pohlavi	<=30	<=40	<=50	<=60	<=70	více než 70	celekm
muž	24	14	23	24	31	16	132
žena	48	17	26	37	42	17	187
Celkový součet	72	31	49	61	73	33	319

Obr. 2.20 Krok 4d) – 6d) Příprava dat pro vytvoření grafu



Obr. 2.21 Krok 7d) Vytvoření grafu

Interpretace grafu: z grafu na obr. 2.21 je patrné, že věkové složení mužů a žen není zcela stejné. U žen je zastoupena věková skupina mladších respondentů více (26 %) než u mužů (18 %). Přesné podíly (%) věkových skupin je možné dopočítat. Analýza, zda věkové složení mužů a žen je významně rozdílné, bude náplní následující kapitoly.

Shrnutí obsahu kapitoly

Údaje (data) o objektech vytváří databázi. Databáze obsahuje veličiny kvalitativní a kvantitativní. Základní operací při zpracování dat je třídění podle kvalitativních veličin. Výsledkem třídění jsou frekvenční tabulky s absolutními i relativními počty. Výsledky třídění se mohou znázornit i graficky.
MS Excel: kontingenční tabulka, funkce KDYŽ(), výsečový a sloupcový graf

Kontrolní otázky:

1. Co je to databáze?
2. Jaké rozlišujeme druhy veličiny a na jakých stupnicích se měří?
3. Co znamená degradace veličiny kvantitativní na kvalitativní?
4. Co obsahují frekvenční tabulky?

Úkol

2.1 Analyzujte data v databázi z příkladu 2.1:

a) Sestavte jednoduché frekvenční tabulky z veličiny vzdělání a kouření, vypočtěte procentuální zastoupení. Sestavte grafy.
b) Vytvořte novou veličinu z veličiny vzdělání sloučením kategorií 3+4 – nová kategorie SŠ s maturitou, 5+6 – VOŠ + VŠ.
c) Vytvořte novou veličinu z veličiny kouření sloučením kategorií 1+2 – nová kategorie - kuřák, 3+4 – nekuřák.
d) Vyjádřete výskyt kuřáků a nekuřáku v závislosti na vzdělání (sloučený znak), znázorněte i graficky.

3 Popisná statistika

V této kapitole se dozvíte:

- Co popisují a jaké používáme charakteristiky střední polohy.
- Co popisují a jaké používáme charakteristiky variability.
- Co je to histogram.

Po jejím prostudování byste měli být schopni:

- Určit a vypočítat charakteristiky střední polohy podle typu znaku.
- Vypočítat charakteristiky variability u metrických znaků.
- Vysvětlit co vyjadřuje směrodatná odchylka.
- Sestavit histogram.

Klíčová slova této kapitoly: modus, medián, aritmetický průměr, kvantil, rozptyl, směrodatná odchylka, histogram

Doba potřebná ke studiu a zpracování úkolů: **2 - 3 hodin**

Průvodce studiem

Tato kapitola předpokládá již zvládnutí základních funkcí uvedených v předchozích kapitolách. Student se seznámí s použitím vybraných statistických funkcí a možností analýzy dat. Náročnost této kapitoly v případě zvládnutí postupů v kapitole 1 – 2 je mírná.
Teorie k uvedeným tématům je probírána v první a třetí přednášce Lékařská biofyzika, výpočetní technika I – Biostatistika.

Úkolem popisné statistiky je provést popis souboru pozorovaných objektů jako celku např. pokud budeme mít informace o tělesné výšce u 100 osob a budeme chtít popsat tento soubor jako celek. K tomuto popisu použijeme ***charakteristiky polohy*** **a** ***variability (rozptýlení)***. Charakteristiky polohy informují, kde se nachází přibližný střed tohoto souboru např. v souboru mužů se střed u tělesné výšky nachází kolem hodnoty 181 cm. Charakteristika variability naopak informuje, jak jsou data rozptýlená např. výška se ve sledovaném souboru pohybuje od 160 do 198 cm.
V následujícím textu je popsán výpočet charakteristik polohy a variability podle typu znaků. Ke znázornění rozložení hodnot v souboru se používají také grafy – histogramy, krabicové grafy.

3.1 Charakteristiky polohy

Modus $\hat{x}$ – jedná se o kategorii nebo hodnotu, do které je zařazen největší počet objektů z pozorovaného souboru.
Určuje se u veličin kvalitativních (nominálních, ordinálních) i kvantitativních.
Medián $\tilde{x}$ – je kategorie nebo hodnota, která rozděluje setřídění soubor na dvě stejně četné části (poloviny).
Určuje se u veličin kvalitativních (ordinálních) i kvantitativních.

Kvantily – výpočet kvantilů vychází z kumulativní relativní četnosti. 100*P*% kvantil je taková kategorie, kdy 100*P* % hodnot patří do kategorie nebo hodnoty nižší nebo rovné tomuto kvantilu. Pro hodnotu *P* platí: $0 \leq P \leq 1$ např. $P = 0{,}25$, pak hovoříme o 25% kvantilu, pokud $P = 0{,}5$, jedná se o medián.
Určuje se u veličin kvalitativních (ordinálních) i kvantitativních.
Aritmetický průměr $\bar{x}$ - součet všech hodnota souboru dělený celkovým počtem pozorování.

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, kde *n* je počet pozorování, x_i jsou jednotlivá pozorování.

Určuje se u veličin kvantitativních.

3.2 Charakteristiky variability

Charakteristiky variability budou uvedeny jen pro znaky kvantitativní (metrické).
Rozpětí *R* – se vypočte jako rozdíl mezi nejvyšší a nejnižší hodnotou souboru.
Rozpětí má nevýhodu v případě, že v souboru se nacházejí odlehlé hodnoty a ostatní pozorování jsou blízko u sebe, pak rozpětí poskytuje zavádějící informaci o variabilitě dat.
Rozptyl s^2 – jedná se o charakteristiku, která je založena na rozdílech jednotlivých pozorování od aritmetického průměru. Pokud jsou pozorování soustředěna kolem svého průměru, je jejich variabilita malá. Pokud jsou naopak roztroušena ve značné vzdálenosti od průměru, pak je hodnota s^2 velká a jejich variabilitu označíme také jako velkou.

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$, kde *n* je počet pozorování

Směrodatná odchylka *s* je odmocnina z rozptylu. Používá se častěji než rozptyl, protože je ve stejné metrické rovině jako veličina *x*.

$s = \sqrt{s^2}$

Směrodatná odchylka je ve stejných jednotkách jako původní hodnoty.
Střední chyba průměru s_e – velikost střední chyby průměru záleží na velikosti rozptylu a na počtu pozorování.

$s_e = \sqrt{\frac{s^2}{n}} = \frac{s}{\sqrt{n}}$

Čím větší rozptyl dat, s tím větší chybou musíme počítat a naopak čím větší počet pozorování při stejném rozptylu, tím menší chyba.
Variační koeficient *v* – vyjadřuje relativní rozptýlení dat. Počítá se jako podíl směrodatné odchylky k průměru v procentech.

$v = \frac{s}{\bar{x}} 100\ \%$

Variační koeficient se používá často i při statistické kontrole laboratorních testů.

Příklad 3.1

Pomocí základních charakteristik střední polohy a variability popište data v databázi S3_Databáze (Preventivní pohlídky – ženy), jedná se o údaje u studentek. Tato databáze obsahuje část informací z dotazníkového šetření, seznam položek dotazníků (veličin) je uveden na obr. 3.1.

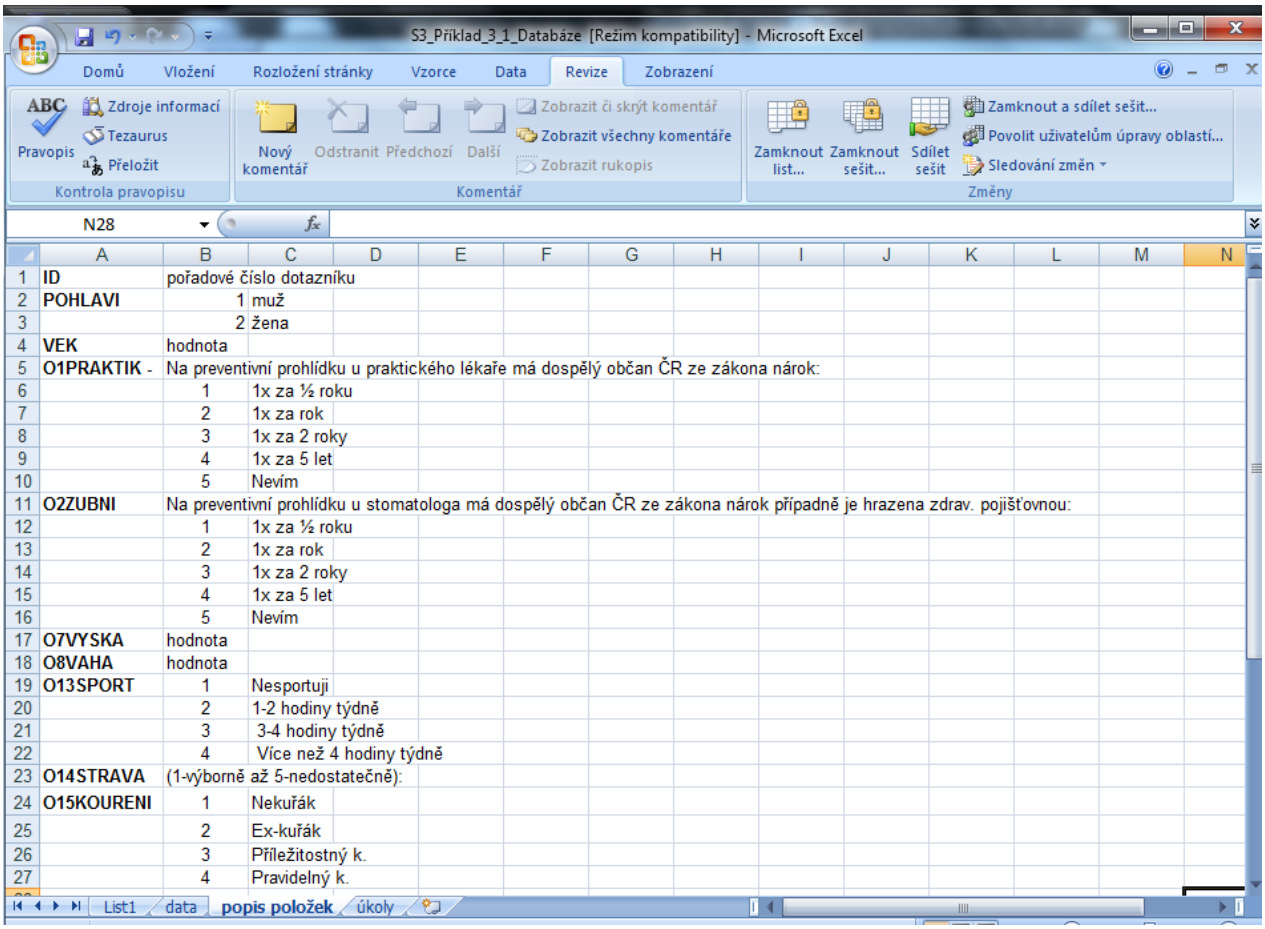
Úkoly:

a) Určete modus a medián u veličiny O13_sportovní aktivita.

b) U položky Výška určete charakteristiky střední polohy i variability, sestavte histogram

c) U položky Výška zjistěte percentily (25%, 50%,75%)

d) Můžete vypočítat aritmetický průměr z veličiny O15_Kouření?

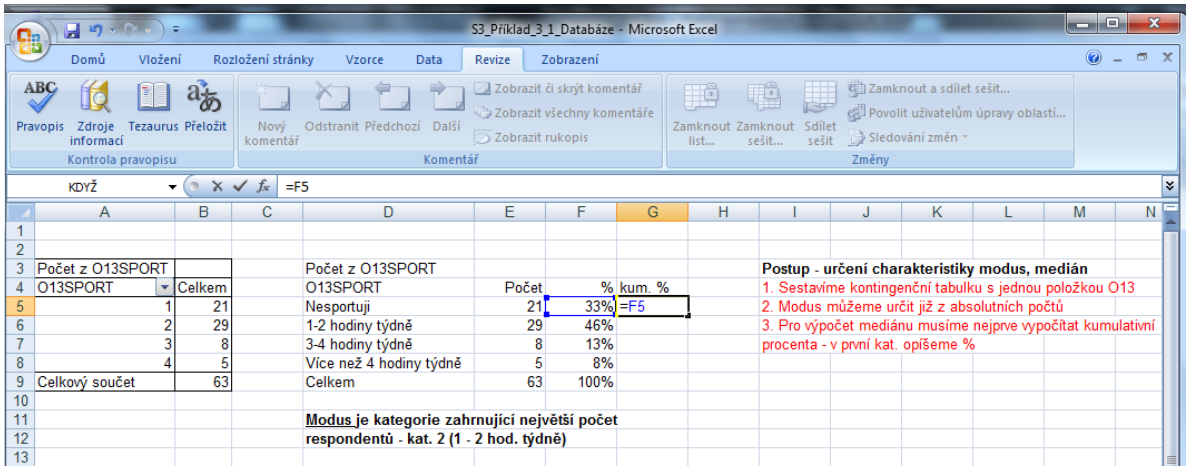


	A	B	C
1	ID	pořadové číslo dotazníku	
2	POHLAVI	1	muž
3		2	žena
4	VEK	hodnota	
5	O1PRAKTIK -	Na preventivní prohlídku u praktického lékaře má dospělý občan ČR ze zákona nárok:	
6		1	1x za ½ roku
7		2	1x za rok
8		3	1x za 2 roky
9		4	1x za 5 let
10		5	Nevím
11	O2ZUBNI	Na preventivní prohlídku u stomatologa má dospělý občan ČR ze zákona nárok případně je hrazena zdrav. pojišťovnou:	
12		1	1x za ½ roku
13		2	1x za rok
14		3	1x za 2 roky
15		4	1x za 5 let
16		5	Nevím
17	O7VYSKA	hodnota	
18	O8VAHA	hodnota	
19	O13SPORT	1	Nesportuji
20		2	1-2 hodiny týdně
21		3	3-4 hodiny týdně
22		4	Více než 4 hodiny týdně
23	O14STRAVA	(1-výborně až 5-nedostatečně):	
24	O15KOURENI	1	Nekuřák
25		2	Ex-kuřák
26		3	Příležitostný k.
27		4	Pravidelný k.

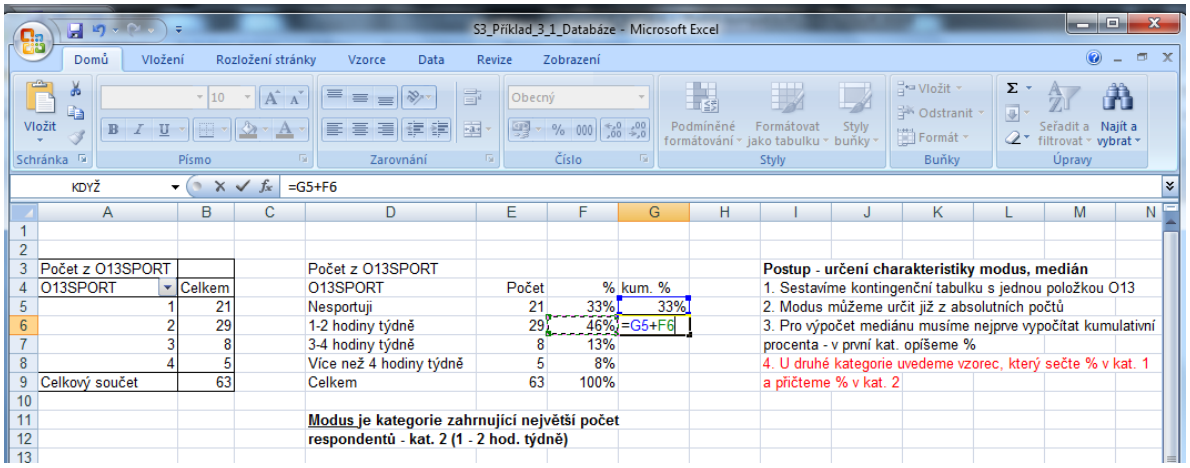
Obr. 3.1 Seznam veličin v datamázi

Řešení

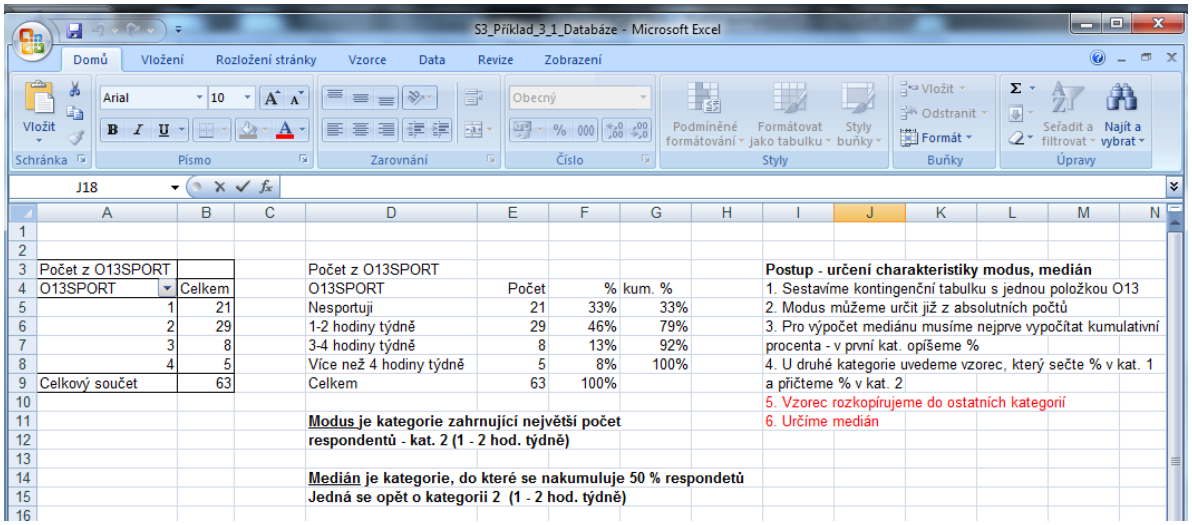
a) Pro výpočet uvedených charakteristik si sestavíme nejprve kontingenční tabulku a vypočteme % (obr. 3.2)



Obr. 3.2 Krok 1a) Označení charakteristiky modus

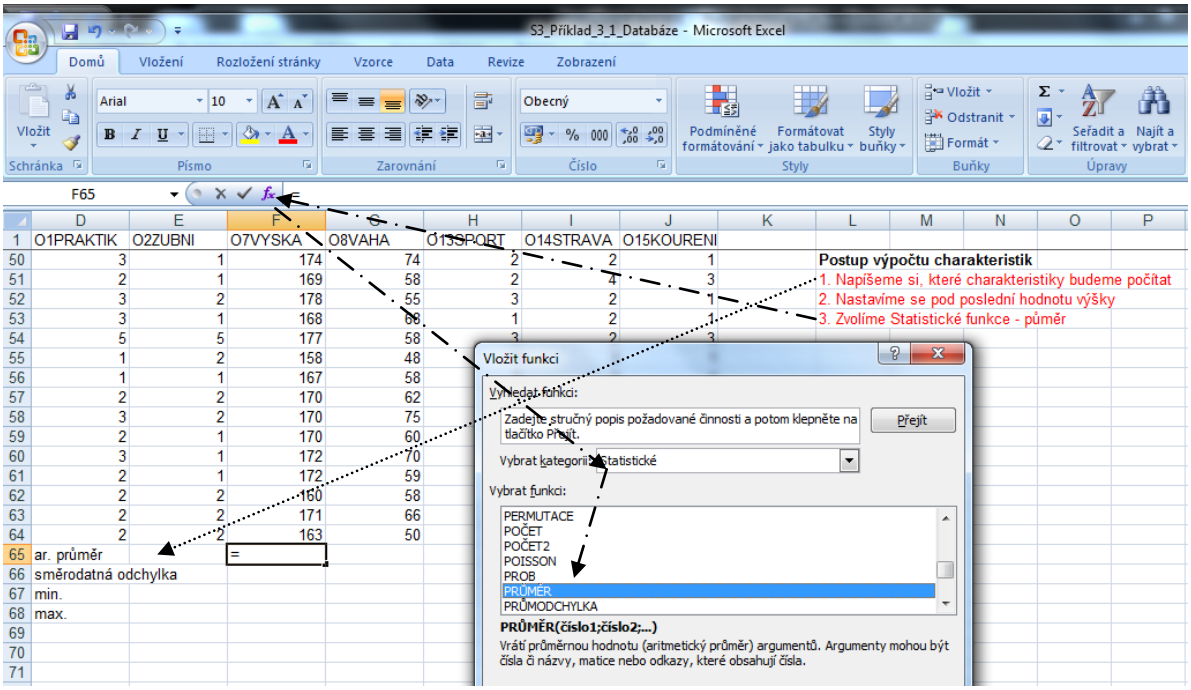


Obr. 3.3 Krok 2a) Výpočet kumulativních procent

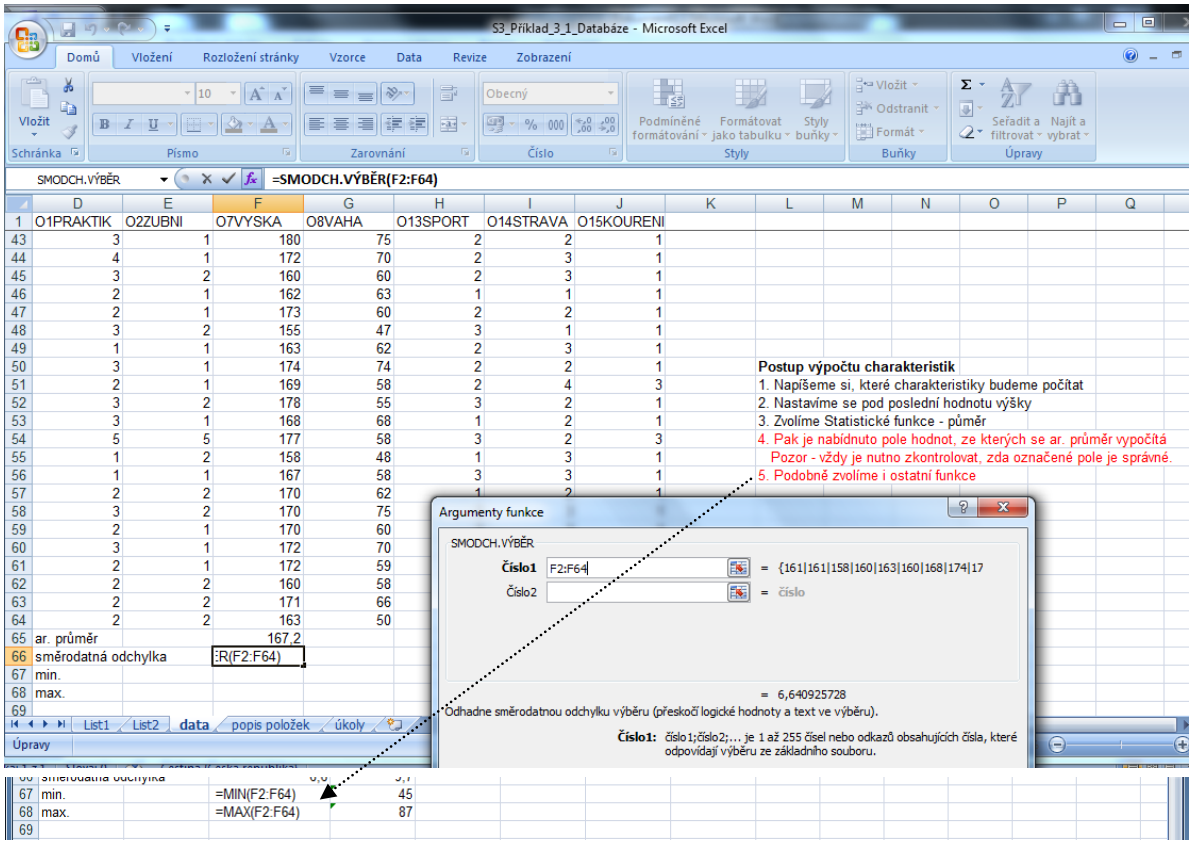


Obr. 3.4 Krok 3a) Označení charakteristiky medián

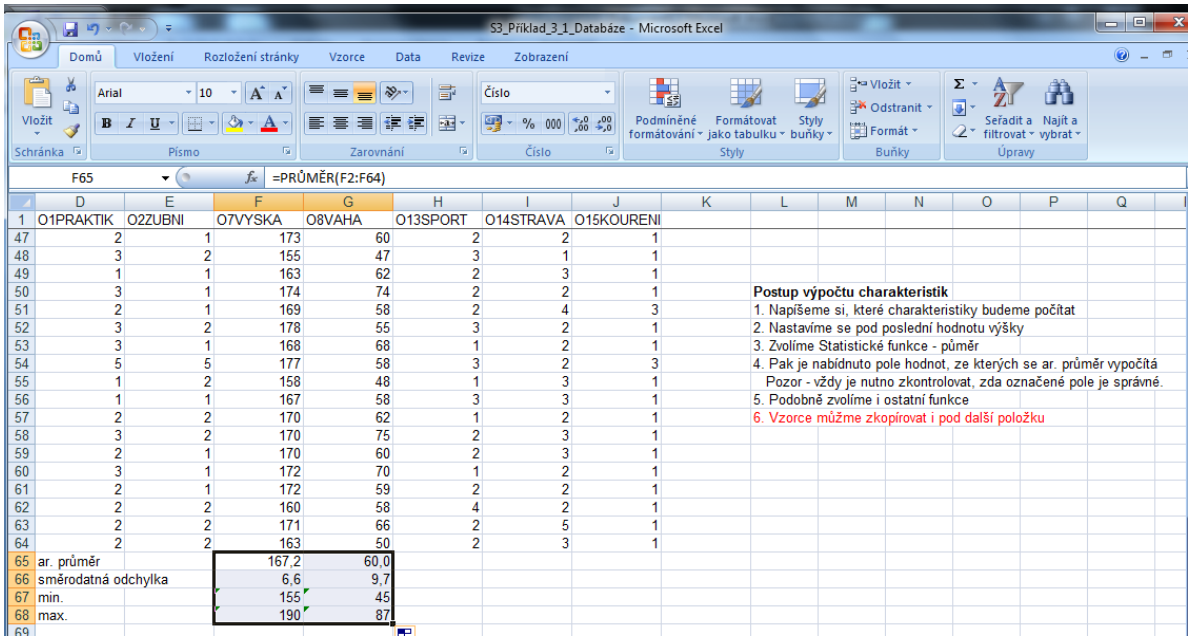
b) Pro výpočet základních charakteristik u znaku Výška použijeme nejprve jednotlivé statistické funkce. Výpočty provedeme na listu Data.



c) *Obr. 3.5 Krok 1-3b) Zvolení Statistické funkce - Průměr()*



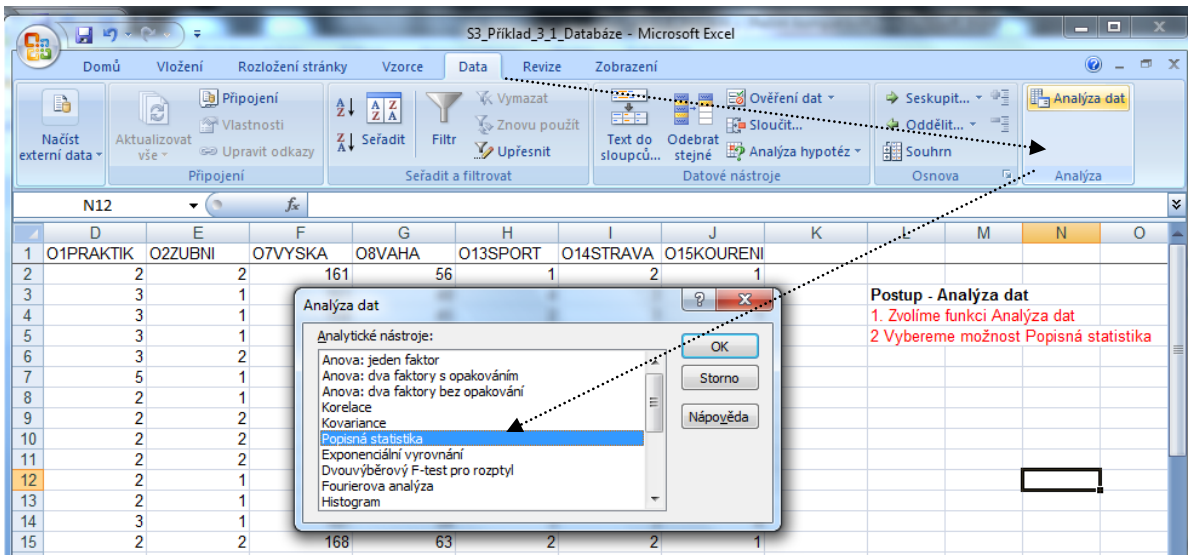
Obr. 3.6 Krok 4 - 5b) Zvolení Statistické funkce – Směrodatná odchylka výběrová(), min(), max()



	D	E	F	G	H	I	J
1	O1PRAKTIK	O2ZUBNI	O7VYSKA	O8VAHA	O13SPORT	O14STRAVA	O15KOURENI
47	2	1	173	60	2	2	1
48	3	2	155	47	3	1	1
49	1	1	163	62	2	3	1
50	3	1	174	74	2	2	1
51	2	1	169	58	2	4	3
52	3	2	178	55	3	2	1
53	3	1	168	68	1	2	1
54	5	5	177	58	3	2	3
55	1	2	158	48	1	3	1
56	1	1	167	58	3	3	1
57	2	2	170	62	1	2	1
58	3	2	170	75	2	3	1
59	2	1	170	60	2	3	1
60	3	1	172	70	1	2	1
61	2	1	172	59	2	2	1
62	2	2	160	58	4	2	1
63	2	2	171	66	2	5	1
64	2	2	163	50	2	3	1
65	ar. průměr		167,2	60,0			
66	směrodatná odchylka		6,6	9,7			
67	min.		155	45			
68	max.		190	87			
69							

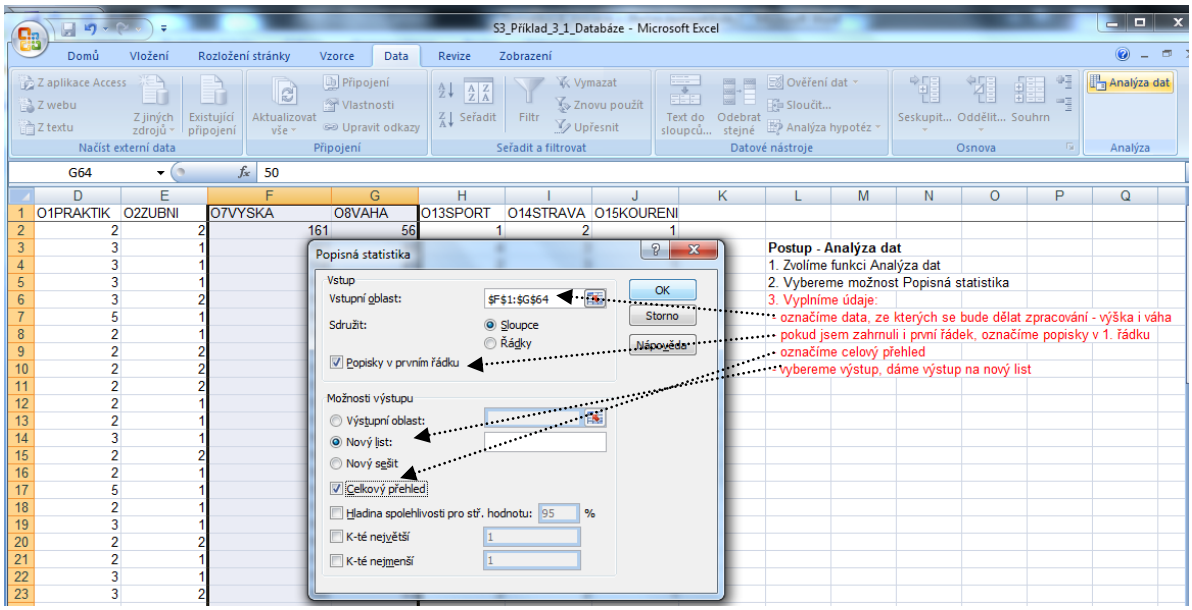
Obr. 3.7 Krok 6b) Zkopírování vzorců pod další položku Váha

Uvedené charakteristiky se dají vypočítat také přes Analýzu dat, která je na liště Data. Pokud tuto funkci nemáte nainstalovanou, v Příloze 1 je uveden podrobný popis instalace. Při spuštění funkce Analýza dat budeme nastaveni na list Data.

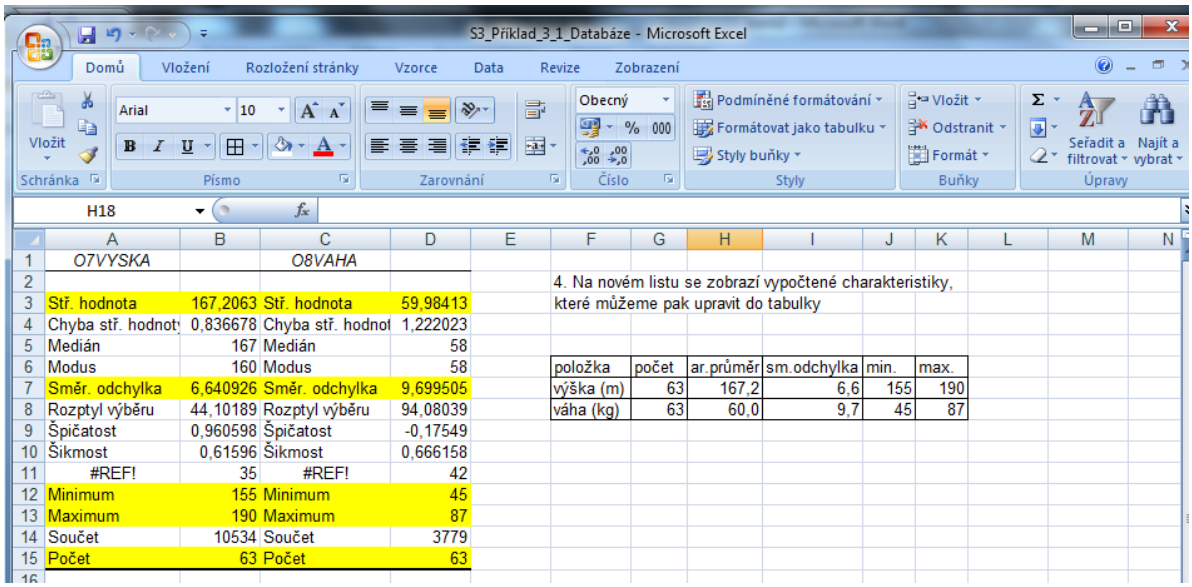


	D	E	F	G	H	I	J
1	O1PRAKTIK	O2ZUBNI	O7VYSKA	O8VAHA	O13SPORT	O14STRAVA	O15KOURENI
2	2	2	161	56	1	2	1
3	3	1					
4	3	1					
5	3	1					
6	3	2					
7	5	1					
8	2	1					
9	2	2					
10	2	2					
11	2	2					
12	2	1					
13	2	1					
14	3	1					
15	2	2	168	63	2	2	1

Obr. 3.8 Krok 1-2b) Výběr funkce – Popisná statistika



Obr. 3.9 Krok 3b) Popisná statistika – zadání vstupních parametrů

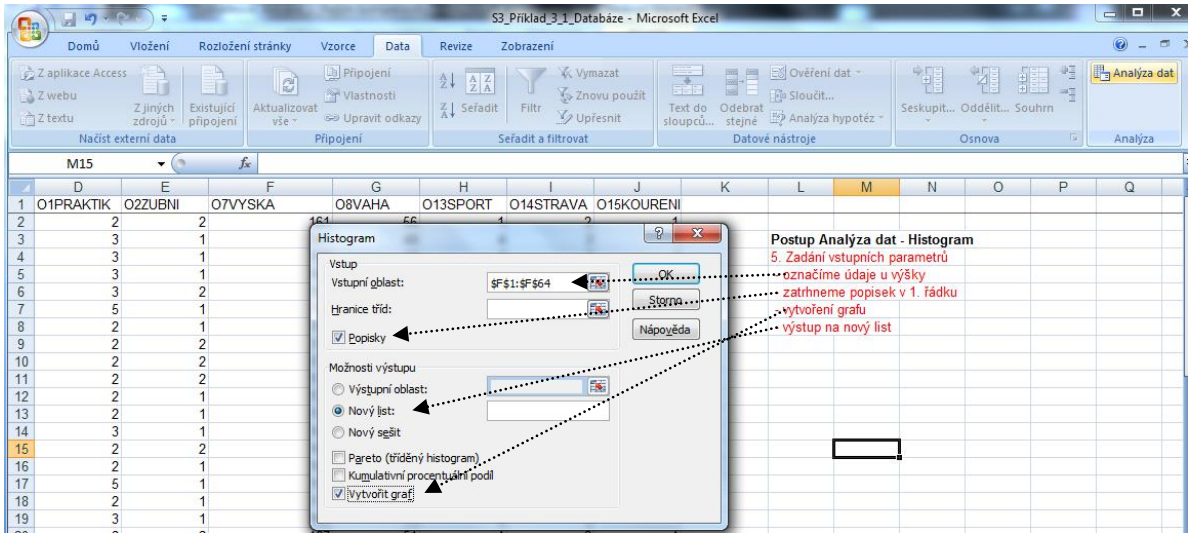


	O7VYSKA		O8VAHA	
3	Stř. hodnota	167,2063	Stř. hodnota	59,98413
4	Chyba stř. hodnoty	0,836678	Chyba stř. hodnoty	1,222023
5	Medián	167	Medián	58
6	Modus	160	Modus	58
7	Směr. odchylka	6,640926	Směr. odchylka	9,699505
8	Rozptyl výběru	44,10189	Rozptyl výběru	94,08039
9	Špičatost	0,960598	Špičatost	-0,17549
10	Šikmost	0,61596	Šikmost	0,666158
11	#REF!	35	#REF!	42
12	Minimum	155	Minimum	45
13	Maximum	190	Maximum	87
14	Součet	10534	Součet	3779
15	Počet	63	Počet	63

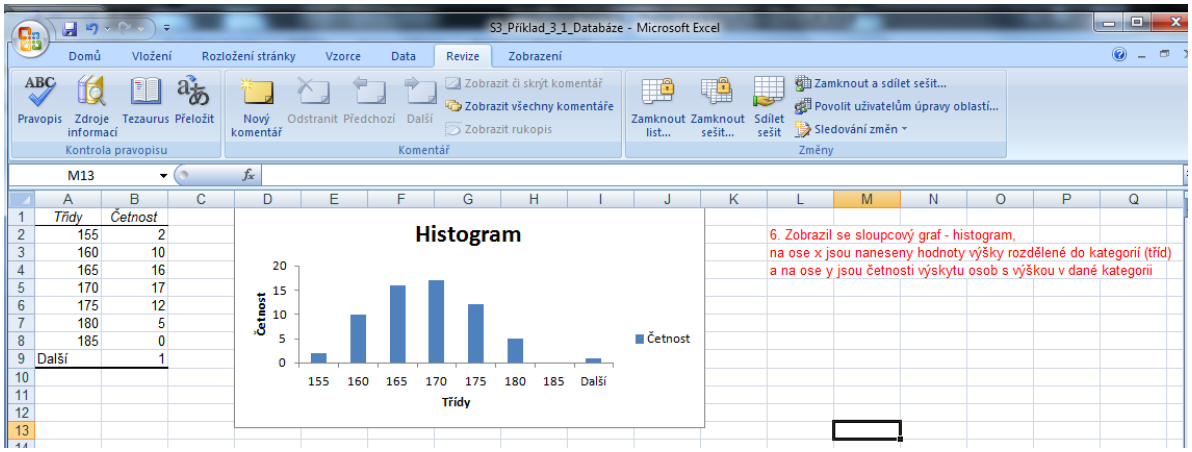
položka	počet	ar.průměr	sm.odchylka	min.	max.
výška (m)	63	167,2	6,6	155	190
váha (kg)	63	60,0	9,7	45	87

Obr. 3.10 Krok 4b) Popisná statistika – zobrazení vypočtených charakteristik

Histogram vytvoříme obdobně přes funkci Analýza dat a také vyplnění vstupních parametrů je obdobné jako u funkce Popisná statistika. Nastavíme se opět na list Data a z Analýzy dat zvolíme Histogram (obr. 3.11 - 3.12).

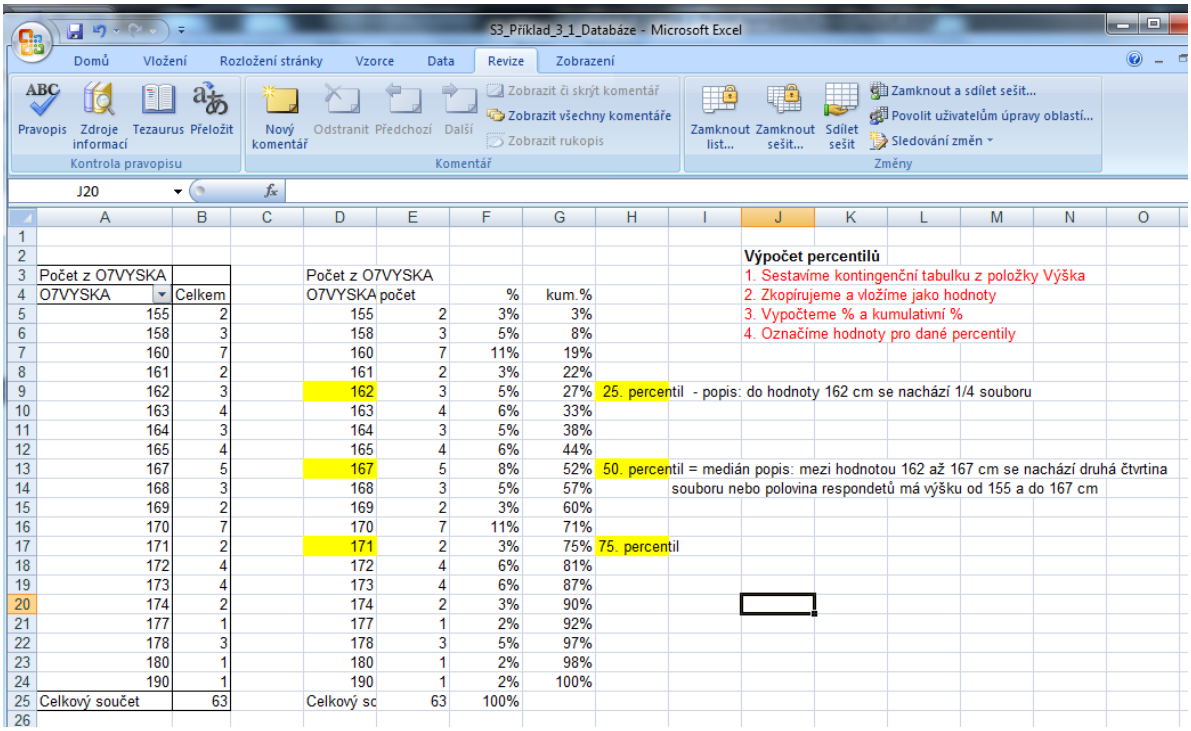


Obr. 3.11 Krok 5b) Histogram – zadání vstupních parametrů



Obr. 3.12 Krok 6b) Histogram – zobrazení výsledku
(pozn. vzhledem k tomu, že výška je spojitá veličina, neměly by být mezery mezi sloupci v histogramu)

c) Pro výpočet percentilů z položky výška, zvolíme výpočet pomocí kumulativních procent, obdobně jako v úkolu a). Sestavíme kontingenční tabulku a vypočteme kumulativní procenta (obr. 3.13 – 3.14).



Počet z O7VYSKA	
O7VYSKA	Celkem
155	2
158	3
160	7
161	2
162	3
163	4
164	3
165	4
167	5
168	3
169	2
170	7
171	2
172	4
173	4
174	2
177	1
178	3
180	1
190	1
Celkový součet	63

Počet z O7VYSKA				
O7VYSKA	počet	%	kum.%	
155	2	3%	3%	
158	3	5%	8%	
160	7	11%	19%	
161	2	3%	22%	
162	3	5%	27%	
163	4	6%	33%	
164	3	5%	38%	
165	4	6%	44%	
167	5	8%	52%	
168	3	5%	57%	
169	2	3%	60%	
170	7	11%	71%	
171	2	3%	75%	
172	4	6%	81%	
173	4	6%	87%	
174	2	3%	90%	
177	1	2%	92%	
178	3	5%	97%	
180	1	2%	98%	
190	1	2%	100%	
Celkový sc	63	100%		

Obr. 3.13 Krok 1-4c) Výpočet percentilů

d) U položky O15_Kouření aritmetický průměr nelze vypočítat, jedná se o kvalitativní veličinu.

Shrnutí obsahu kapitoly

Popisné statistiky popisují soubor pozorovaných objektů jako celek. K tomuto popisu se slouží ***charakteristiky polohy*** **a** ***variability (rozptýlení)***. Ke znázornění rozložení hodnot v souboru se používají grafy – histogramy.

MS Excel – statistické funkce Počet(), Průměr(), Smodch.výběr(), Min(), Max(), v Analýze dat – Popisná statistika, Histogram

Kontrolní otázky:

1. Jaké určujeme charakteristiky polohy u znaků kvalitativních?
2. Jaké určujeme charakteristiky střední polohy a variability u znaků kvantitativních?
3. Co vyjadřuje směrodatná odchylka a proč se užívá častěji než rozptyl?
4. Co popisuje histogram?

Úkol

3.1 Pokračujte v analýze položek v databázi S3_Databáze.

a) Vypočtěte BMI (Body Mass Index – kg/m^2) a rozdělte do skupin do 25, do 30, 30 a více (nová položka skBMI)
b) Určete druhy znaků
c) Podle druhu znaků určete charakteristiky střední polohy a variability, sestavte histogram
d) U znaku BMI si ověřte výpočet směrodatné odchylky

4 Statistické testy pro kvalitativní data

V této kapitole se dozvíte:

- Jaký je princip χ^2 testu pro dva a více výběrů a veličin), jak zní nulová hypotéza.
- Jak lze χ^2 test provést pomocí funkce MS Excelu
- Jak lze χ^2 test provést pomocí programu OpenEpi.
- Jak se interpretuje výsledek.

Po jejím prostudování byste měli být schopni:

- Určit, kdy použijete pro tetování nulové hypotézy χ^2 test.
- Provést výpočet χ^2 testu pomocí funkcí MS Excelu a programu Open Epi.
- Interpretovat výsledek statistického testu.

Klíčová slova této kapitoly: statistický test, nulová hypotéza, χ^2 test, testové kritérium, p-hodnota

Doba potřebná ke studiu a zpracování úkolů: 5-6 hodin

Průvodce studiem

Tato kapitola navazuje na obecné principy testování statistických hypotéz (přednáška 4). Bez pochopení základního principu testování statistických hypotéz bude mít student velké problémy s probíraným tématem.
Pro pochopení podstaty χ^2 testu pro dva a více výběrů (veličin) bude nejprve probrán manuální výpočet testového kritéria a pak budou studenti seznámeni s výpočtem tohoto testu pomocí programu OpenEpi.
Náročnost této kapitoly je vysoká.
Teorie k uvedeným tématům je probírána ve čtvrté přednášce Lékařská biofyzika, výpočetní technika I – Biostatistika.
Na cvičení nejsou probírány neparametrické testy pro jeden výběr a pro párová data.

4.1 Test homogenity pro kvalitativní znaky

Pearsonovo testové kritérium χ^2 se používá v případě, kdy má být ověřena hypotéza, že se sledovaná kvalitativní veličina se v několika různých situacích řídí stejným pravděpodobnostním rozdělením.

Toto rozdělení není nulovou hypotézou jednoznačně specifikováno; hypotéza H_o tvrdí, že jde o jedno rozdělení. Výběry ve všech uvažovaných situacích jsou vzorky z tohoto rozdělení.

Data tvoří četnostní tabulky (tab. 4.1) o *r* řádcích a *s* sloupcích, kde *r* je počet srovnávaných situací (výběrů) a *s* počet úrovní sledované veličiny *X*. Údaje v posledním sloupci vpravo jsou rozsahy výběrů.

Tab. 4.1 Četnostní tabulka

Výběr	Kategorie znaku X				Celkem
	x_1	x_2	.....	x_s	
1	n_{11}	n_{12}	.....	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	.....	n_{2s}	$n_{2.}$
.	.			.	.
.	.			.	.
.	.			.	.
.	.			.	.
r	n_{r1}	n_{r2}	.....	n_{rs}	$n_{r.}$
Celkem	$n_{.1}$	$n_{.2}$	.....	$n_{.1}$	$n_{..}$

Četnost n_{ij} vyjadřuje četnost kategorie j ve výběru i, kde i = 1,2,….,r a j = 1,2, ……s.

Nulová hypotéza H_o předpokládá pro všech r výběrů stejné pravděpodobnostní rozdělení p_1, p_2, ..., p_s odhadované jako:

$$p_j = \frac{n_{.j}}{n_{..}}, j = 1, 2,…., s$$

Platnost H_0 se prověřuje testovým kritériem:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s}\frac{\left(n_{ij} - e_{ij}\right)^2}{e_{ij}} \quad , \quad i = 1, 2,…., r; \quad j = 1, 2,…., s$$

kde

$e_{ij} = p_j n_i$ jsou očekávané počty.

Testové kritérium má za platnosti H_0 pro dostatečně velké výběry přibližně pravděpodobnostní rozdělení χ^2 s (r-1)(s-1) stupni volnosti. Podmínka použití χ^2 testu je, že očekávané počty e_{ij} >5. Kritický obor testu tvoří relativně vysoké hodnoty testového kritéria. Pokud vypočtená hodnota χ^2 je vyšší než kritická hodnota, H_0 se zamítá. Kritickou hodnotu nalezneme v závislosti na počtu stupňů volnosti a zvolené hladině významnosti (α=0,05) např. pomocí funkce v Excelu – CHIINV().

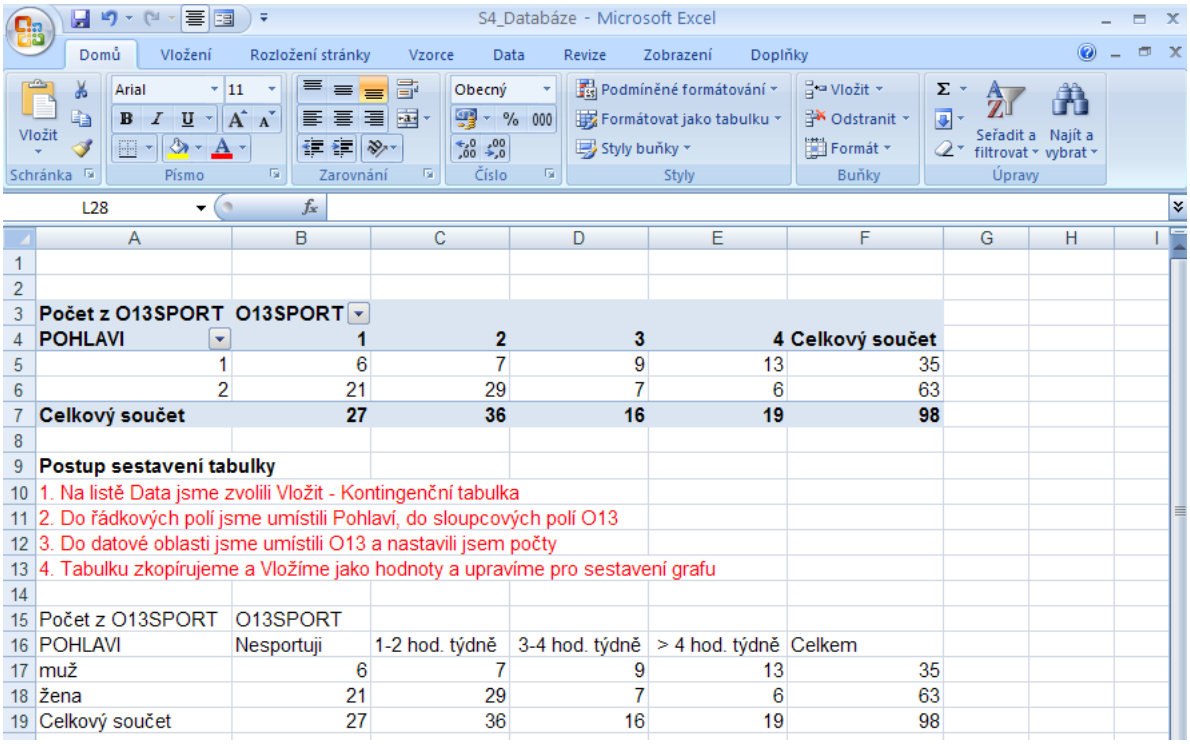
Příklad 4.1

Z údajů v databázi S4_Databáze (Preventivní prohlídky – muži i ženy), zjistěte zda se liší muži a ženy ve sportovní aktivitě. (Položky v databázi jsou shodné jako v databázi S3 – obr. 3.1)

a) Sestavte kontingenční tabulku se znaky pohlaví a O13sport.
b) Sestavte na základě tabulky graf.
c) Formulujte H_0, H_a, zvolte hladinu významnosti.
d) Vypočtěte testové kritérium χ^2, stupně volnosti, kritickou hodnotu – funkce CHIINV(), výsledek interpretujte.
e) Pro výpočet p-hodnoty použijte funkci CHITEST(), výsledek interpretujte.
f) Pro výpočet χ^2 – použijte funkci programu OpenEpi.

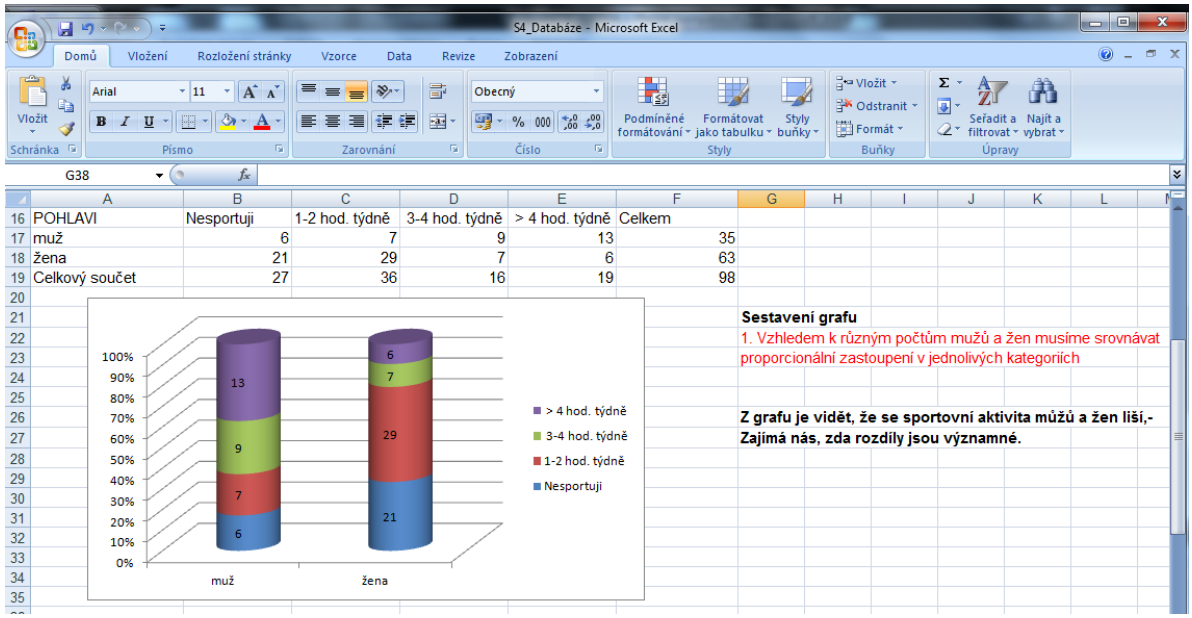
Řešení

a), b) Sestavení kontingenční tabulky se dvěma veličinami a sestavení grafu je pospáno v kapitole 2. Výsledek úkolů a), b) je na obr. 4.1 a 4.2.



	A	B	C	D	E	F
3	**Počet z O13SPORT**	**O13SPORT**				
4	**POHLAVI**	**1**	**2**	**3**	**4**	**Celkový součet**
5	1	6	7	9	13	35
6	2	21	29	7	6	63
7	**Celkový součet**	**27**	**36**	**16**	**19**	**98**
9	**Postup sestavení tabulky**					
10	1. Na listě Data jsme zvolili Vložit - Kontingenční tabulka					
11	2. Do řádkových polí jsme umístili Pohlaví, do sloupcových polí O13					
12	3. Do datové oblasti jsme umístili O13 a nastavili jsem počty					
13	4. Tabulku zkopírujeme a Vložíme jako hodnoty a upravíme pro sestavení grafu					
15	Počet z O13SPORT	O13SPORT				
16	POHLAVI	Nesportuji	1-2 hod. týdně	3-4 hod. týdně	> 4 hod. týdně	Celkem
17	muž	6	7	9	13	35
18	žena	21	29	7	6	63
19	Celkový součet	27	36	16	19	98

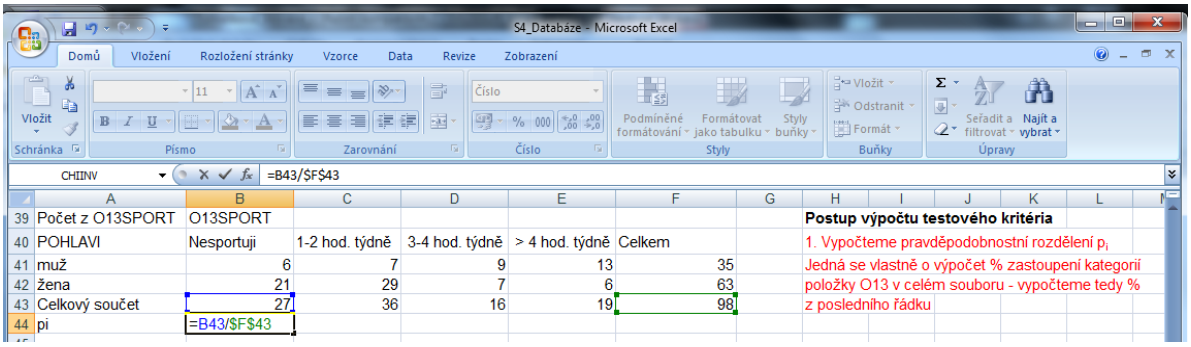
Obr. 4.1 Krok 1-4a) Sestavení kontingenční tabulky



Obr. 4.2 Krok 1b) Sestavení grafu

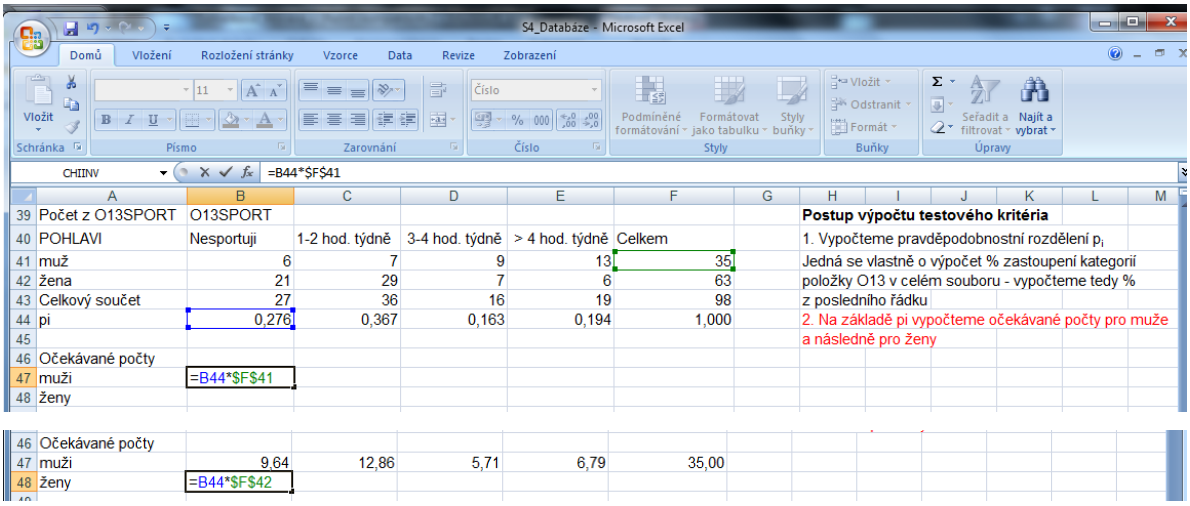
c) H_0: není rozdíl v rozložení kategorií položky O13_sport u mužů a žen.
H_a: je rozdíl rozložení kategorií položky O13_sport u mužů a žen.
Hladinu významnosti volíme obvykle 5 % ($\alpha = 0{,}05$).

d) Pro výpočet testového kritéria χ^2 použijeme uvedený vzorec v kapitole 4.1. Výpočet provedeme na základě tabulky, kterou jsme použili pro sestavení grafu (obr. 4.3).



	A	B	C	D	E	F
39	Počet z O13SPORT	O13SPORT				
40	POHLAVI	Nesportuji	1-2 hod. týdně	3-4 hod. týdně	> 4 hod. týdně	Celkem
41	muž	6	7	9	13	35
42	žena	21	29	7	6	63
43	Celkový součet	27	36	16	19	98
44	pi	=B43/F43				

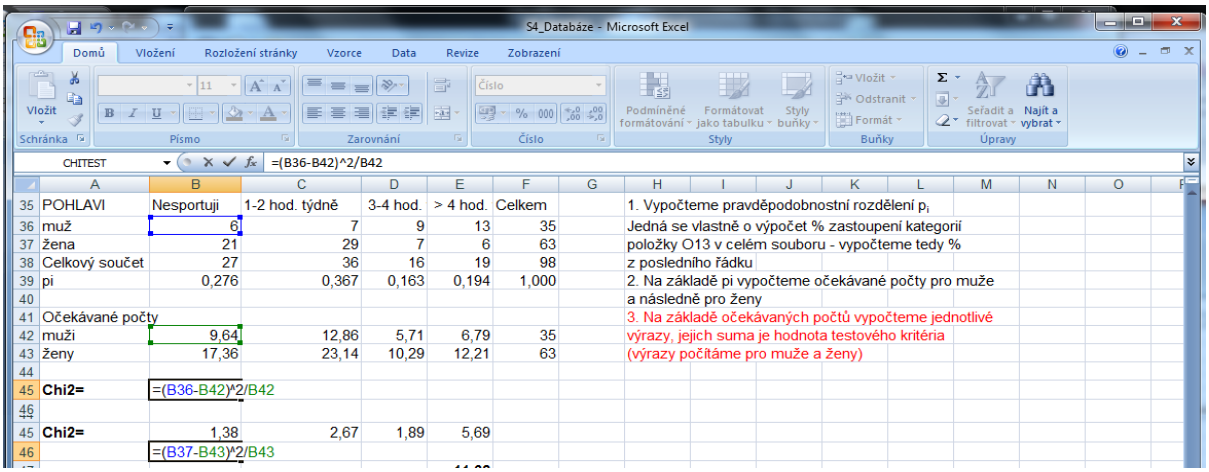
Obr. 4.3 Krok 1d) Výpočet p_i



	A	B	C	D	E	F
39	Počet z O13SPORT	O13SPORT				
40	POHLAVI	Nesportuji	1-2 hod. týdně	3-4 hod. týdně	> 4 hod. týdně	Celkem
41	muž	6	7	9	13	35
42	žena	21	29	7	6	63
43	Celkový součet	27	36	16	19	98
44	pi	0,276	0,367	0,163	0,194	1,000
45						
46	Očekávané počty					
47	muži	=B44*F41				
48	ženy					

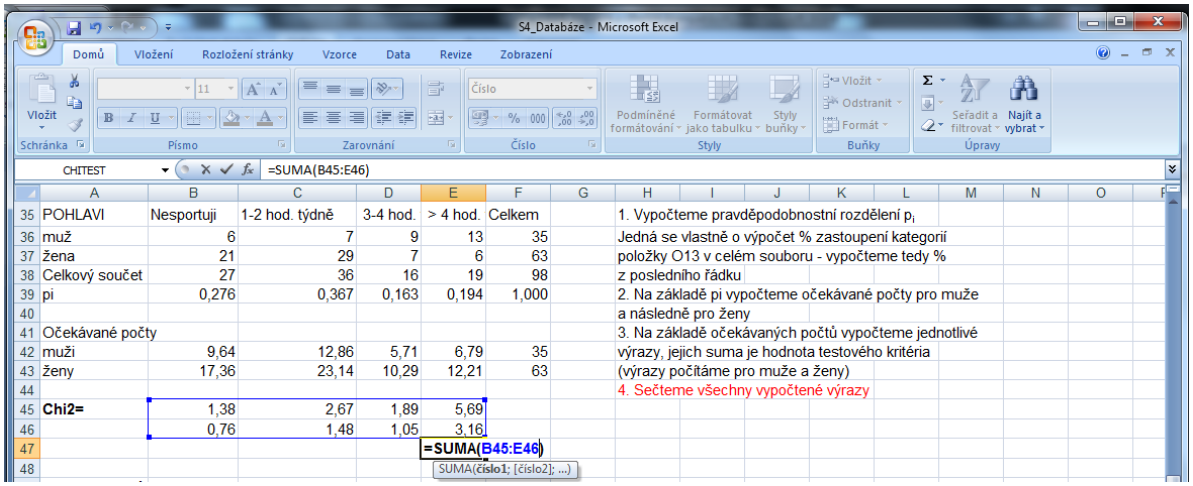
46	Očekávané počty				
47	muži	9,64	12,86	5,71	6,79	35,00
48	ženy	=B44*F42				

Obr. 4.4 Krok 2d) Výpočet očekávaných počtů pro muže a pro ženy

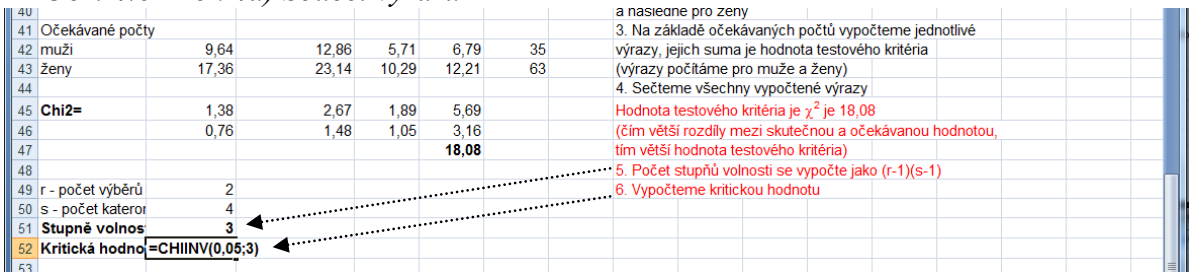


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
35	POHLAVI	Nesportuji	1-2 hod. týdně	3-4 hod.	> 4 hod.	Celkem		1. Vypočteme pravděpodobnostní rozdělení pi							
36	muž	6	7	9	13	35		Jedná se vlastně o výpočet % zastoupení kategorií							
37	žena	21	29	7	6	63		položky O13 v celém souboru - vypočteme tedy %							
38	Celkový součet	27	36	16	19	98		z posledního řádku							
39	pi	0,276	0,367	0,163	0,194	1,000		2. Na základě pi vypočteme očekávané počty pro muže							
40								a následně pro ženy							
41	Očekávané počty							3. Na základě očekávaných počtů vypočteme jednotlivé							
42	muži	9,64	12,86	5,71	6,79	35		výrazy, jejich suma je hodnota testového kritéria							
43	ženy	17,36	23,14	10,29	12,21	63		(výrazy počítáme pro muže a ženy)							
44															
45	**Chi2=**	=(B36-B42)^2/B42													
46															
45	**Chi2=**	1,38	2,67	1,89	5,69										
46		=(B37-B43)^2/B43													

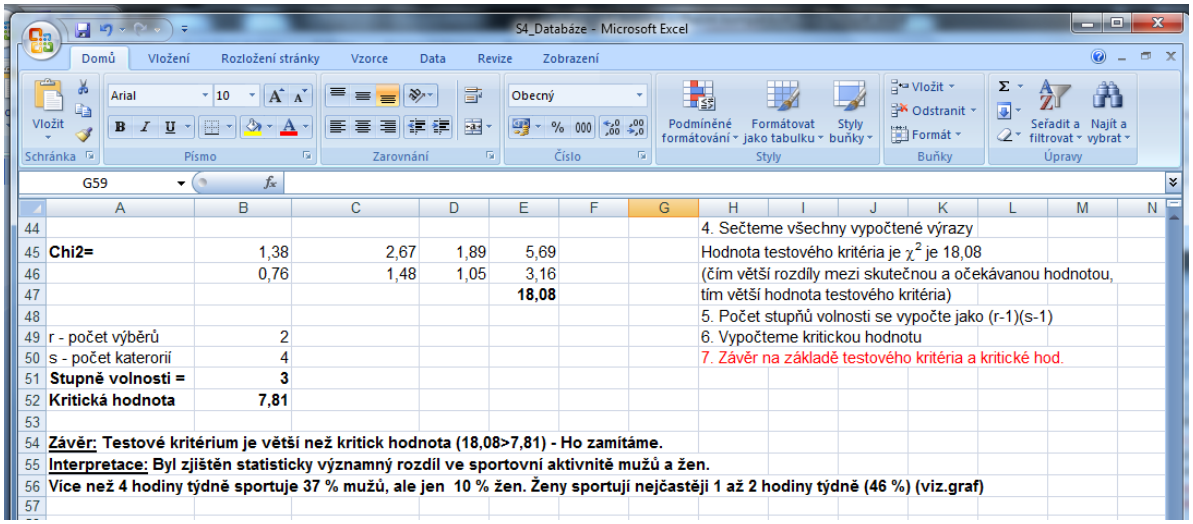
Obr. 4.5 Krok 3d) Výpočet jednotlivých výrazů testového kritéria



Obr. 4.6 Krok 4d) Součet výrazů

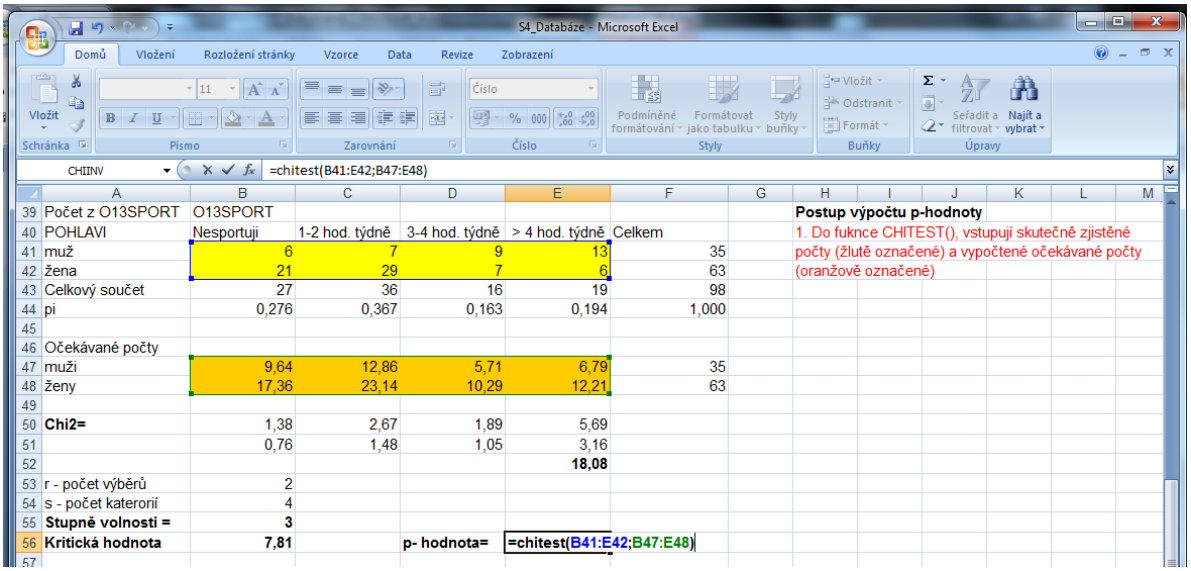


Obr. 4.7 Krok 5-6d) Výpočet stupňů volnosti a kritické hodnoty

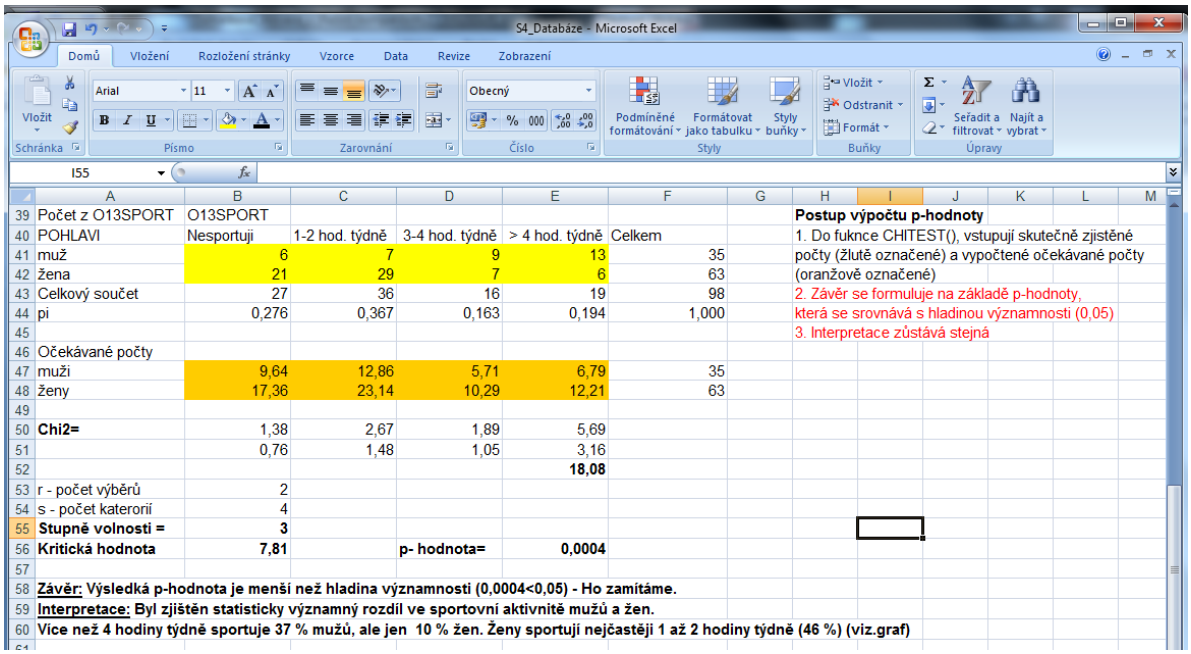


Obr. 4.8 Krok 7d) Interpretace výsledků

e) V bodě d) byl proveden manuální výpočet testového kritéria včetně kritické hodnoty. Pomocí funkce CHITEST() je možné na základě očekávaných hodnot vypočítat přímo p-hodnotu (obr. 4.9).

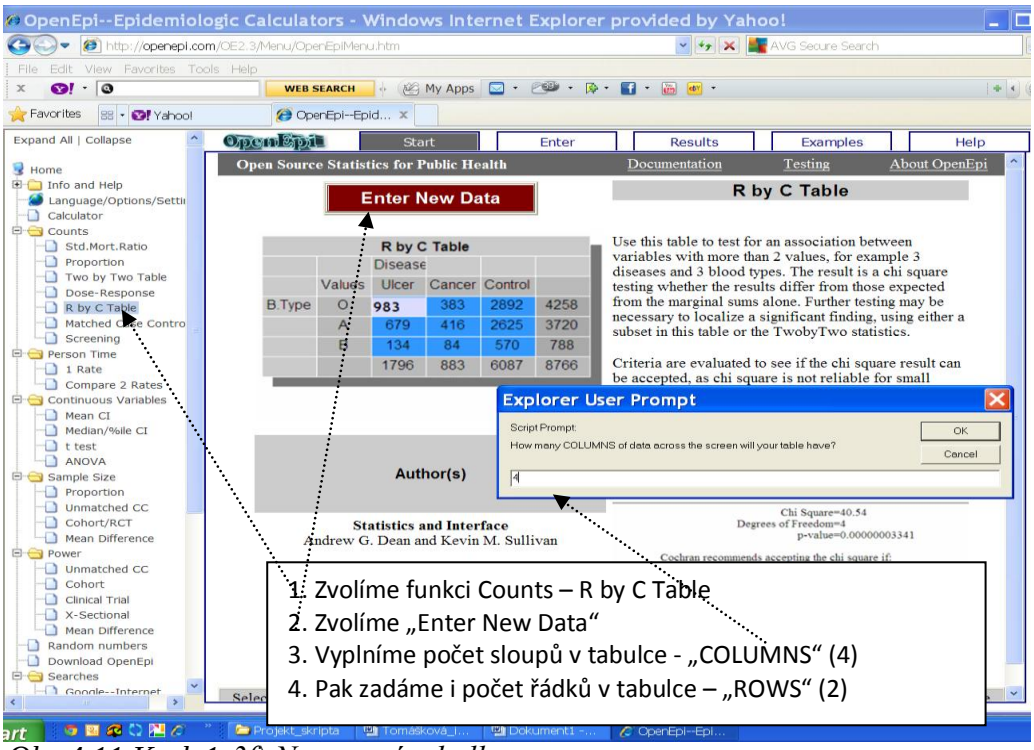


Obr. 4.9 Krok 1e) Výpočet p-hodnoty pomocí funkce CHITEST()

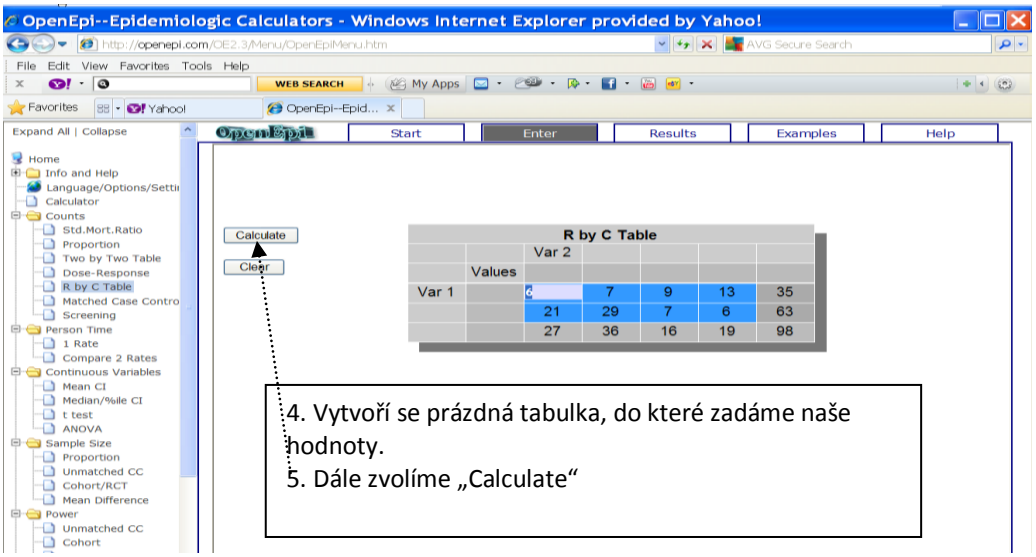


Obr. 4.10 Krok 2-3e) Interpretace na základě p-hodnoty

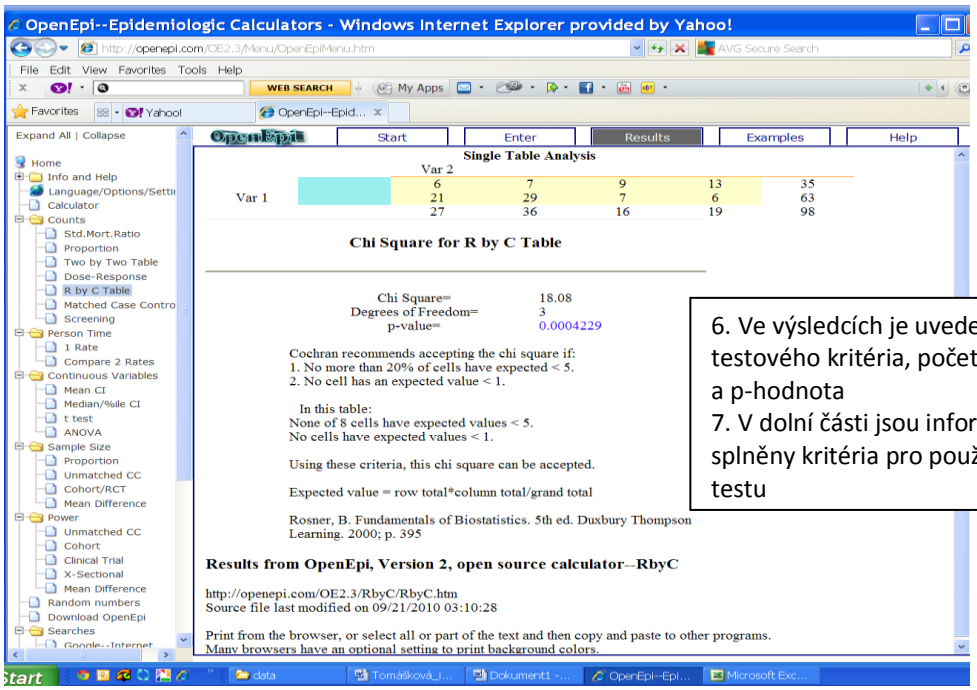
f) Pro výpočet testového kritéria χ^2 i p-hodnoty lze použít program OpenEpi (http://openepi.com/OE2.3/Menu/OpenEpiMenu.htm). Pro jeho výpočet budeme vycházet z hodnot ze sestavené kontingenční tabulky (Obr. 4.11 – 4.13).



Obr.4.11 Krok 1-3f) Nastavení tabulky



Obr. 4.12 Krok 4-5f) Vyplnění tabulky a výpočet



6. Ve výsledcích je uvedena hodnota testového kritéria, počet stupňů volnosti a p-hodnota
7. V dolní části jsou informace, zda byly splněny kritéria pro použití uvedeného testu

Obr. 4.13 Krok 6-7f) Výsledek χ^2 testu pomocí programu OpenEpi

Shrnutí obsahu kapitoly

Pearsonovo testové kritérium χ^2 se používá v případě, kdy má být ověřena hypotéza, že se sledovaná kvalitativní veličina se v několika různých situacích řídí stejným pravděpodobnostním rozdělením.
MS Excel – statistické funkce CHIINV(), CHITEST()
OpenEpi – funkce Counts – R by C Table

Kontrolní otázky:

1. Kdy použijeme χ^2 test pro dva a více výběrů, jak zní Ho?
2. Co vyjadřuje hladina statistické významnosti?
3. Jak budete interpretovat výsledek, když p-hodnota bude 0,03?
4. Jak budete interpretovat výsledek, když p-hodnota bude 0,33?

Úkol

4.1 Pokračujte v analýze údajů v databázi S4_Databáze

a) Vytvořte novou položku k otázce O1Praktik, ve které uvedete, zda studenti opověděli na tuto otázku správně nebo chybně (správná odpověď je 3).
b) Zjistěte, zda existují rozdíly mezi muži a ženami ve znalostech o preventivních prohlídkách u praktického lékaře (viz. nová položka bod a)).
c) Zjistěte, zda existují rozdíly mezi muži a ženami v kouření (bude nutné sloučit kategorie, aby byla splněna podmínka použití χ^2 testu).

5 Statistické testy pro kvantitativní data

V této kapitole se dozvíte:

- Jaký je rozdíl mezi dvouvýběrovým t-testem a párovým t-test, jak zní nulové hypotézy.
- Jak vypočítat p-hodnotu pro dvouvýběrový t-test pomocí funkce MS Excelu a programu OpenEpi.
- Jak vypočítat p-hodnotu pro párový t-test pomocí funkce MS Excelu
- Jak se interpretují výsledky.
- Jaký je rozdíl mezi jednostrannou a oboustrannou hypotézou

Po jejím prostudování byste měli být schopni:

- Určit, zda použijete dvouvýběrový t-test nebo párový t-test.
- Rozhodnout, zda použijete jednostrannou nebo oboustrannou hypotézu.
- Rozhodnout, zda zvolíte v případě dvouvýběrového t-testu, variantu se shodnými nebo neshodnými rozptyly.
- Provést výpočet t-testu pomocí funkcí MS Excelu a programu Open Epi.
- Interpretovat výsledky statistických testů.

Klíčová slova této kapitoly: dvouvýběrový t-test, párový t-test, F-test, jednostranná hypotéze, oboustranná hypotéza

Doba potřebná ke studiu a zpracování úkolů: 4 hodiny

Průvodce studiem

Pokud student pochopil základní princip testování statistických hypotéz, pak tato kapitola nebude pro něj náročná. V případě kvantitativních dat se nejčastější řeší problém srovnání středních hodnot mezi dvěma výběry (např. muži, ženy) nebo změna hodnot v závislosti na nějaké intervenci nebo čase (např. léčba), v tomto případě se jedná o data párová. Je nutné si určit při definování H_0, který test v závislosti na datech má být proveden.
Pro výpočet budou použity jak funkce MS Excelu, tak program OpenEpi.
Náročnost této kapitoly je střední.
Teorie k uvedeným tématům je probírána ve čtvrté a šesté přednášce Lékařská biofyzika, výpočetní technika I – Biostatistika.

5.1 Dvouvýběrový t-test

Pokud je sledován metrický znak s přibližně normální rozložením ve dvou výběrech, můžeme pomocí dvouvýběrového t-test testovat H_0, která předpokládá, že střední hodnoty těchto dvou populací se neliší.
H_0: $\mu_1=\mu_2$, kde μ je populační parametr označující střední hodnotu. Alternativní hypotéza může být definována jako oboustranná H_a: $\mu_1\neq\mu_2$ nebo jednostranná H_a: $\mu_1<$ (resp. $>$)μ_2.
Testovým kritériem je statistika *t* (přednáška 6). Do vzorce pro výpočet testového kritéria vstupují základní charakteristiky obou výběru: rozsah výběrů

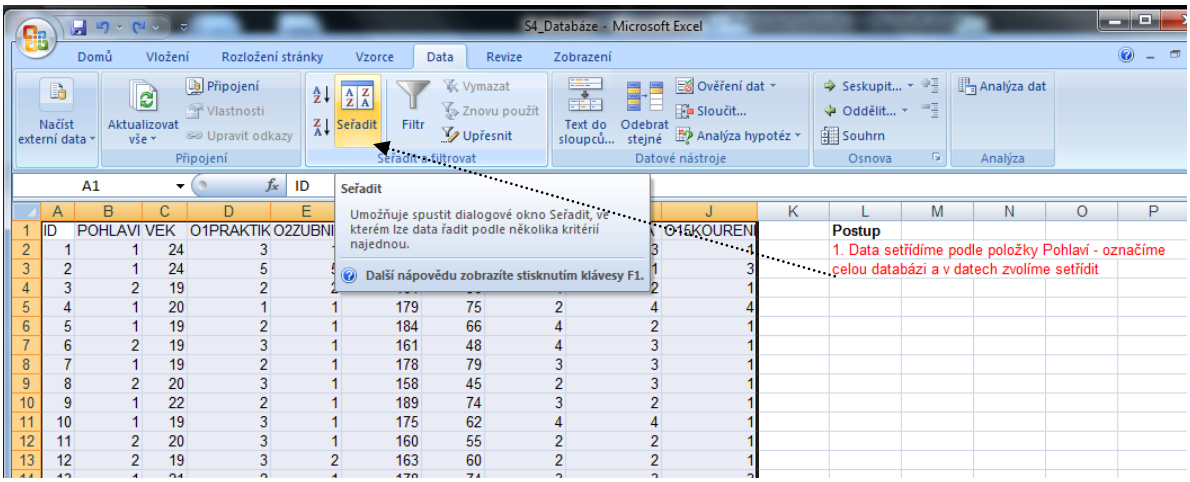
n_1 , n_2, aritmetické průměry $\overline{x_1}$, $\overline{x_2}$ a rozptyly s_1^2, s_2^2. Na základě vztahu mezi rozptyly zvolíme *t*-test pro dva výběry se shodnými rozptyly nebo *t*-test pro dva výběry s různými rozptyly. Pro testování H_0: $\sigma_1^2=\sigma_2^2$, použijeme *F*-test (σ^2 – populační rozptyl).
Výsledkem testu bude *p*-hodnota. Na základě srovnání *p*-hodnoty a hladiny významnosti (α=0,05) vyslovíme závěr. Pokud je *p*-hodnota menší než hladina významnosti, pak H_0 zamítáme, v opačném případě H_0 nezamítáme.

Příklad 5.1
Z údajů v databázi S4_Databáze (Preventivní prohlídky – muži i ženy) zjistěte, zda se liší muži a ženy v průměrné hodnotě výšky.

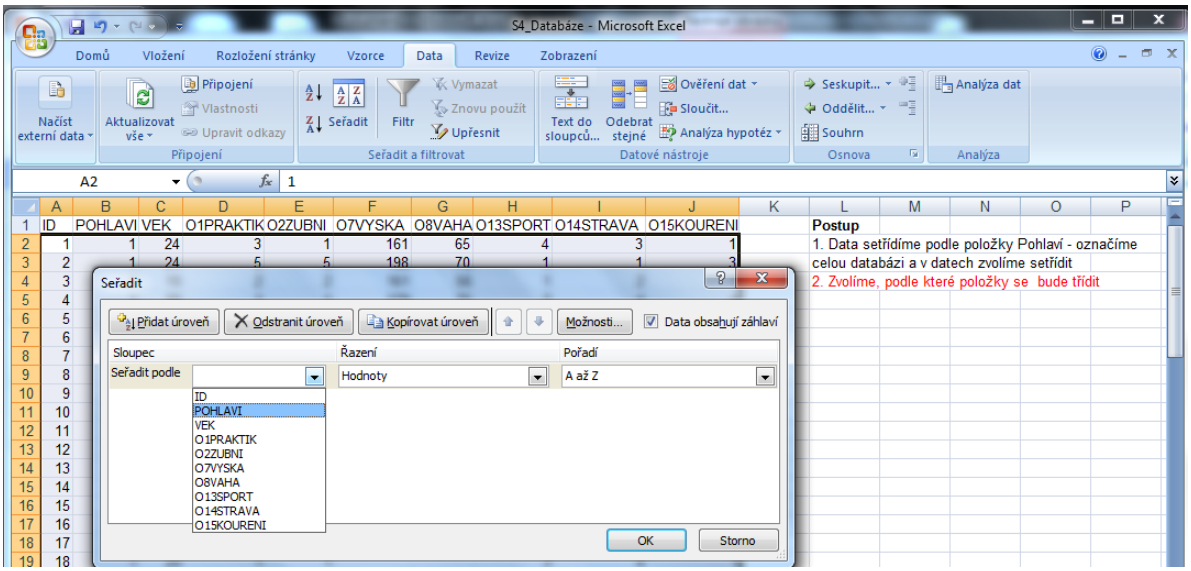
a) Vypočtěte základní charakteristiky pro výšku u mužů a žen.
b) Formulujte H_0 a H_a a stanovte hladinu významnosti.
c) Pro testování H_0: $\sigma_z^2=\sigma_m^2$ použijte *F*-test.
d) Proveďte *t*-test a výsledek interpretujte.

Řešení

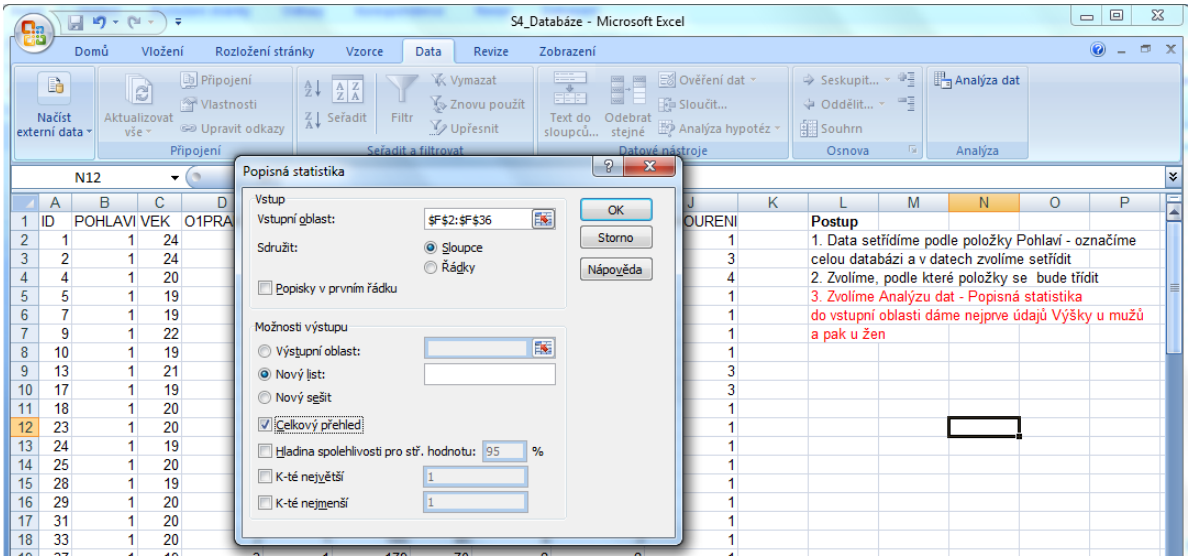
a) Pro výpočet základních charakteristik zvolíme funkci Analýza dat – Popisná statistika. Do Analýzy dat musí vstupovat blok hodnot tzn. výška mužů a výška žen, toho docílíme setříděním databáze. Databázi tedy nejprve setřídit podle položky pohlaví (obr. 5.1 – 5.4). Nastavíme se do listu Data a použijeme funkci Seřadit.



Obr. 5.1 Krok 1a) Výběr funkce setřídění



Obr. 5.2 Krok 2a) Výběr položky podle které se bude třídit



Obr. 5.3 Krok 3a) Zvolení funkce Popisné statistika

b) Formulace H_0 a H_a.

H_0: není rozdíl v průměrné výšce u mužů a žen

H_a: je rozdíl v průměrné výšce u mužů a žen (oboustranná hypotéza viz. Strany)

Hladina významnosti 5 % (0,05)

Uvedenou H_0 budeme testovat pomocí t-testu, použijeme statistickou funkci Ttest(pole1, pole2, strany, typ) kde,

Pole 1 – hodnoty výšky u mužů

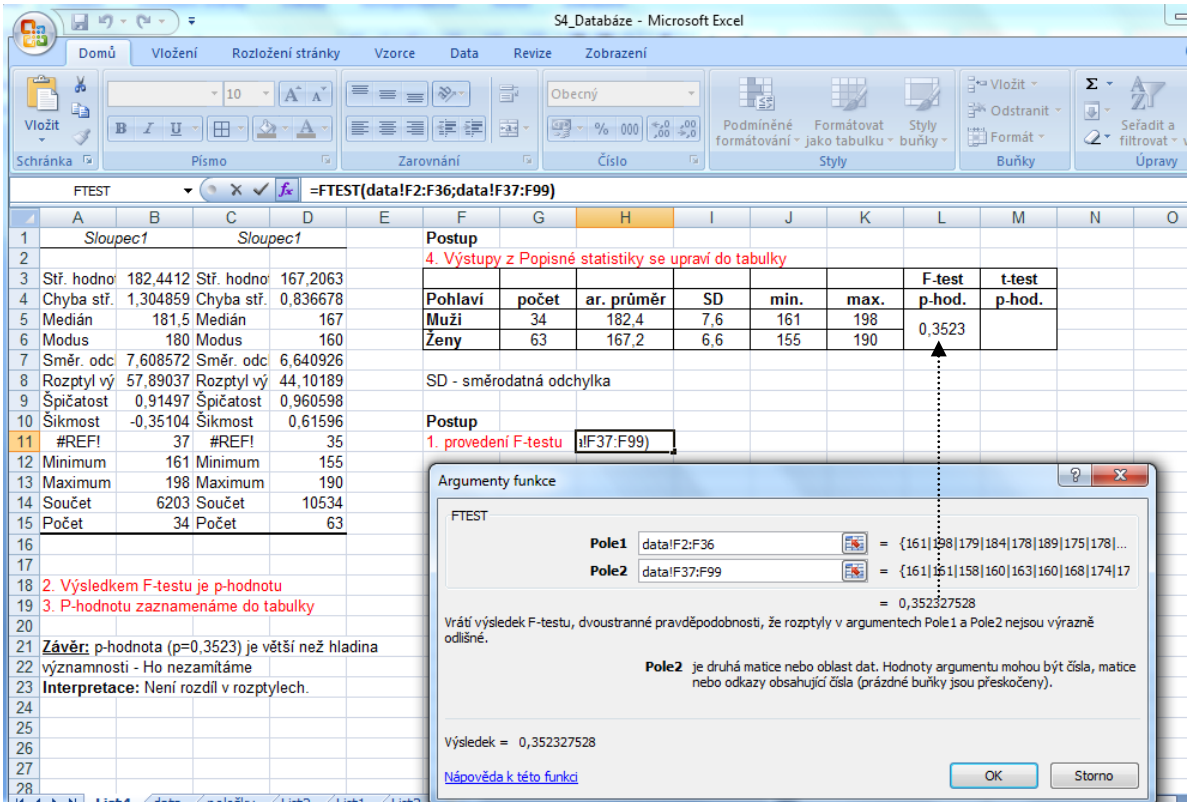
Pole 2 – hodnoty výšky u žen

Strany – 1 jednostranné rozdělení, 2 oboustranné rozdělení (zvolíme na základě typ H_a, v našem případě 2)

Typ – 1 spárované výběry, 2 dva výběry se shodným rozptylem, 3 dva výběry s různým rozptylem

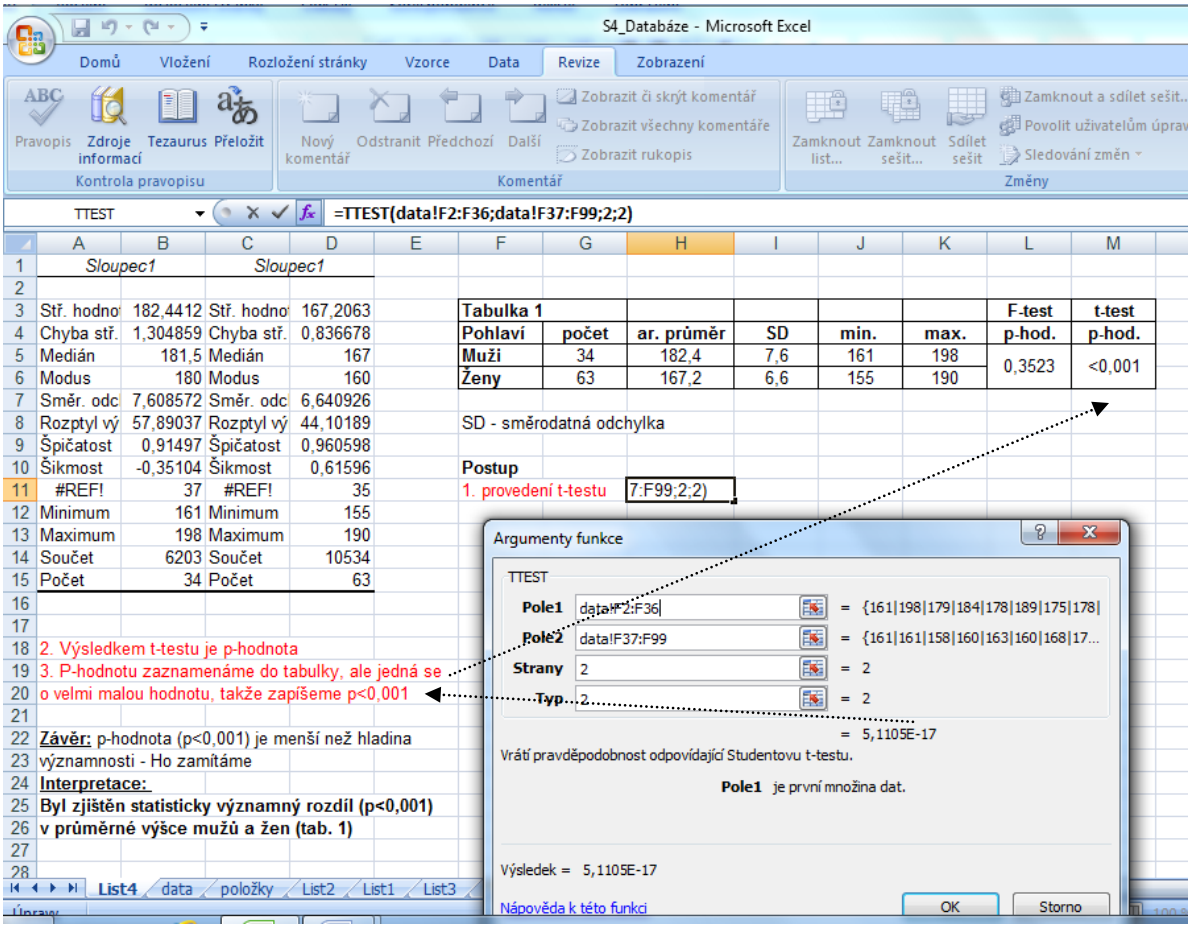
Typ 2 nebo 3 můžeme zvolit až na základě provedení F-testu.

c) *F*-test testuje:
H_0: není rozdíl v rozptylech u výšky u mužů a žen
H_a: je rozdíl v rozptylech
provedeme pomocí statistické funkce Ftest(pole1, pole2)
Pole1 – výška u mužů
Pole 2 – výška u žen
Výsledem *F*-testu je *p*-hodnota (obr. 5.4).



Obr. 5.4 Krok 4a +1-3c) Výpočet F-testu

d) Na základě výsledku *F*-testu zvolíme Typ (v zadání Ttestu()) 2 - dva výběry se shodným rozptylem (obr. 5.5).



	A	B	C	D
1	*Sloupec1*		*Sloupec1*	
2				
3	Stř. hodno	182,4412	Stř. hodno	167,2063
4	Chyba stř.	1,304859	Chyba stř.	0,836678
5	Medián	181,5	Medián	167
6	Modus	180	Modus	160
7	Směr. odc	7,608572	Směr. odc	6,640926
8	Rozptyl vý	57,89037	Rozptyl vý	44,10189
9	Špičatost	0,91497	Špičatost	0,960598
10	Šikmost	-0,35104	Šikmost	0,61596
11	#REF!	37	#REF!	35
12	Minimum	161	Minimum	155
13	Maximum	198	Maximum	190
14	Součet	6203	Součet	10534
15	Počet	34	Počet	63

Tabulka 1

Pohlaví	počet	ar. průměr	SD	min.	max.	F-test p-hod.	t-test p-hod.
Muži	34	182,4	7,6	161	198	0,3523	<0,001
Ženy	63	167,2	6,6	155	190		

SD - směrodatná odchylka

Postup

1. provedení t-testu 7:F99;2;2)

2. Výsledkem t-testu je p-hodnota
3. P-hodnotu zaznamenáme do tabulky, ale jedná se o velmi malou hodnotu, takže zapíšeme p<0,001

Závěr: p-hodnota (p<0,001) je menší než hladina významnosti - Ho zamítáme

Interpretace:

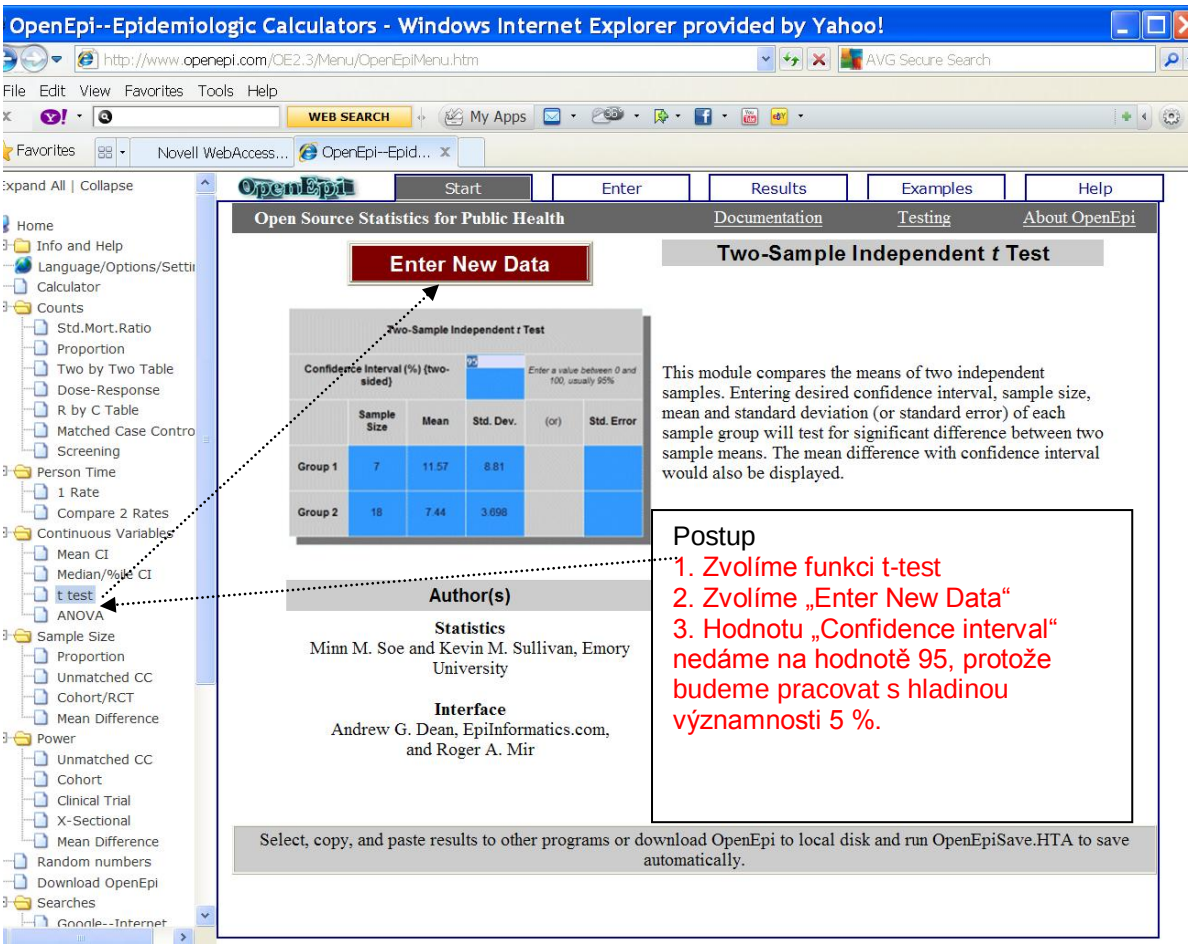
Byl zjištěn statisticky významný rozdíl (p<0,001) v průměrné výšce mužů a žen (tab. 1)

Obr. 5.5 Krok 1-3d) Výpočet t-testu a interpretace

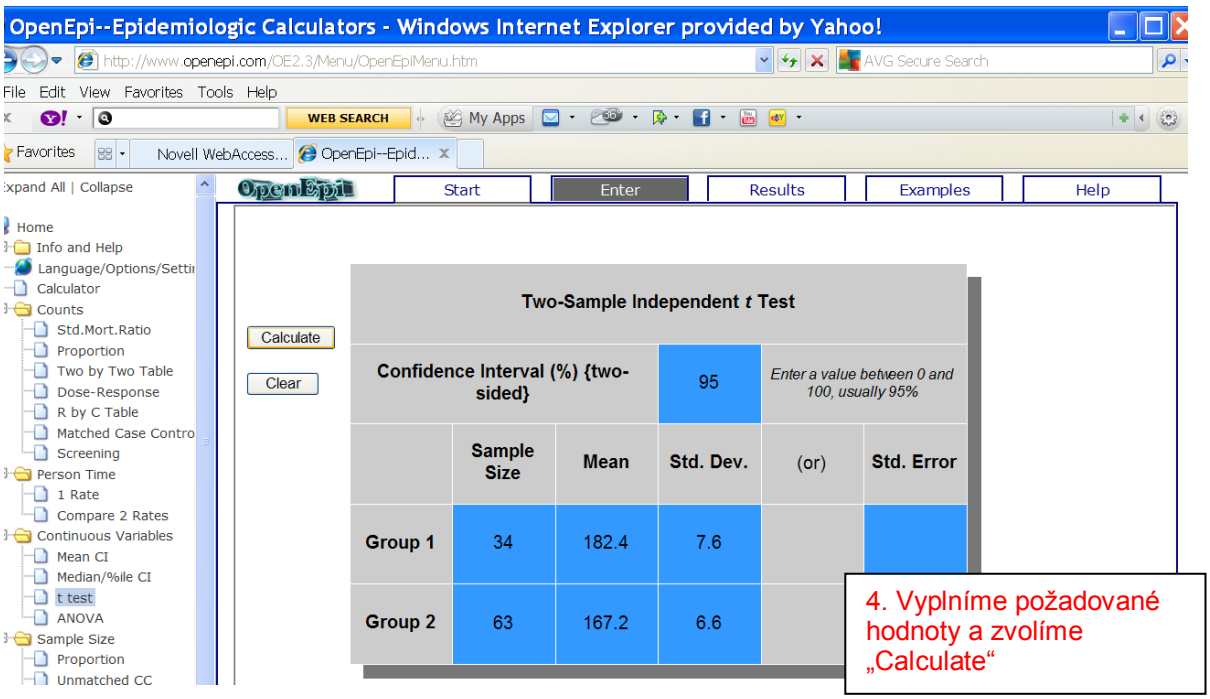
Výpočet F-test a t-test můžeme také provést pomocí programu OpenEpi přes funkci Continuous Variables – t test

Vstupními údaji jsou základní charakteristiky za oba výběry (Group 1- muži, Group 2 –ženy): Sample size – počet respondentů, mean – aritmetický průměr, Std. Dev. – směrodatná odchylka.

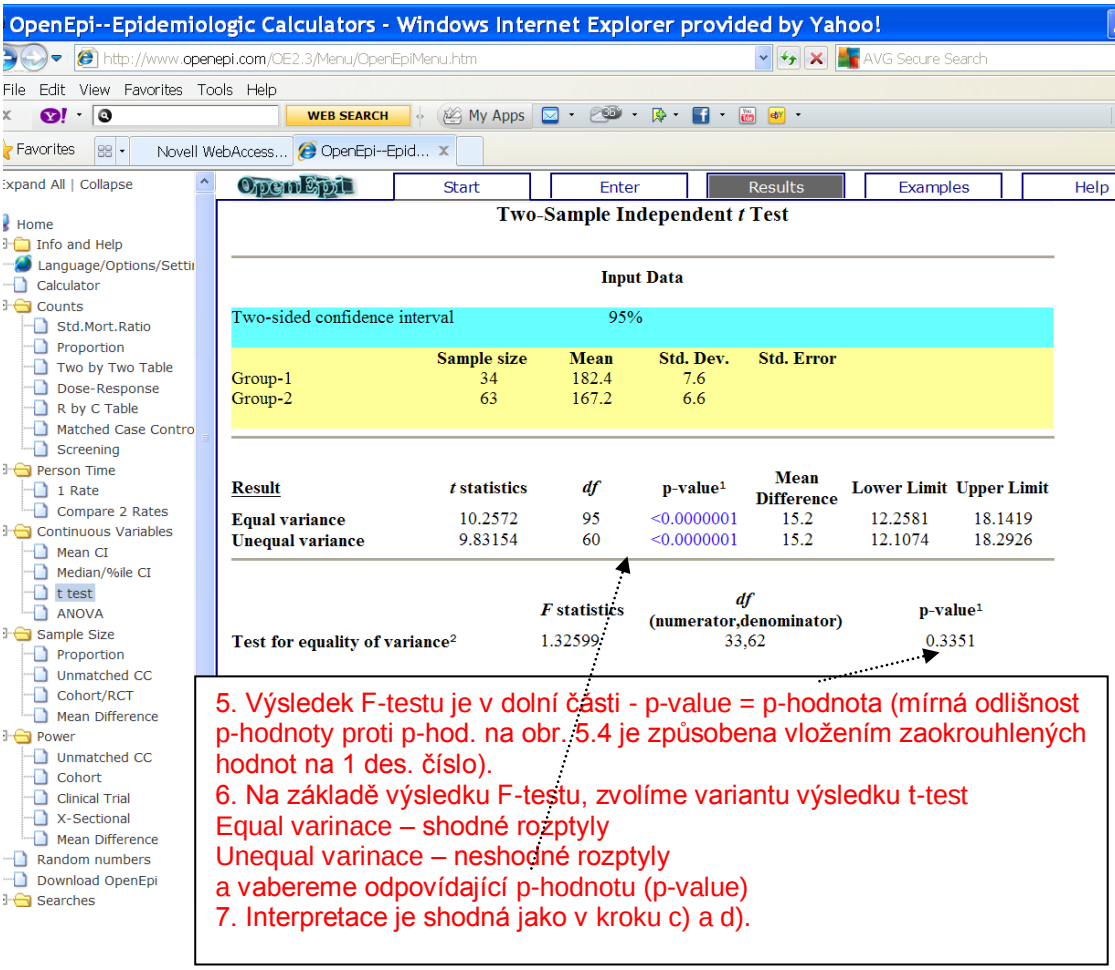
Pozor! Desetinná čísla se musí zadávat pomocí desetinné tečky, ne čárky.



Obr. 5.6 Krok 1-3 Výpočet t-testu pomocí programu OpenEpi



Obr. 5.7 Krok 4 Výpočet t-testu pomocí programu OpenEpi



Obr. 5.8 Krok 4-7 Výsledek t-testu pomocí programu OpenEpi

5.2 Párový t-test

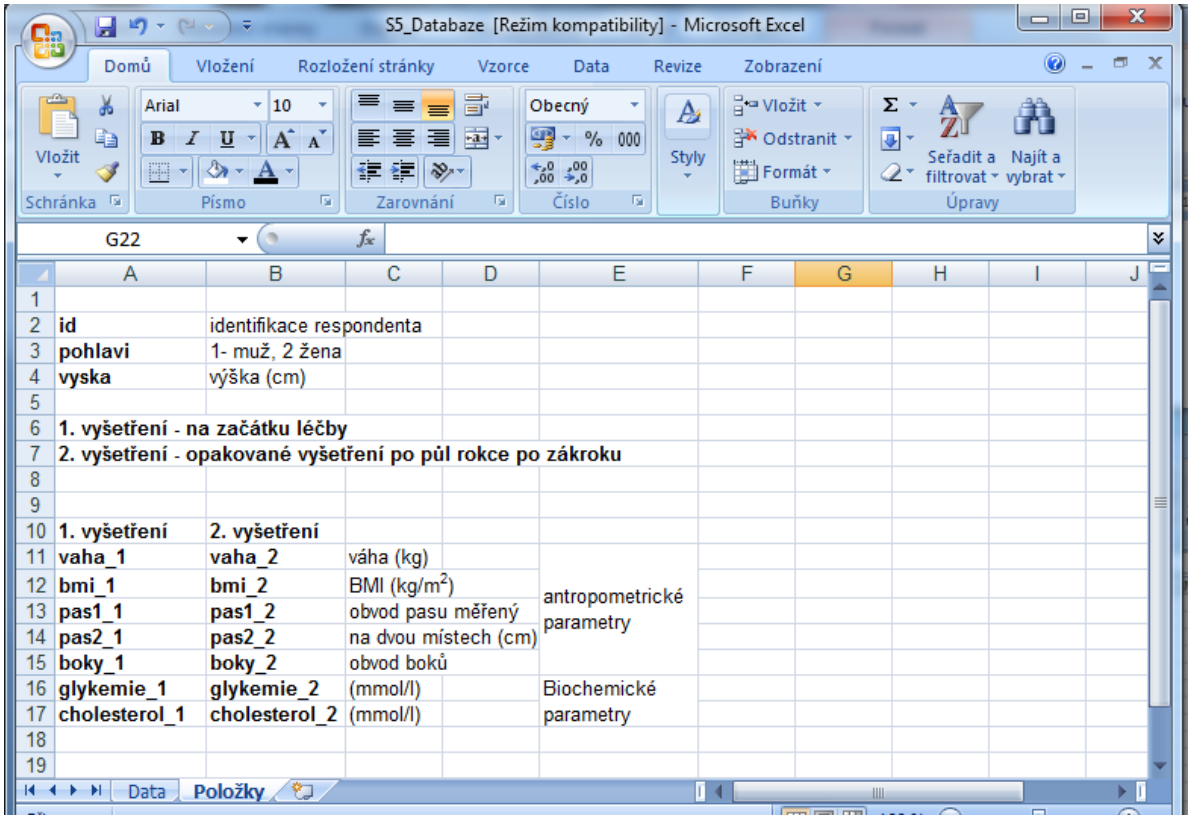
V případě, že sledujeme metrický znak s přibližně normálním rozdělením u jednoho výběru ve dvou situacích (např. před léčbou a po léčbě), použijeme pro testování H_0 párový *t*-test.

Pro výpočet testového kritéria *t* musíme vypočítat aritmetický průměr z diferencí hodnot v prvním a druhém měření (x_{1i}-x_{2i}) a rozptyl z těchto diferencí (přednáška 4). Nulová hypotéza většinou předpokládá, že H_0: $\mu_d=0$ (μ_d – střední hodnota z diferencí), alternativní hypotéza může být opět sestavena jako oboustranná H_0: $\mu_d\neq0$ nebo jednostranná H_0: $\mu_d<$ (resp. >) 0. Testové kritérium má Studentovo *t*-rozdělení o *n-1* stupních volnosti (*n* je rozsah výběru). Když testové kritérium je větší než kritická hodnota, pak H_0 zamítáme. Nebo využijeme *p*-hodnotu. A závěr provedeme na základě srovnání *p*-hodnoty a hladiny významnosti (α=0,05). Pokud je *p*-hodnota menší než hladina významnosti, pak H_0 zamítáme.

Příklad 5.2

Z údajů v databázi S5_Databáze (Redukce hmotnosti), zjistěte, zda se liší váha před terapií a za půl roku po terapii. (Jedná se o soubor pacientů, kteří podstoupili chirurgickou léčbu obezity a byly u nich sledovány různé parametry – antropometrické i biochemické (obr. 5.9)

a) Vypočtěte diferenci mezi váhou před a po terapii.
b) Vypočtěte základní charakteristiky pro váhu před a po terapii a pro diferenci váhy.
c) Formulujte H_0 a H_a, stanovte hladinu významnosti, vypočtěte p-hodnotu.
d) Výsledek interpretujte.

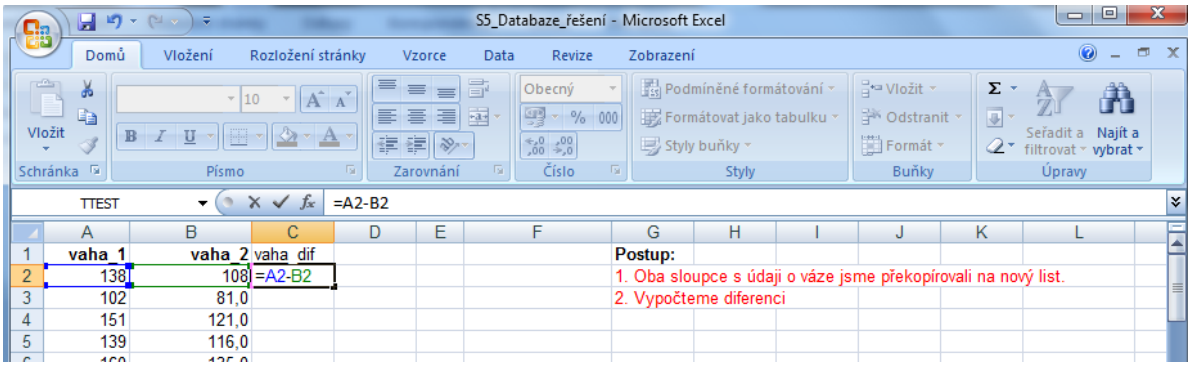


	A	B	C	D	E	F	G	H	I	J
1										
2	**id**	identifikace respondenta								
3	**pohlavi**	1- muž, 2 žena								
4	**vyska**	výška (cm)								
5										
6	**1. vyšetření - na začátku léčby**									
7	**2. vyšetření - opakované vyšetření po půl rokce po zákroku**									
8										
9										
10	**1. vyšetření**	**2. vyšetření**								
11	**vaha_1**	**vaha_2**	váha (kg)							
12	**bmi_1**	**bmi_2**	BMI (kg/m^2)		antropometrické					
13	**pas1_1**	**pas1_2**	obvod pasu měřený		parametry					
14	**pas2_1**	**pas2_2**	na dvou místech (cm)							
15	**boky_1**	**boky_2**	obvod boků							
16	**glykemie_1**	**glykemie_2**	(mmol/l)		Biochemické					
17	**cholesterol_1**	**cholesterol_2**	(mmol/l)		parametry					
18										
19										

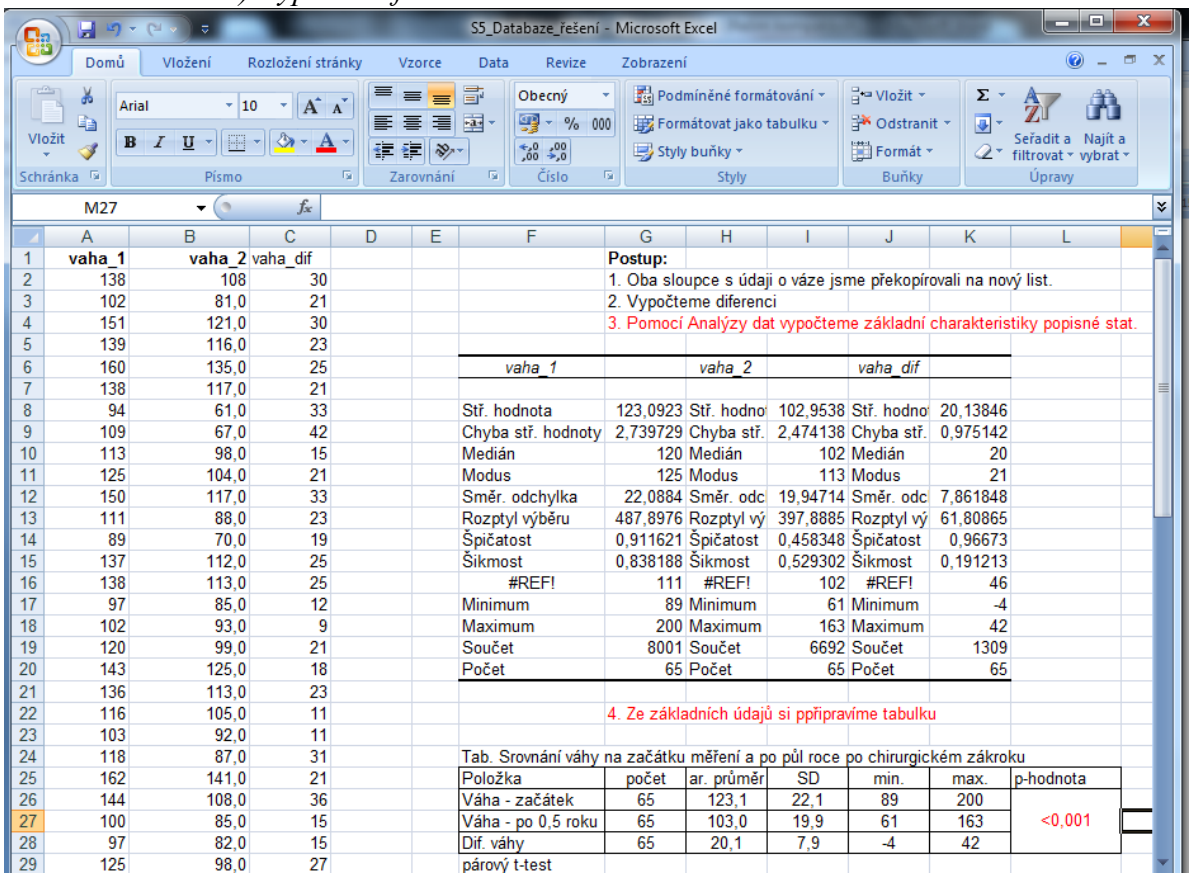
Obr. 5.9 Položky databáze S5

Řešení

Před samotným výpočtem si zkopírujeme oba sloupce s váhou na nový list. Tento krok není nutný, naopak všechny zásahy do databáze tzn. přidávání položek, mají být provedeny na listu Data, ale vzhledem k názornosti jsme data zkopírovali a budeme dále pracovat na novém listu Váha (obr. 5.10 – 5.13).



Obr. 5.10 Kok 1a) Výpočet diference



	vaha_1	*vaha_2*	*vaha_dif*
Stř. hodnota	123,0923	102,9538	20,13846
Chyba stř. hodnoty	2,739729	2,474138	0,975142
Medián	120	102	20
Modus	125	113	21
Směr. odchylka	22,0884	19,94714	7,861848
Rozptyl výběru	487,8976	397,8885	61,80865
Špičatost	0,911621	0,458348	0,96673
Šikmost	0,838188	0,529302	0,191213
#REF!	111	102	46
Minimum	89	61	-4
Maximum	200	163	42
Součet	8001	6692	1309
Počet	65	65	65

Tab. Srovnání váhy na začátku měření a po půl roce po chirurgickém zákroku

Položka	počet	ar. průměr	SD	min.	max.	p-hodnota
Váha - začátek	65	123,1	22,1	89	200	
Váha - po 0,5 roku	65	103,0	19,9	61	163	<0,001
Dif. váhy	65	20,1	7,9	-4	42	
párový t-test						

Obr. 5.11 Krok 2a) Výpočet popisných charakteristik

c) Formulace H_0 a H_a

H_0: průměrná diference (rozdíl) váhy na začátku měření a po půl roce se neliší od nuly.

H_{a1}: průměrná diference se liší od nuly (oboustranná hypotéza viz. Strany)

H_{a2}: průměrná diference je větší než nula (jednostranná hypotéza viz. Strany) – zvolíme jednostrannou hypotézu, protože předpokládáme, že chirurgický zákrok je účinný.

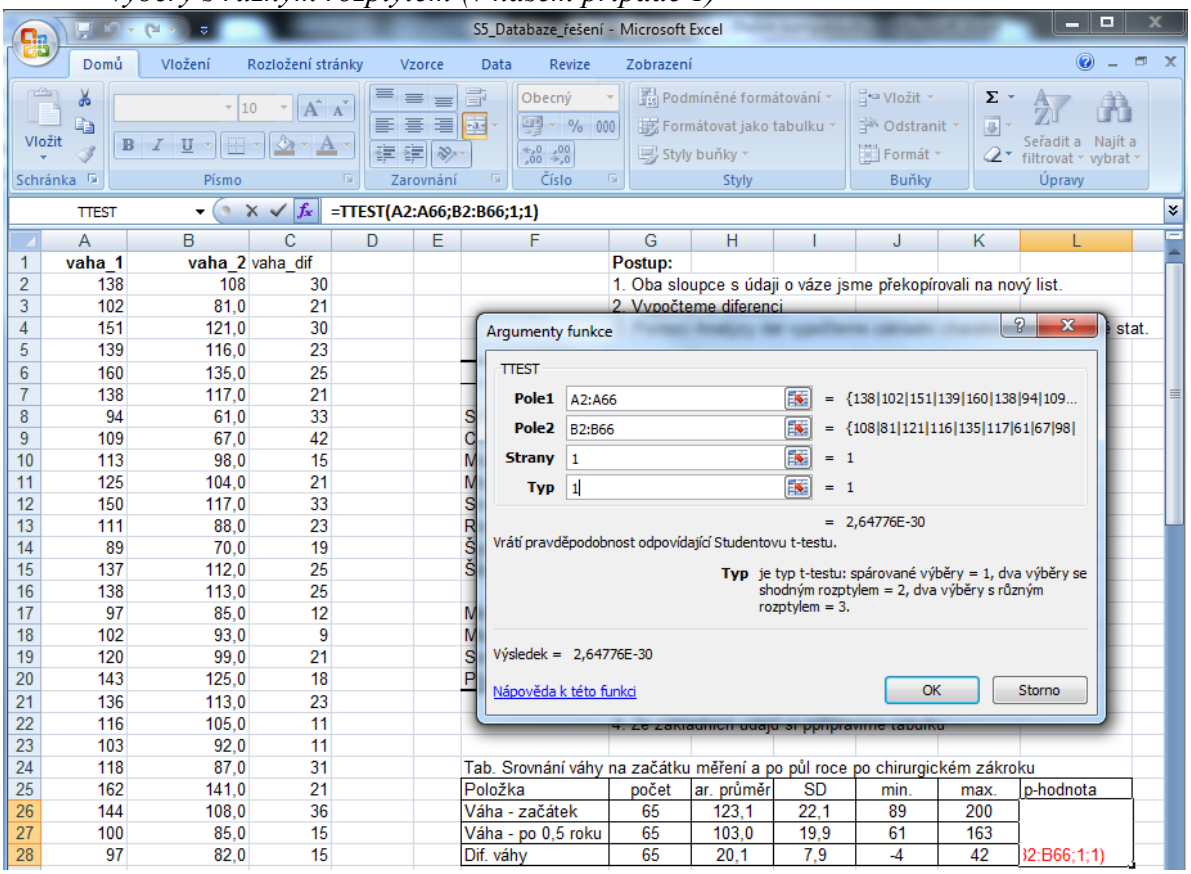
Hladina významnosti 5 % (0,05)

Uvedenou H_0 budeme testovat pomocí t-testu, použijeme statistickou funkci Ttest(pole1, pole2, strany, typ) obdobně jako v kapitole 5.1 (obr. 5.12)

Pole 1 – hodnoty váhy1

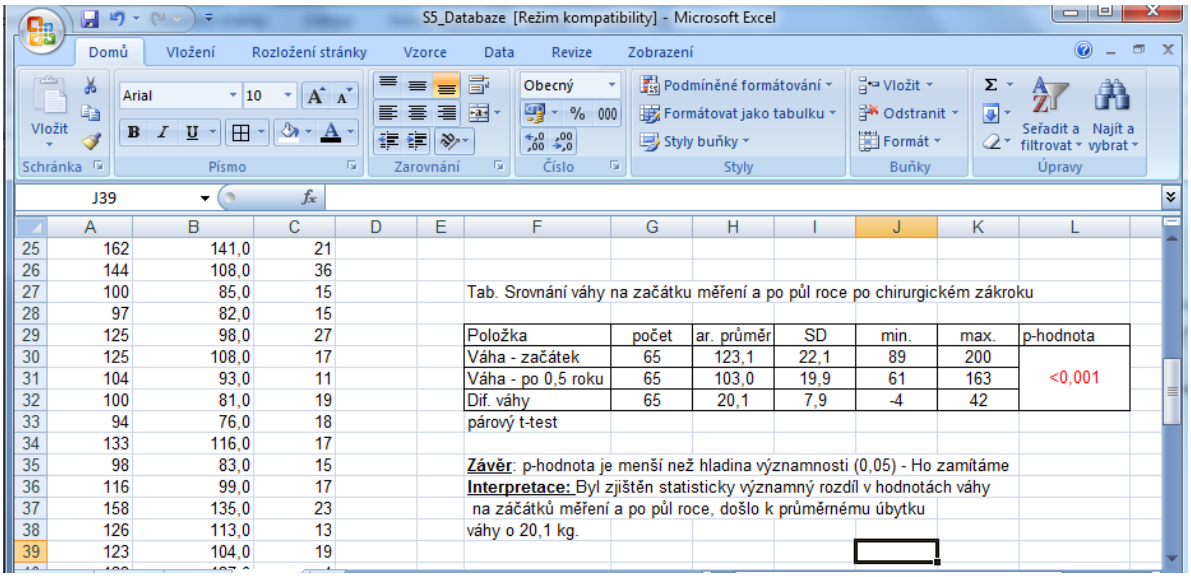
Pole 2 – hodnoty váhy2

Strany – 1 jednostranné rozdělení, 2 oboustranné rozdělení (zvolme na základě typ H_a, v našem případě 1)
Typ – 1 spárované výběry, 2 dva výběry se shodným rozptylem, 3 dva výběry s různým rozptylem (v našem případě 1)



Obr. 5.12 Krok 1c) Výpočet p-hodnoty ($p=2,6477.10^{-30}$)

d) Jak vidíme p hodnota je velmi malá, pomocí matematického formátu je vyjádřena jako *2,64776E-30*, znamená to, že první platná číslice je až na 30. pozici za desetinnou tečkou. Výsledek tedy do tabulky zapíšeme ve tvaru <0,001 (obr. 5.13).



Obr. 5.13 Krok 1d) Závěr a interpretace

Shrnutí obsahu kapitoly

V případě kvantitativních dat se nejčastější řeší problém srovnání středních hodnot mezi dvěma výběry (např. muži, ženy) nebo změna hodnot v závislosti na nějaké intervenci nebo čase (např. léčba), v tomto případě se jedná o data párová. Pokud srovnáváme střední hodnoty dvou výběrů, použijeme dvouvýběrový t-test, v případě párových dat použijeme párový t-test.

MS Excel – statistické funkce Ftest(), ttest()
OpenEpi – funkce Continuous Variables – t test

Kontrolní otázky:

1. Jaký je rozdíl mezi dvouvýběrovým t-test a párovým t-testem?
2. Jaký test použijeme pro testování shody rozptylů?
3. Jaká budete interpretovat výsledek, když p-hodnota bude 0,03?
4. Jaká budete interpretovat výsledek, když p-hodnota bude 0,33?

Úkol

5.1 Pokračujte v analýze údajů v databázi S4_Databáze:

a) Vypočtěte BMI (kg/m^2)
b) Zjistěte, zda existují rozdíly v průměrné hodnotě BMI mezi muži a ženami.
c) Formulujte H_0 a H_a, stanovte hladinu významnosti a vypočtěte p-hodnotu.
d) Výsledky interpretujte.

5.2 Pokračujte v analýze údajů v databázi S5_Databáze:

a) Zjistěte, zda a jak se změnil obvod boků (před a po terapii).
b) Zjistěte, zda a jak se změnily hodnoty glykemie (před a po terapii).
c) Formulujte H_0 a H_a, stanovte hladinu významnosti, vypočtěte p-hodnotu.
d) Výsledky interpretujte.

6 Analýza vztahu dvou metrických veličin

V této kapitole se dozvíte:

- Jak se graficky znázorňuje závislost dvou metrických (spojitých) veličin
- Co popisuje korelační koeficient a jakých může nabývat hodnot
- Jak můžeme na základě znalosti jedné veličiny odhadnout jinou veličinu
- Jak se vypočítá korelační koeficient pomocí funkce MS Excelu
- Jak lze vyjádřit závislost dvou metrických veličin pomocí funkcí MS Excelu

Po jejím prostudování byste měli být schopni:

- Sestavit bodový graf, který popisuje vztah dvou metrických veličin.
- Vypočítat korelační koeficient.
- Pomocí spojnice trendu vyjádřit závislost dvou metrických veličin.

Klíčová slova této kapitoly: korelační koeficient, lineární regrese

Doba potřebná ke studiu a zpracování úkolů: **4 hodiny**

Průvodce studiem

V předchozích dvou kapitolách byla analyzována jedna veličina (kategoriální nebo metrická) ve dvou výběrech nebo ve dvou situacích. Pokud se sleduje vztah dvou nebo více veličin, pak hovoříme o závislosti. Jedná se o značně náročné téma, ale v rámci cvičení Biostatistiky bude probírána závislost jen dvou metrických (spojitých) veličin. Pro výpočet budou použity funkce MS Excelu.
Vzhledem k tomu, že úkolem této kapitoly je jen studenta seznámit, jak se posuzuje závislost dvou metrických veličin. Náročnost této kapitoly je střední.
Teorie k uvedeným tématům je probírána v páté přednášce Lékařská biofyzika, výpočetní technika I – Biostatistika.

6.1 Pearsonův korelační koeficient

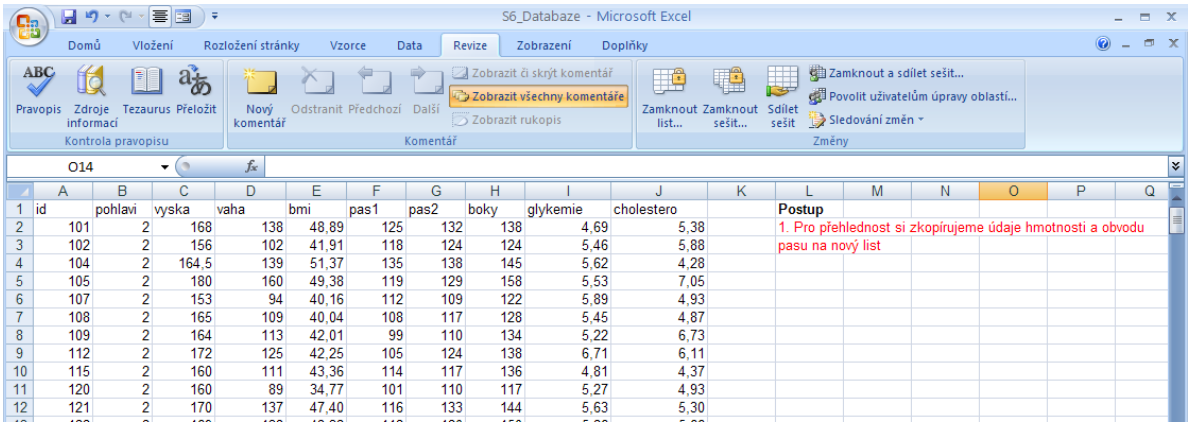
Korelační koeficient (Pearsonův - *r*) je nejpoužívanější mírou těsnosti vztahu dvou spojitých veličin.
Je mírou linearity vztahu (jak těsně body přimykají k přímce), a proto u výrazně nelineálních vztahů selhává.
Hodnota korelačního koeficientu je v rozmezí od -1 do 1. Hodnoty -1 nebo 1 nabývá tehdy, pokud všechny body leží na přímce. Nule je roven v případě, že neexistuje lineární vztah mezi sledovanými veličinami. Korelační koeficient však může být nulový i v případě, že veličiny jsou funkčně závislé, ale závislost není lineární. Proto je při užití Pearsonova korelačního koeficientu vždy nutné posoudit, zda je jeho aplikace vhodná. Při měření lineární závislosti je znaménko kladné, pokud obě veličiny *X* a *Y* zároveň rostou nebo obě

zároveň klesají (přímá úměra). Záporné znaménko znamená, že jedna veličina roste a druhá klesá (nepřímá úměra).
Vždy je nutné se zamyslet nad praktickým významem. Pro rozsáhlé výběry se bude i velmi malá hodnota korelačního koeficientu statisticky významně lišit od nuly. Je důležité si uvědomit, že korelace neznamená příčinnost.

Postup při zjišťování závislosti mezi dvěma znaky:
1. Sestavení bodového graf *XY*.
2. Výpočet korelačního koeficientu.
3. Posouzení znaménka korelačního koeficientu.
4. Podívat se na velikost korelačního koeficientu.
5. Provést interpretaci, ale pozor korelace neznamená příčinnost.

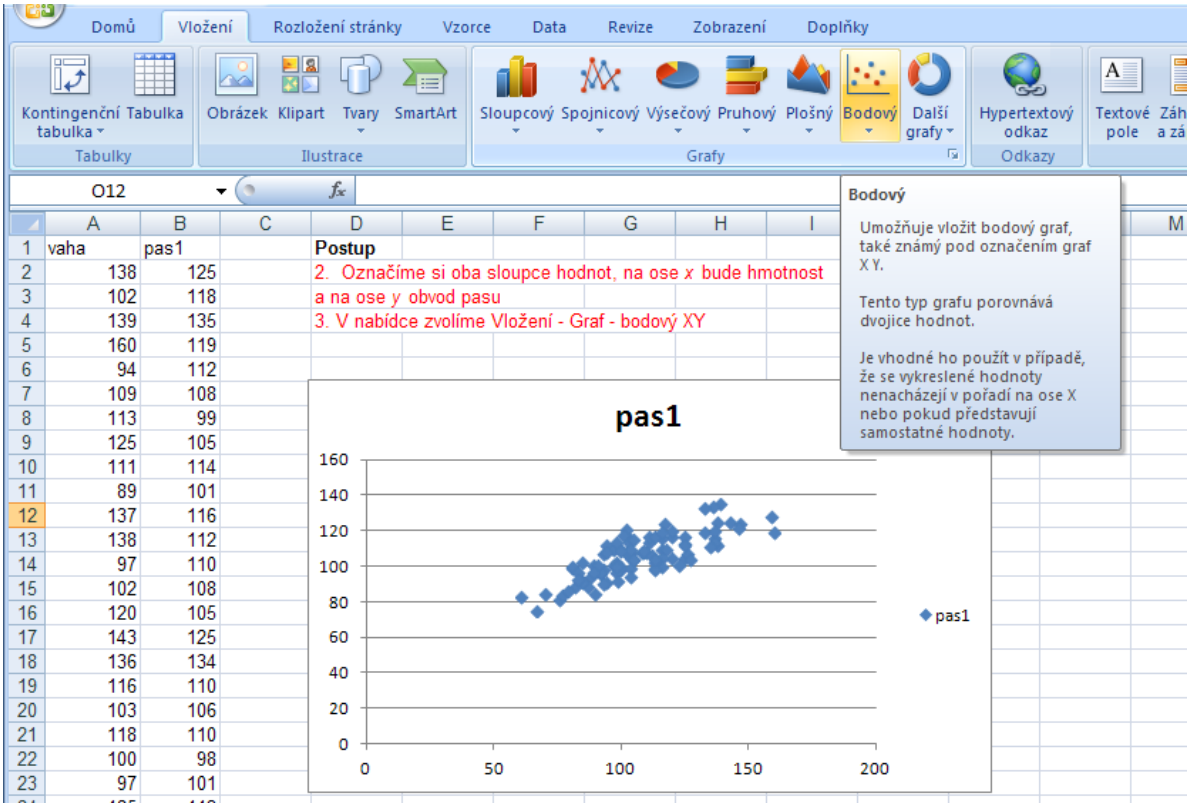
Příklad 6.1
Zjistěte, zda existuje závislost obvodu pasu (Pas1) na hmotnosti (Vaha) a jak silná je tato závislost. Použijte data z databáze S6 (Jedná se o data ze stejné studie jako v kapitole 5, ale jsou zde údaje žen za první a druhé vyšetření).
a) Sestavte bodový graf XY (X – hmotnost, Y – obvod pasu).
b) Vypočtěte korelační koeficient.
c) Výsledek interpretujte.

Řešení
Řešení je popsáno na obr. 6.1 – 6.4.



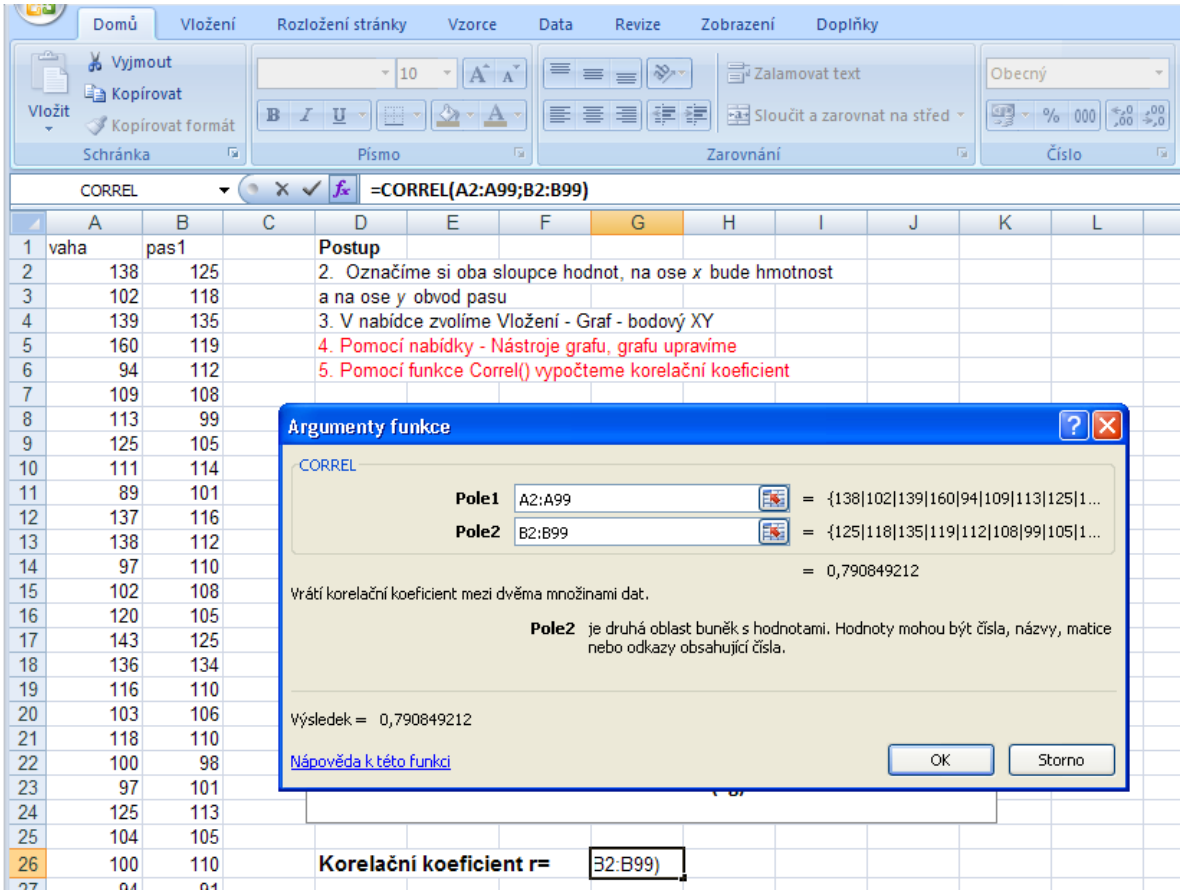
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1	id	pohlavi	vyska	vaha	bmi	pas1	pas2	boky	glykemie	cholestero		Postup					
2	101	2	168	138	48,89	125	132	138	4,69	5,38		1. Pro přehlednost si zkopírujeme údaje hmotnosti a obvodu					
3	102	2	156	102	41,91	118	124	124	5,46	5,88		pasu na nový list					
4	104	2	164,5	139	51,37	135	138	145	5,62	4,28							
5	105	2	180	160	49,38	119	129	158	5,53	7,05							
6	107	2	153	94	40,16	112	109	122	5,89	4,93							
7	108	2	165	109	40,04	108	117	128	5,45	4,87							
8	109	2	164	113	42,01	99	110	134	5,22	6,73							
9	112	2	172	125	42,25	105	124	138	6,71	6,11							
10	115	2	160	111	43,36	114	117	136	4,81	4,37							
11	120	2	160	89	34,77	101	110	117	5,27	4,93							
12	121	2	170	137	47,40	116	133	144	5,63	5,30							

Obr. 6.1 Krok 1a) Příprava pro sestavní grafu



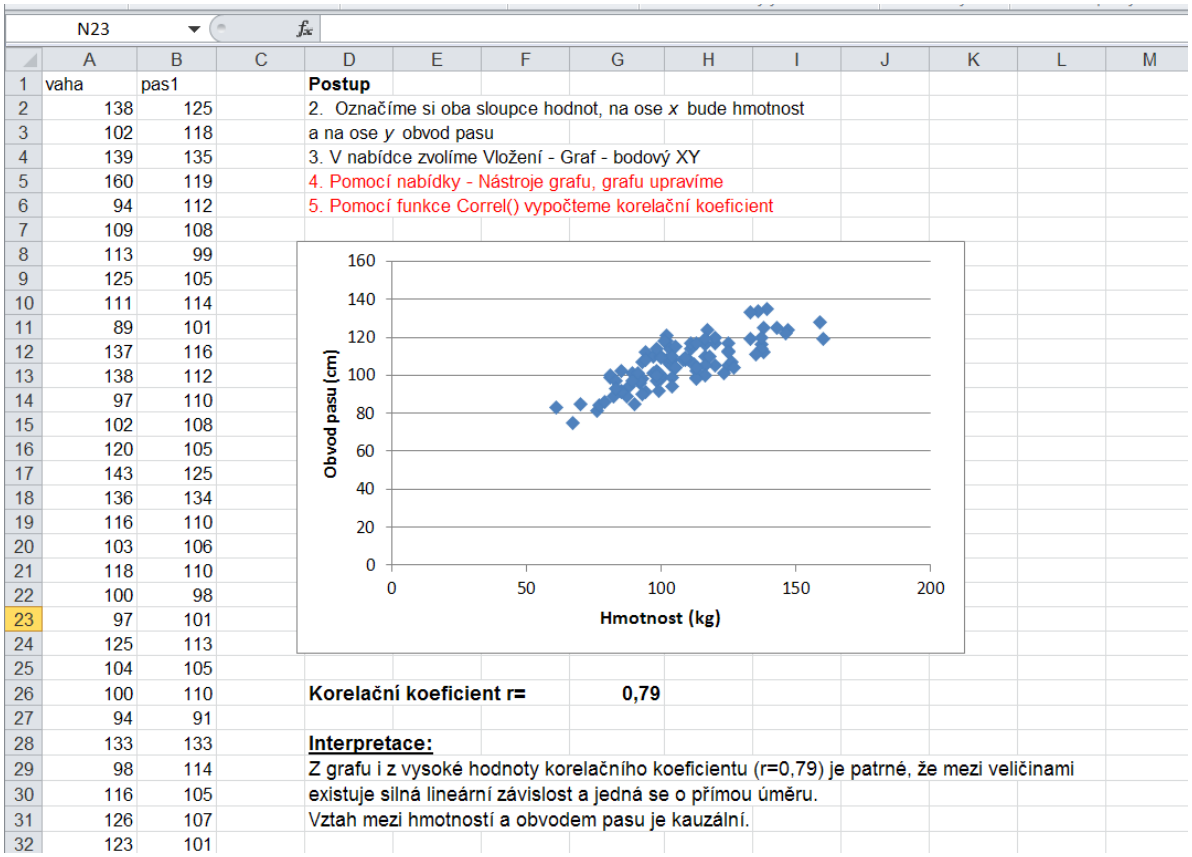
	A	B	C	D
1	vaha	pas1		Postup
2	138	125		2. Označíme si oba sloupce hodnot, na ose *x* bude hmotnost
3	102	118		a na ose *y* obvod pasu
4	139	135		3. V nabídce zvolíme Vložení - Graf - bodový XY
5	160	119		
6	94	112		
7	109	108		
8	113	99		
9	125	105		
10	111	114		
11	89	101		
12	137	116		
13	138	112		
14	97	110		
15	102	108		
16	120	105		
17	143	125		
18	136	134		
19	116	110		
20	103	106		
21	118	110		
22	100	98		
23	97	101		

Obr. 6.2 Krok 2a) Sestavení grafu (ve vytvořeném grafu je nutné upravit popis os obr. 6.4)



	A	B	C	D	E	F	G	H	I	J	K	L
1	vaha	pas1		Postup								
2	138	125		2. Označíme si oba sloupce hodnot, na ose *x* bude hmotnost								
3	102	118		a na ose *y* obvod pasu								
4	139	135		3. V nabídce zvolíme Vložení - Graf - bodový XY								
5	160	119		4. Pomocí nabídky - Nástroje grafu, grafu upravíme								
6	94	112		5. Pomocí funkce Correl() vypočteme korelační koeficient								
7	109	108										
8	113	99										
9	125	105										
10	111	114										
11	89	101										
12	137	116										
13	138	112										
14	97	110										
15	102	108										
16	120	105										
17	143	125										
18	136	134										
19	116	110										
20	103	106										
21	118	110										
22	100	98										
23	97	101										
24	125	113										
25	104	105										
26	100	110		Korelační koeficient r=			B2:B99)					

Obr. 6.3 Krok 1b) Výpočet korelačního koeficientu



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	vaha	pas1											
2	138	125											
3	102	118											
4	139	135											
5	160	119											
6	94	112											
7	109	108											
8	113	99											
9	125	105											
10	111	114											
11	89	101											
12	137	116											
13	138	112											
14	97	110											
15	102	108											
16	120	105											
17	143	125											
18	136	134											
19	116	110											
20	103	106											
21	118	110											
22	100	98											
23	97	101											
24	125	113											
25	104	105											
26	100	110											
27	94	91											
28	133	133											
29	98	114											
30	116	105											
31	126	107											
32	123	101											

Obr. 6.4 Krok 1c) Interpretace výsledku

6.2 Lineární regrese

Máme dvě veličiny *X* a *Y*, přičemž veličina *X* představuje snadno dostupná měření (nezávisle proměnná) a *Y* je měřitelná jen s obtížemi (závisle proměnná). Pokud mezi veličinami *X* a *Y* existuje lineární závislost, je možné na základě principu nejmenších čtverců (přednáška 5) tento vztah vyjádřit lineární regresní funkcí:

$y = a+bx$,

kde *b* je regresní koeficient, *a* je absolutní člen, *x* označujeme jako nezávislou proměnnou a *y* je závisle proměnná.

Základní otázkou v regresní analýze je, zda hodnoty *X* opravdu ovlivňují hodnoty *Y*. Koeficient *b* představuje změnu *Y* při změně *X* o jedničku. Ptáme se tedy, zda se *b* významně liší od nuly. Nebo přesněji, zda zamítáme hypotézu H_0: $\beta = 0$, kde β je koeficient teoretické regresní přímky. Regresní koeficient i absolutní člen by měl být uváděn s intervaly spolehlivosti, na jejichž základě můžeme rozhodnout o zamítnutí H_0. Interval spolehlivosti je interval, ve kterém bude hodnota sledovaného parametru ležet s určitou pravděpodobností (nejčastěji se volí 95 % na základě alfa).

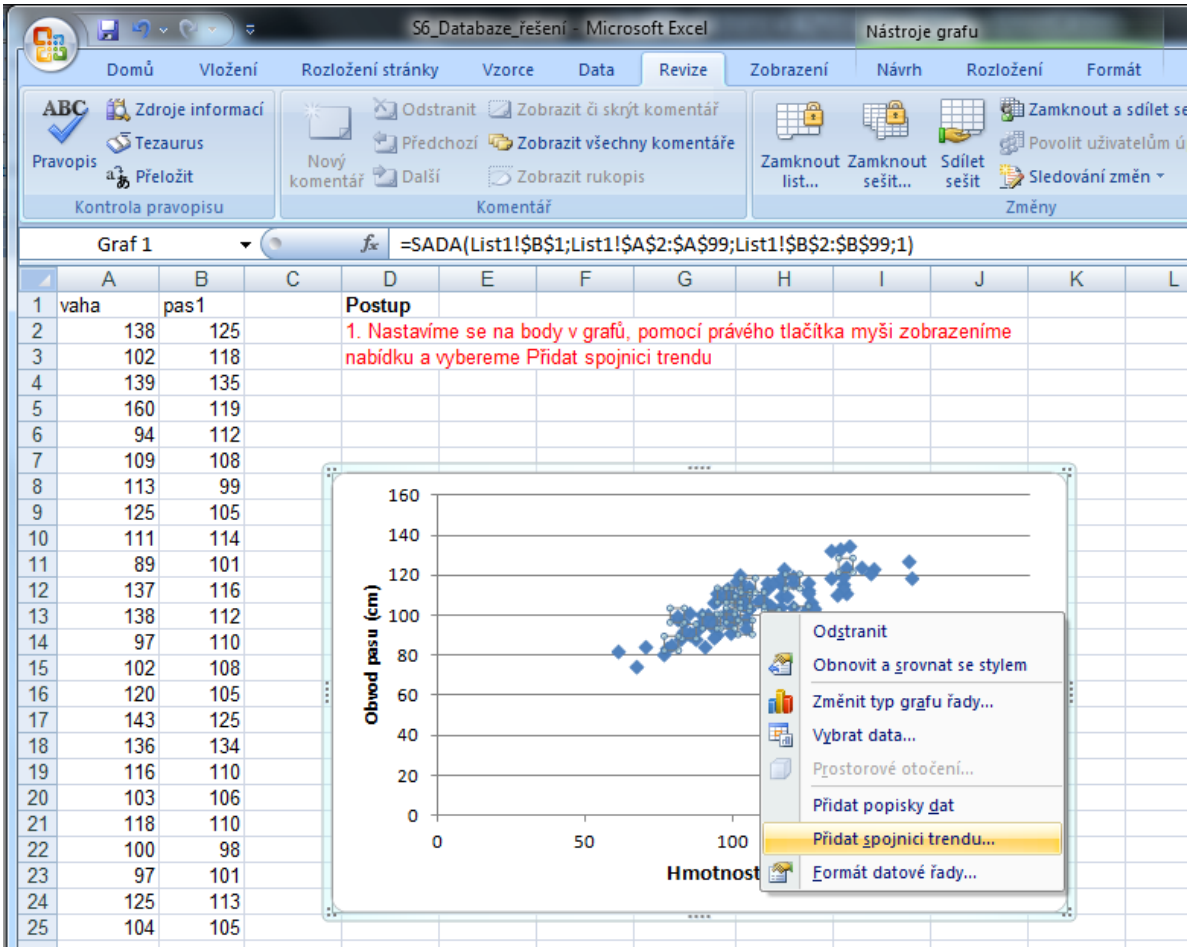
Příklad 6.2

Pokračujte v příkladu 6.1. Pokud jste zjistili, že existuje závislost mezi obvodem pasu (Pas1) na hmotností (Vaha), vyjádřete tento vztah pomocí lineární regresní funkce

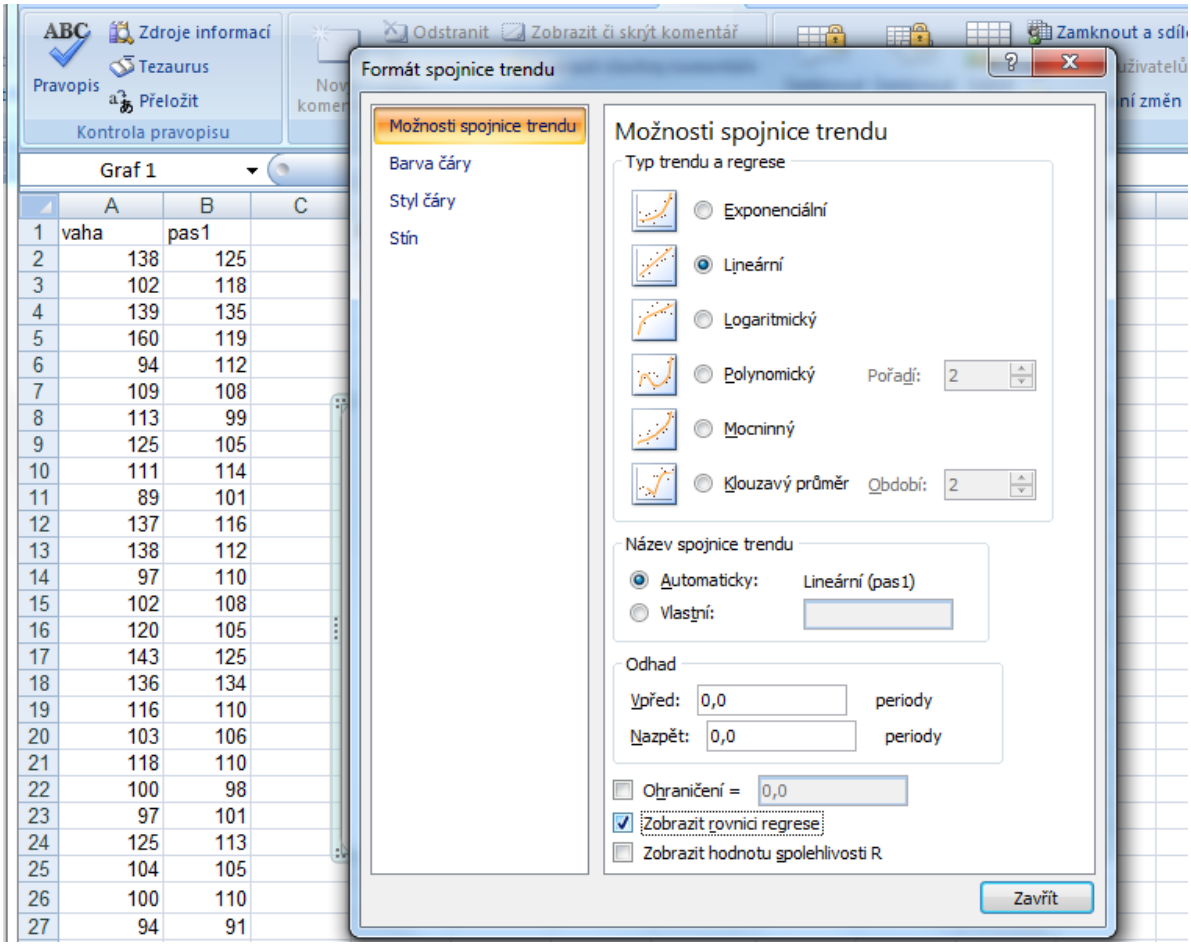
a) Do grafu doplňte spojnici trendu.
b) Vypočtěte parametry rovnice (*a*, *b*).
c) Výsledek interpretujte.

Řešení

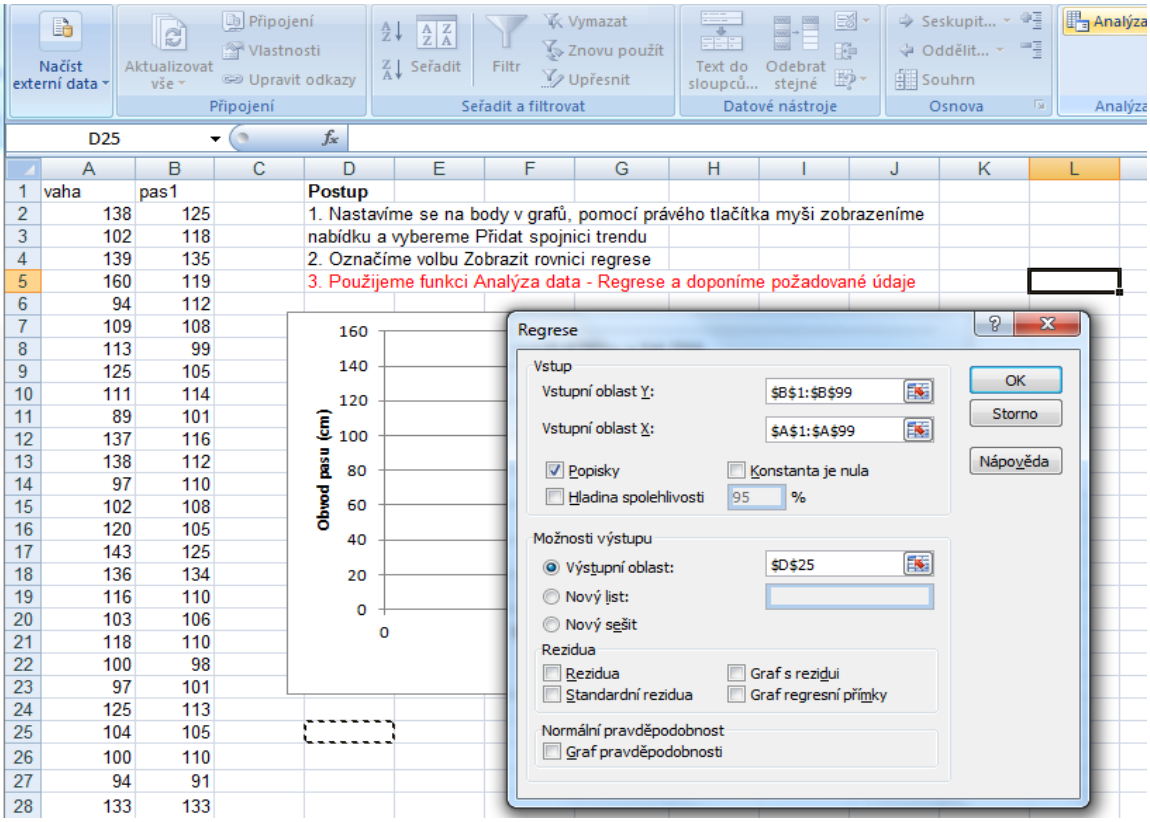
Budeme pokračovat v příkladu 6.1, do grafu doplníme spojnici trendu i doplněním rovnic křivky (obr. 6.5 - 6.8).



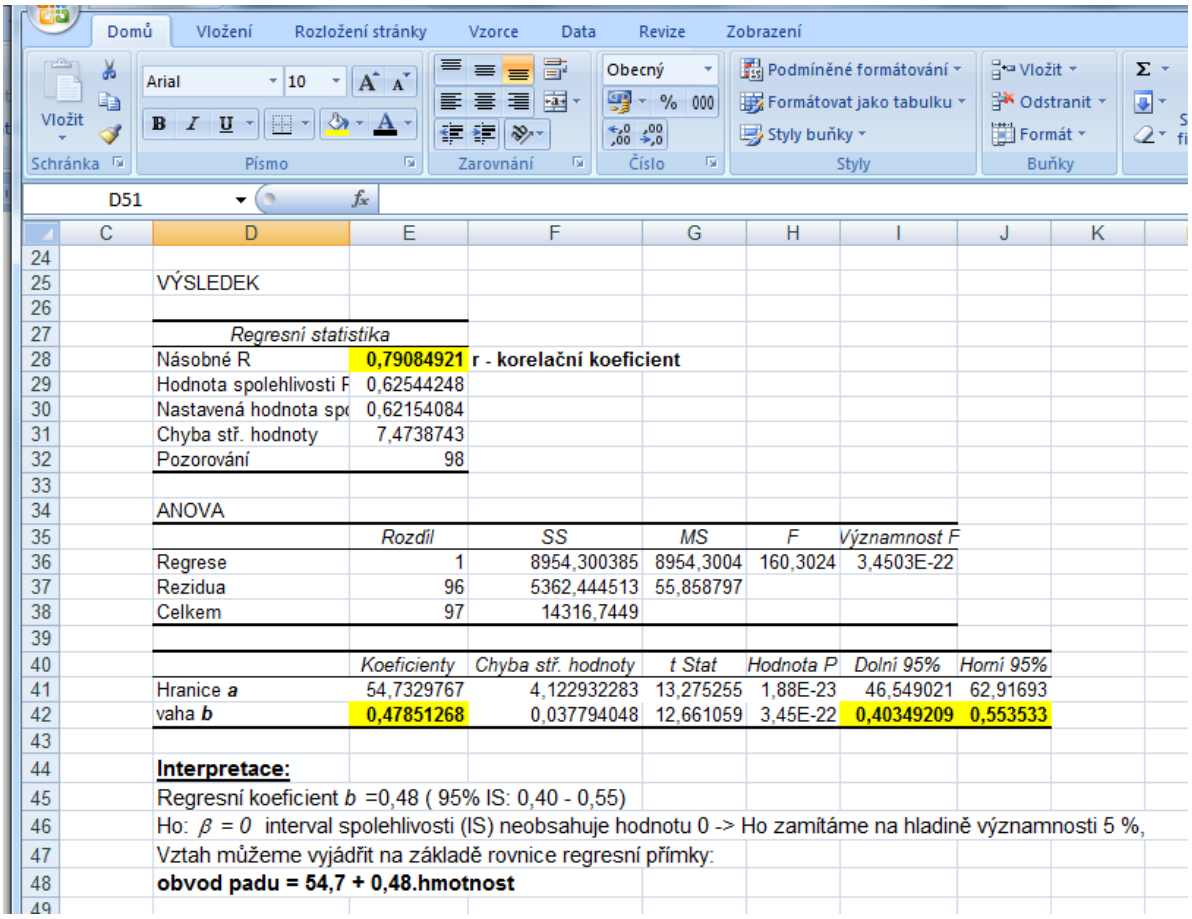
Obr. 6.5 Krok 1a) Přidání spojnice trendu



Obr. 6.6 Krok 2a) Zobrazení rovnice regrese



Obr. 6.7 Krok 1b) Výpočet parametrů regresní přímky



	C	D	E	F	G	H	I	J	K
24									
25		VÝSLEDEK							
26									
27		*Regresní statistika*							
28		Násobné R	**0,79084921**	**r - korelační koeficient**					
29		Hodnota spolehlivosti F	0,62544248						
30		Nastavená hodnota spo	0,62154084						
31		Chyba stř. hodnoty	7,4738743						
32		Pozorování	98						
33									
34		ANOVA							
35			*Rozdíl*	*SS*	*MS*	*F*	*Významnost F*		
36		Regrese	1	8954,300385	8954,3004	160,3024	3,4503E-22		
37		Rezidua	96	5362,444513	55,858797				
38		Celkem	97	14316,7449					
39									
40			*Koeficienty*	*Chyba stř. hodnoty*	*t Stat*	*Hodnota P*	*Dolní 95%*	*Horní 95%*	
41		Hranice ***a***	54,7329767	4,122932283	13,275255	1,88E-23	46,549021	62,91693	
42		vaha ***b***	**0,47851268**	0,037794048	12,661059	3,45E-22	**0,40349209**	**0,553533**	
43									
44		**Interpretace:**							
45		Regresní koeficient *b* =0,48 (95% IS: 0,40 - 0,55)							
46		Ho: $\beta = 0$ interval spolehlivosti (IS) neobsahuje hodnotu 0 -> Ho zamítáme na hladině významnosti 5 %,							
47		Vztah můžeme vyjádřit na základě rovnice regresní přímky:							
48		**obvod padu = 54,7 + 0,48.hmotnost**							
49									

Obr. 6.8 Krok 1c)Interpretace

Shrnutí obsahu kapitoly

Pokud sledujeme vztah dvou metrických veličin, tak hovoříme o závislosti. Lineární závislost mezi dvěma metrickými (spojitými) veličinami zjišťujeme pomocí Pearsonova korelačního koeficientu. Pokud mezi veličinami existuje lineární závislost, můžeme na základě znalosti jedné veličiny odhadnout druhou veličinu pomocí lineární regresní funkce.

MS Excel – bodový graf XY a spojnice trendu, statistická funkce Correl(), Analýza data – Regrese

Kontrolní otázky:

1. Co zjišťuje korelační koeficient?
2. Jakých hodnot může korelační koeficient nabývat?
3. Co je výsledkem regresní analýzy?
4. Jak budete interpretovat vztah mezi dvěma veličinami, pokud korelační koeficient bude r = -0,85?

Úkol

6.1 Pokračujte v analýze údajů v databázi S6_Databáze:

a) Zjistěte, zda existuje závislost obvodu Pasu (Pas2) na hmotnosti (Vaha)
b) Která závislosti je silnější – Pas1 a Vaha nebo Pas2 a Vaha?
c) Vyjádřete zjištěnou závislost pomocí lineární regresní funkce.
d) Výsledky interpretujte

6.2 Pokračujte v analýze údajů v databázi S6_Databáze:

a) Zjistěte, zda existuje závislost obvodu boků (Boky) na hmotnosti (Vaha).
b) Vyjádřete zjištěnou závislosti pomocí lineární regresní funkce.
c) Výsledky interpretujte.

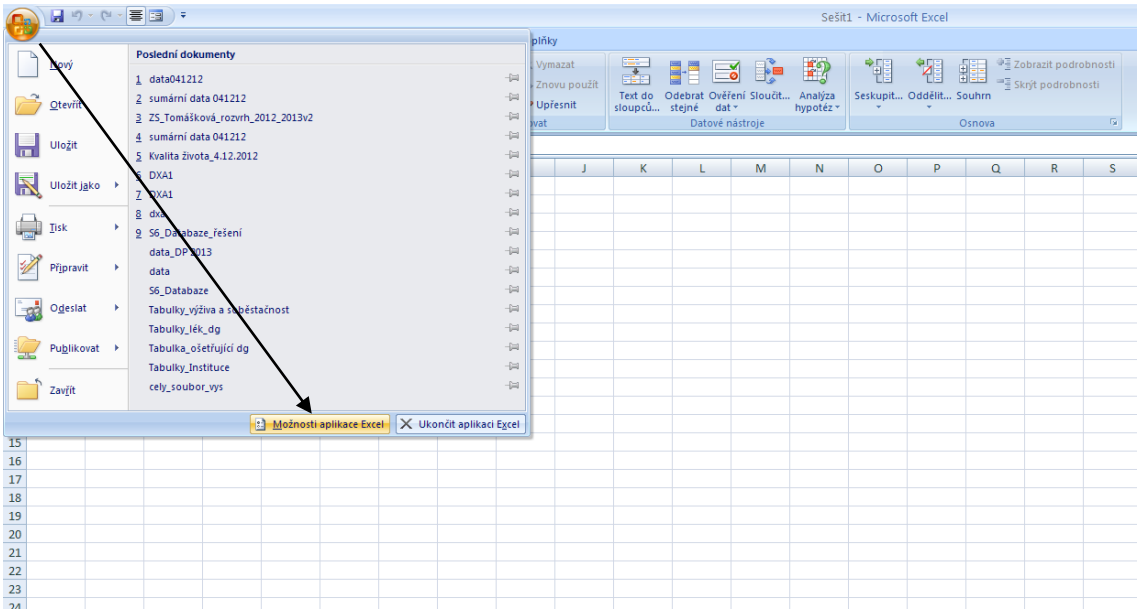
6.3 Pokračujte v analýze údajů v databázi S6_Databáze:

a) Existuje přímá závislost hodnot glykemie a cholesterolu?

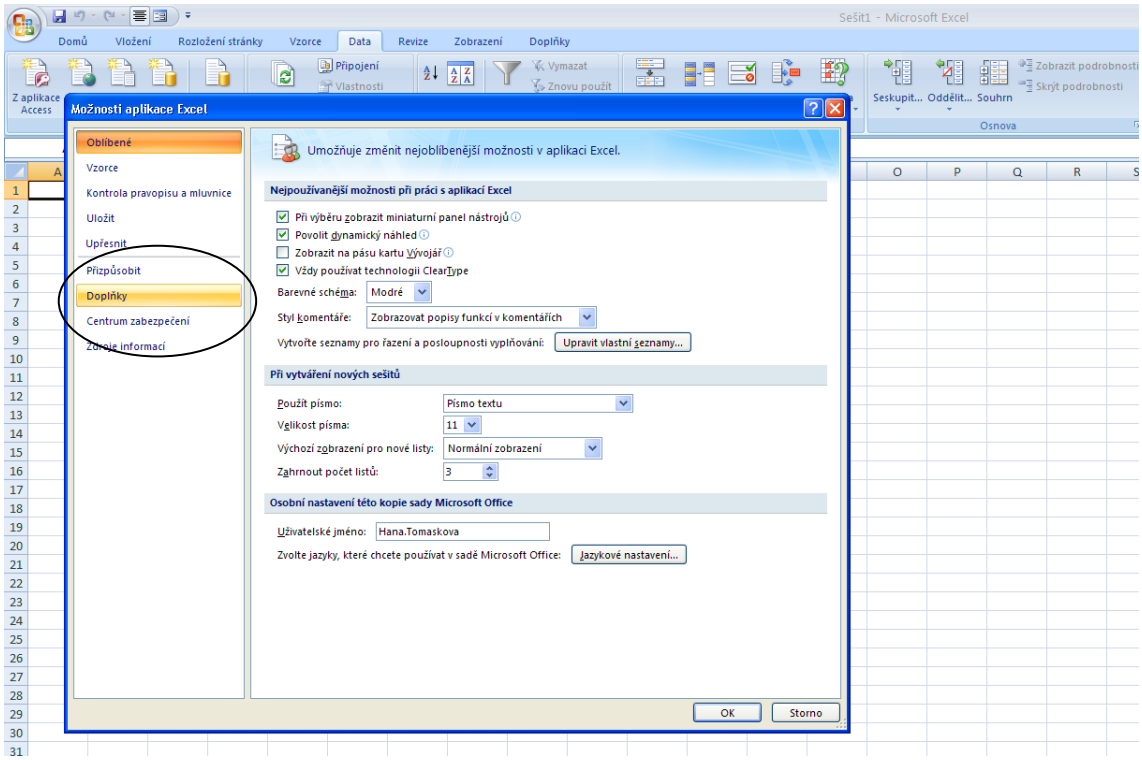
Příloha

Příloha 1 – Instalace funkce Analýza dat

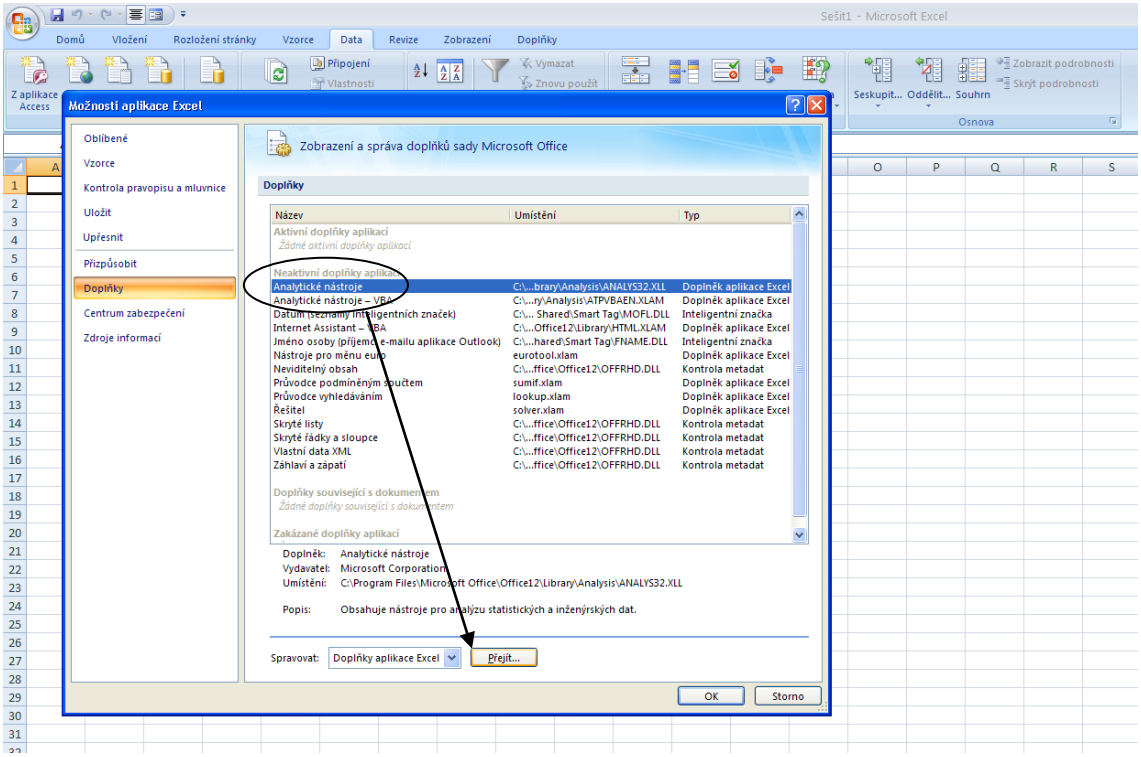
Instalace dolňku Analýza dat v MS Excelu 2007 je znázorněna na obr. 1 – 5.



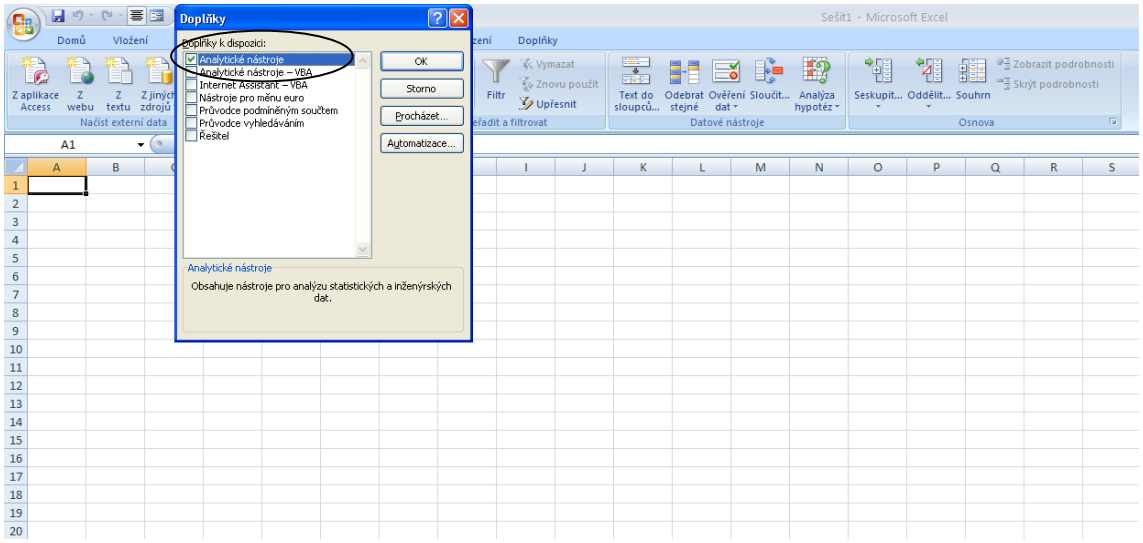
Obr. 1 Zvolíme Tlačítko Office a vybereme Možnosti aplikace Excel



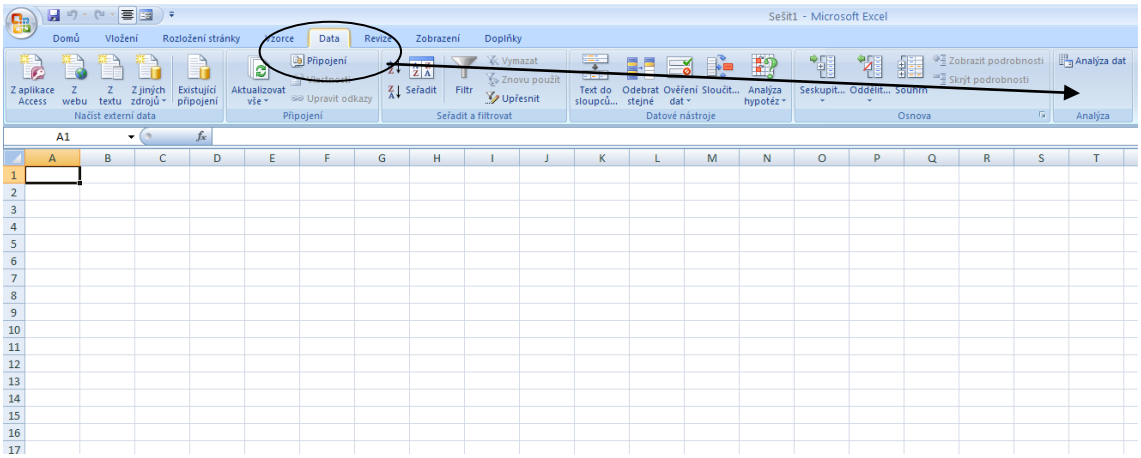
Obr. 2 Zvolíme Doplňky



Obr. 3 Zvolíme Analytické nástroje a Přejít



Obr. 4 Zaškrtneme Analytické nástroje a OK



Obr. 5 Analýza dat je pod nabídkou Data