Biostatistika

Kontingenční tabulky v Excelu
Základní popisné statistiky
Představení programu Statistica
Import a základní popis dat ve Statistice

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

I. Kontingenční tabulky v Excelu

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

Kontingenční tabulka

- Frekvenční sumarizace dvou kategoriálních proměnných (binárních, nominálních nebo ordinálních proměnných).
- Obecně: **R x C kontingenční tabulka** (R – počet kategorií jedné proměnné, C – počet kategorií druhé proměnné).
- Speciální případ: 2 x 2 tabulka = čtyřpolní tabulka.
- Kontingenční tabulky: **absolutních četností, celkových procent, řádkových/sloupcových četností**

- Př.: Sumarizace vyšetřených osob podle pohlaví a výsledku diagnostického testu.

	Výsledek vyšetření		
Pohlaví	Nemocný	Zdravý	**Celkem**
Muž	45	11	56
Žena	25	6	31
Celkem	70	17	87

Jsou více nemocní muži nebo ženy?

Ukázka kontingenční tabulky

Kontingenční tabulka absolutních četností

	Nemocný	Zdravý	Celkem
Muž	45	11	56
Žena	25	6	31
Celkem	70	17	87

Větší počet nemocných mužů, který je dán pouze vyšším zastoupení mužů v celkovém vzorku (56 z 87)

Kontingenční tabulka řádkových procent

	Nemocný	Zdravý	Celkem
Muž	80,4 %	19,6 %	100,0 %
Žena	80,6 %	19,4 %	100,0 %

Po výpočtu relativních četností vidíme, že se muži a ženy neliší ve výskytu onemocnění.

Jsou více nemocní muži nebo ženy?

Kontingenční tabulky v Excelu: zdroj dat a příprava dat

Kontingenční tabulka se dá vytvořit:

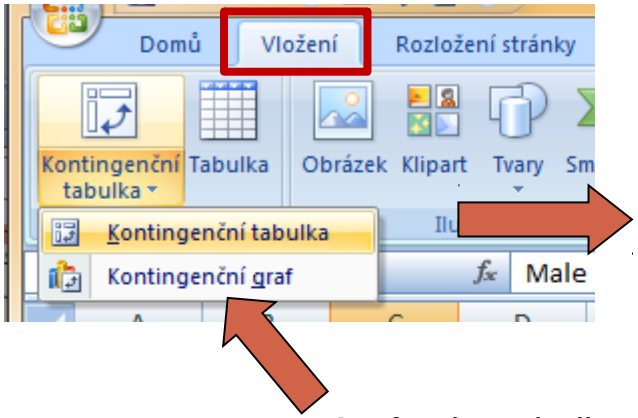
1. z tabulky v daném sešitě
2. z dat z jiného sešitu Excelu
3. z externích dat (např. MS Access)
4. ze sloučených dat z více oblastí - z různých listů nebo různých sešitů
5. z jiné kontingenční tabulky

Data musí být uspořádána formou standardního databázového seznamu:

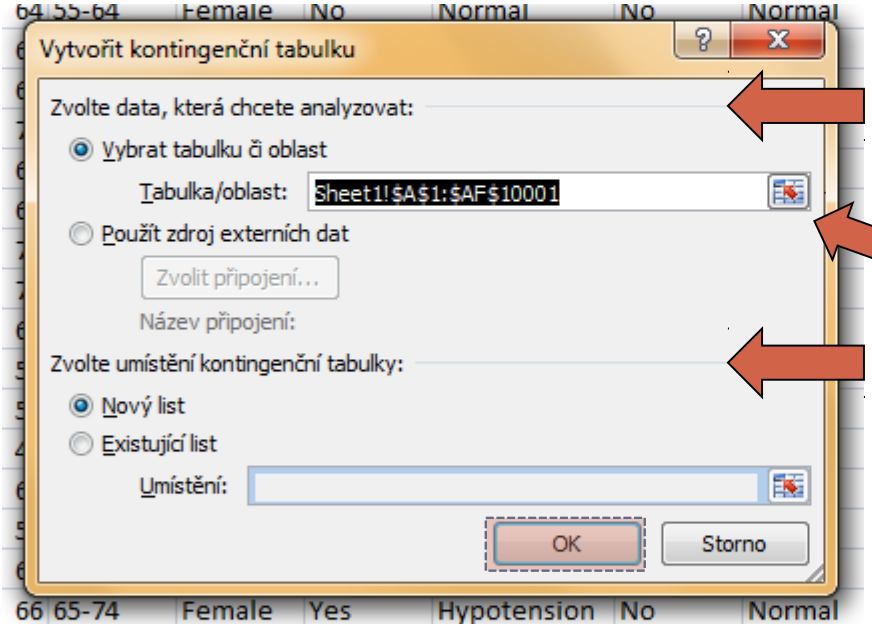
- V prvním řádku: názvy polí
- Další řádky: data

Vzhled tabulky: karta **Domů** → **Formátovat jako tabulku**

Vytvoření kontingenční tabulky v Excelu



Graf nebo tabulka

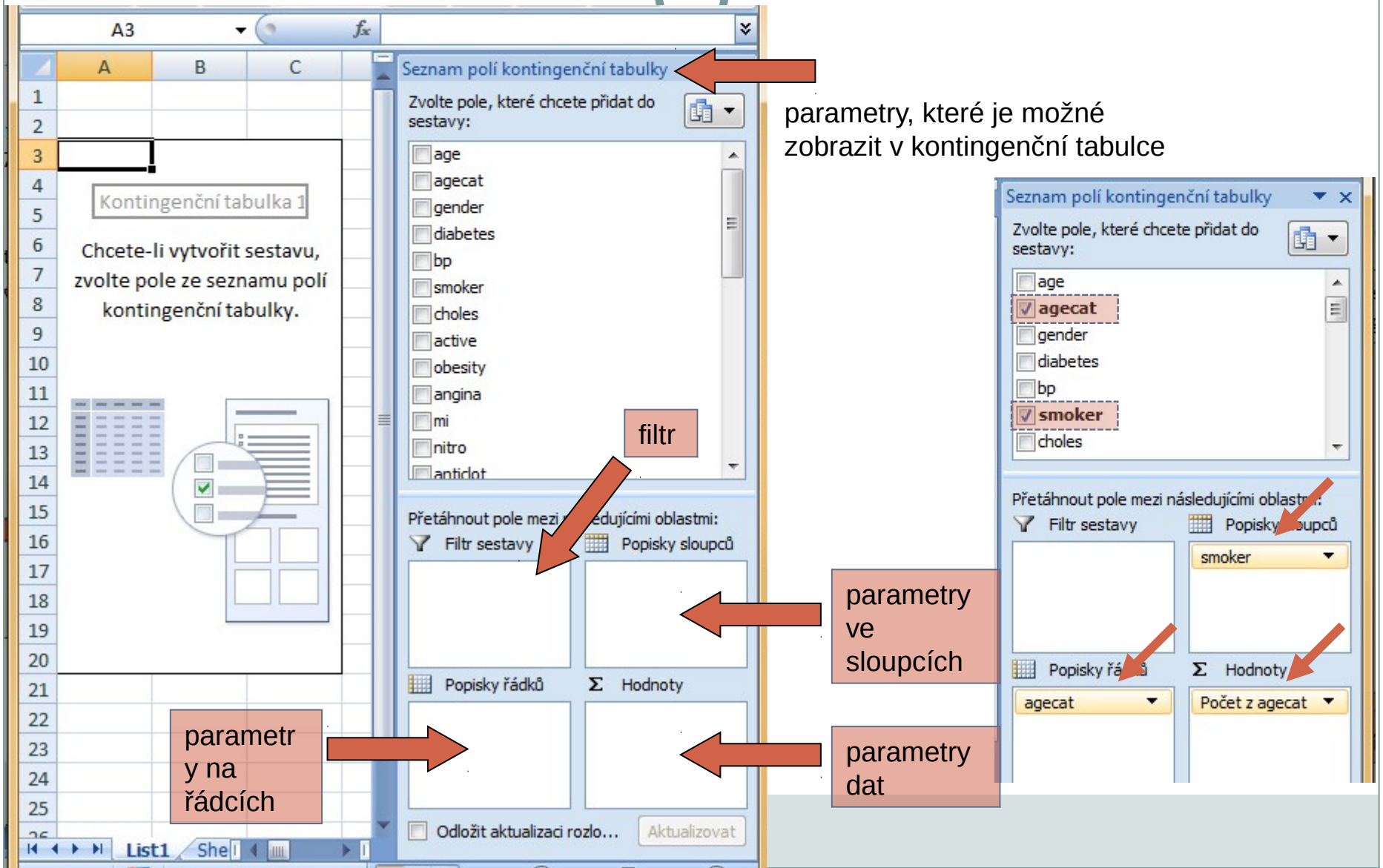


Zdroj dat (kromě Excelu i např. externí databáze)

Zdrojová oblast dat

Umístění tabulky

Kontingenční tabulky – rozvržení

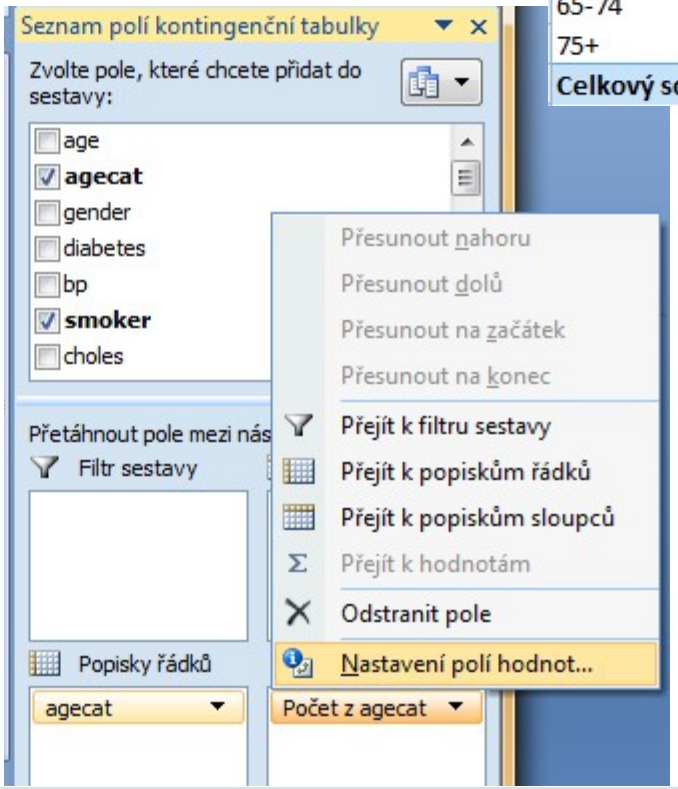


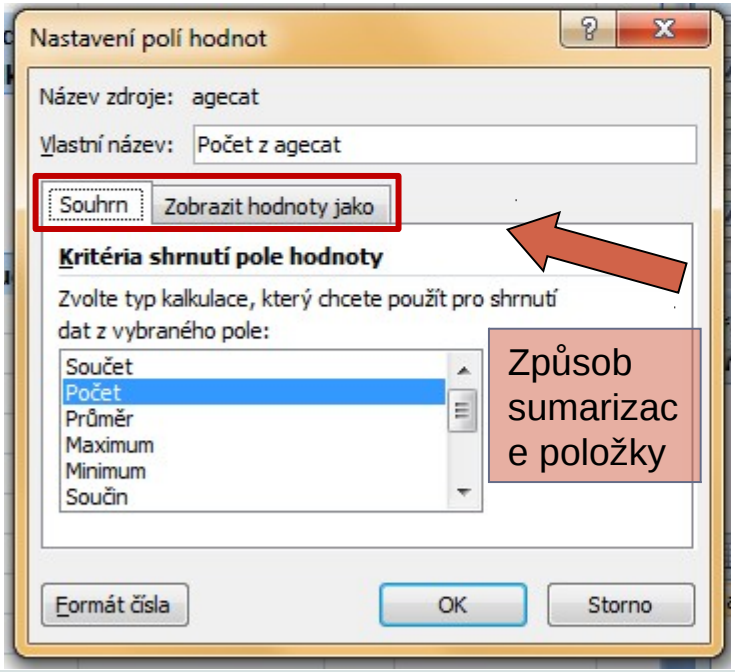
parametry, které je možné zobrazit v kontingenční tabulce

Kontingenční tabulky – nastavení II.

Kontingenční tabulka

Počet z agecat			
Popisky řádků	No	Yes	Celkový součet
45-54	1694	501	2195
55-64	3015	863	3878
65-74	2200	661	2861
75+	816	250	1066
Celkový součet	**7725**	**2275**	**10000**





Způsob sumarizace položky

Aktualizace dat v kontingenční tabulce

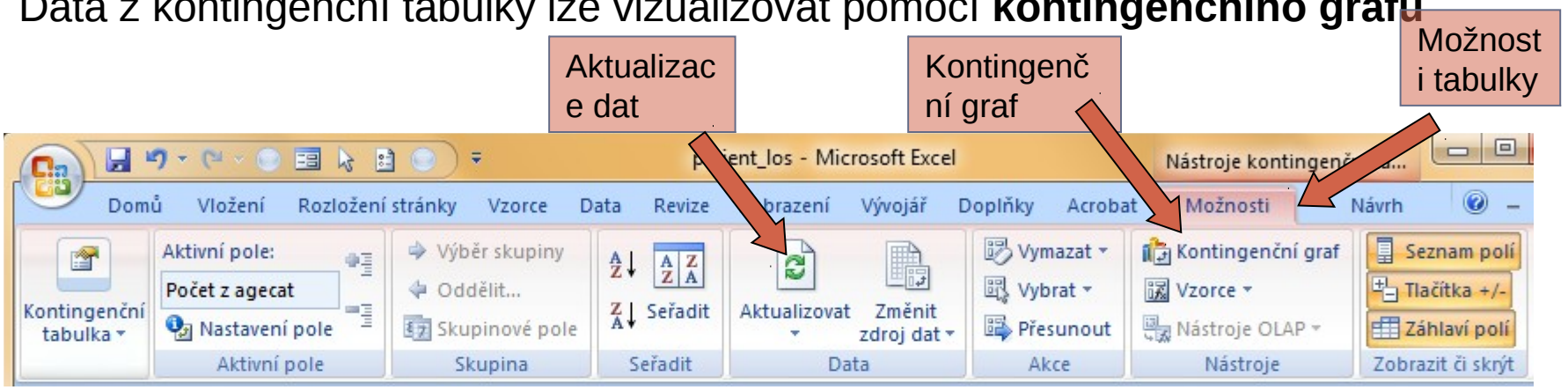
Při změně dat v tabulce se zdrojovými daty **nedojde** automaticky k aktualizaci dat

v kontingenční tabulce.

Musíte provést aktualizaci dat.

1. Stůjte kdekoliv v kontingenční tabulce
2. Na kartě **Možnosti** ve skupině **Data** klikněte na **Aktualizovat** (Alt+F5), nebo na **Aktualizovat vše** (Ctrl+Alt+F5)

Data z kontingenční tabulky lze vizualizovat pomocí **kontingenčního grafu**



Rozložení kontingenční tabulky

Po vytvoření se kontingenční tabulka zobrazí v tzv. **kompaktním formátu**. Lze ji zobrazit

ale i ve formě **tabulky**, nebo ve formě **osnovy**.

1. Stůjte kdekoliv v kontingenční tabulce
2. Na kartě **Návrh** vyberte tlačítko **Rozložení sestavy** a volbu **Zobrazit ve formě osnovy nebo zobrazit ve formě tabulky**

Kompaktní formát - uspořádání tabulky aby zabírala co nejméně místa

Forma osnovy - řádková pole nižší úrovně je od vyšších úrovní odsazena, řádky nejsou

odděleny čarami

Forma tabulky - klasická forma tabulky, pole nižší úrovně jsou v dalším sloupci

Vyzkoušej!

II. Základy popisné statistiky

Kvalitativní (kategoriální) proměnná

lze ji řadit do kategorií, ale nelze ji kvantifikovat

Příklad: ??

Kvantitativní (numerická) proměnná

můžeme ji přiřadit číselnou hodnotu

Příklad: ??

Kvalitativní (kategoriální) proměnná
lze ji řadit do kategorií, ale nelze ji kvantifikovat
Příklady: *pohlaví, HIV status, barva vlasů ...*

Kvantitativní (numerická) proměnná
můžeme ji přiřadit číselnou hodnotu
Příklady: *výška, váha, teplota, počet hospitalizací ...*

Intervalové znaky: interpretace rozdílu dvou hodnot (stejný interval mezi jednou a druhou dvojicí hodnot vyjadřuje i stejný rozdíl v intenzitě zkoumané vlastnosti). Společný znak intervalových znaků: nula byla stanovena uměle, tedy pouhou konvencí. *Příklad: teplota měřená ve stupních Celsia, letopočet.*

Den	Teplota	Rozdíl 1	Podíl 1
1.	2 °C	-	-
2.	4 °C	+2	2x
3.	6 °C	+2	1.5x

1 Srovnání s měřením z předchozího dne

← 1.5krát vyšší teplota ve srovnání s 2. dnem, přičemž došlo ke stejnému nárůstu teploty jako při srovnání 2. a 1. dne

Poměrové znaky: kromě rozdílu interpretujeme i podíl dvou hodnot.
Příklady: výška v cm, váha v kg, ...

Popisné statistiky

Charakteristiky polohy (míry střední hodnoty, míry centrální tendence)

- Udávají, kolem jaké hodnoty se data centrují, resp. které hodnoty jsou nejčastější, popis „těžiště“ – míry polohy
- Aritmetický průměr, medián, modus, geometrický průměr

Charakteristiky variability (proměnlivosti)

- Zachycují rozptýlení hodnot v souboru (proměnlivost dat)
- Variační rozpětí, rozptyl, směrodatná odchylka, variační koeficient, střední chyba průměru

Charakteristiky polohy

Charakteristiky polohy u nominálních znaků

- **Modus**: nejčastěji se vyskytující hodnota proměnné v souboru.

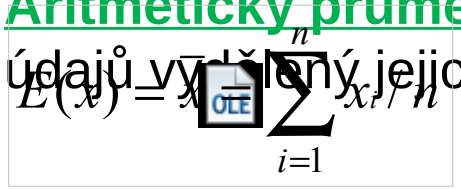
Charakteristiky polohy u ordinálních znaků

- **α-kvantil**: je-li α Є (0,1), pak α-kvantil xα je číslo, které rozděluje uspořádaný datový soubor na dolní úsek, obsahující aspoň podíl α všech dat a na horní úsek obsahující aspoň podíl 1-α všech dat.
- x0,50- **medián**, x0,25- **dolní kvartil**, x0,75-**horní kvartil**, x0,1…. x0,9-**decily**
- **Medián**: hodnota, jež dělí řadu podle velikosti seřazených výsledků na dvě stejně početné poloviny.

Charakteristiky polohy

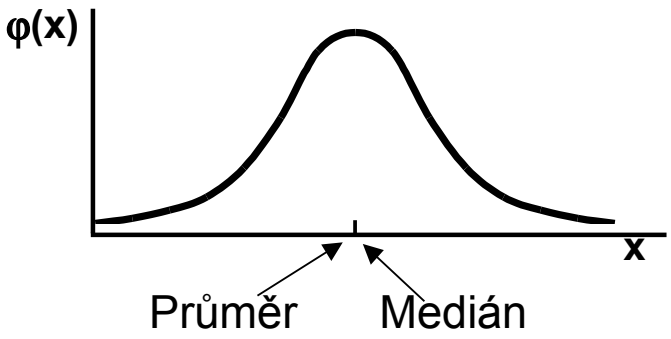
Charakteristiky polohy u intervalových a poměrových znaků

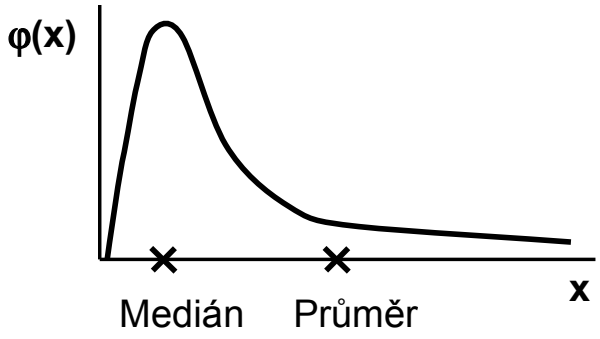
- **Aritmetický průměr**: je definován jako součet všech naměřených údajů vydělený jejich počtem,



kde *xi* jsou jednotlivé hodnoty a *n* jejich počet

- **Geometrický průměr**: *n* kladných hodnot *xi*,  , má smysl všude, kde má nějaký informační smysl součin hodnot proměnné. Z praktického hlediska platí, že logaritmus geometrického průměru je roven aritmetickému průměru logaritmovaných hodnot souboru.


φ(x)
x
Průměr
Medián


φ(x)
x
Medián
Průměr

Charakteristiky variability u ordinálních znaků

- **Kvartilové rozpětí (odchylka):** q = x0,75 - x0,25

Charakteristiky variability u intervalových a poměrových znaků

Rozptyl (variance) je ukazatelem šířky rozložení získaný na základě odchylky jednotlivých hodnot od průměru.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

Jeho vypovídací schopnost nejvyšší v případě symetrického/normálního rozložení

Směrodatná odchylka je druhá odmocnina z rozptylu

Koeficient variance - podíl SD ku průměru, u poměrových znaků, umožňuje porovnat variabilitu několika znaků (vyjadřuje se v %)

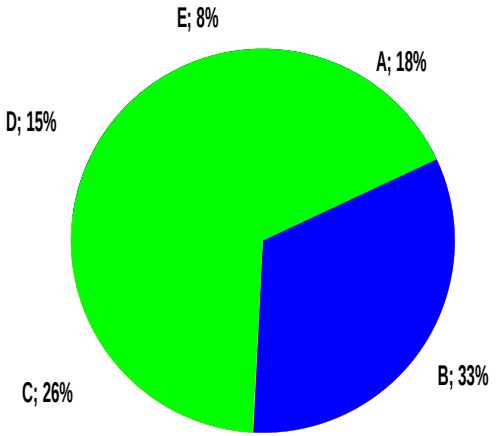
- **Popis kvalitativních dat:** frekvence jednotlivých kategorií
- **Vizualizace kvalitativních dat:** nejčastěji koláčový nebo sloupcový graf

Příklad: Známka z biostatistiky (podzim 2014)

Frekvenční tabulka

Známka	n	%
A	11	18,0
B	20	32,8
C	16	26,2
D	9	14,8
E	5	8,2
F	0	0,0
Celkem	61	100,0

Koláčový graf



Sloupcový graf

- **Popis kvantitativních dat:** charakteristika středu (průměr, medián aj.), charakteristika variability (rozptyl, rozsah hodnot, interkvartilové rozpětí aj.)

Příklad: Popis výšky (cm) pacientů

Popisné statistiky

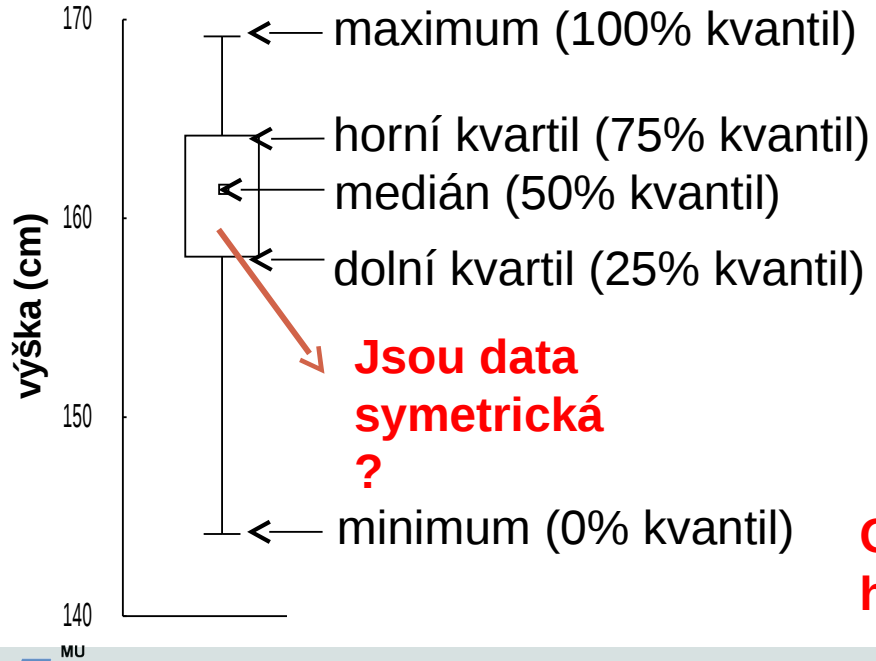
Charakteristika	
N	61
Průměr (cm)	161,0
Medián (cm)	161,5
Sm. odchylka (cm)	4,7
Rozptyl (cm2)	22,2
min-max (cm)	144 – 169
dolní-horní kvartil (cm)	158 - 164

Průměr a medián se téměř shodují. Co nám to říká?

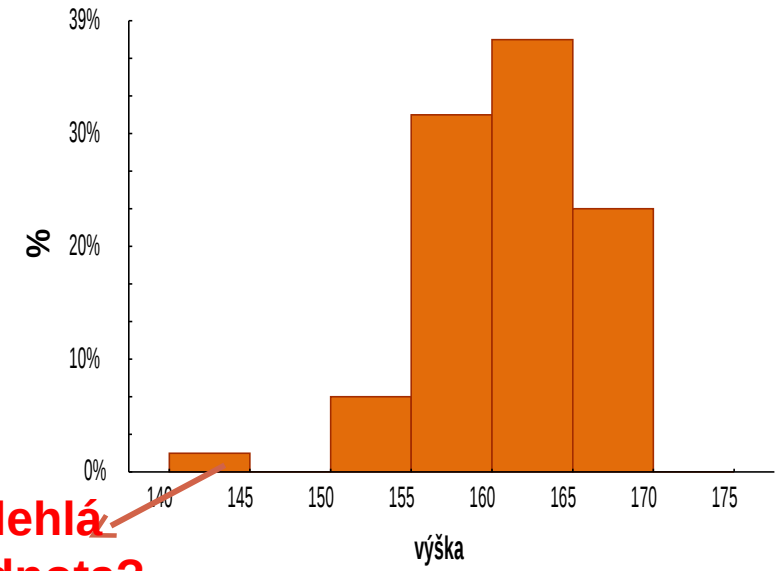
- **Vizualizace kvantitativních dat:** nejčastěji pomocí krabicového grafu nebo histogramu

Příklad: Popis výšky (cm) pacientů

Krabicový graf



Histogram



III. Cvičení v programu Statistica

Základní popisné statistiky v programu Statistica
Datový soubor pacienti.sta
Datový soubor studenti.sta

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

Program Statistica

Jak získat program Statistica:

https://inet.muni.cz

Login a heslo: UČO a primární heslo jako do IS-u.

V ponuke kliknout: **Provozní služby – Software – Nabídka softwaru**

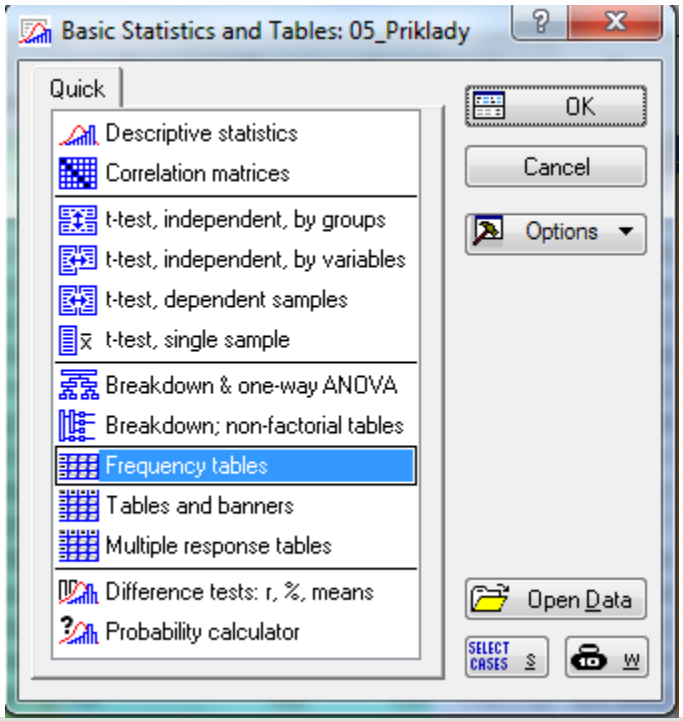
Nalézt: **Statistica 13** – kliknout **Získat**

Postupovat dle návodu

Základy popisné statistiky: soubor pacienti.sta

Načtěte soubor **pacienti.sta**, který obsahuje údaje o 61 pacientech.

- **Nejprve budeme pracovat s *kategoriální proměnnou*.**
- ***Pro proměnnou pohlaví zjistěte: absolutní, relativní četnost, dále absolutní a relativní kumulativní četnost***

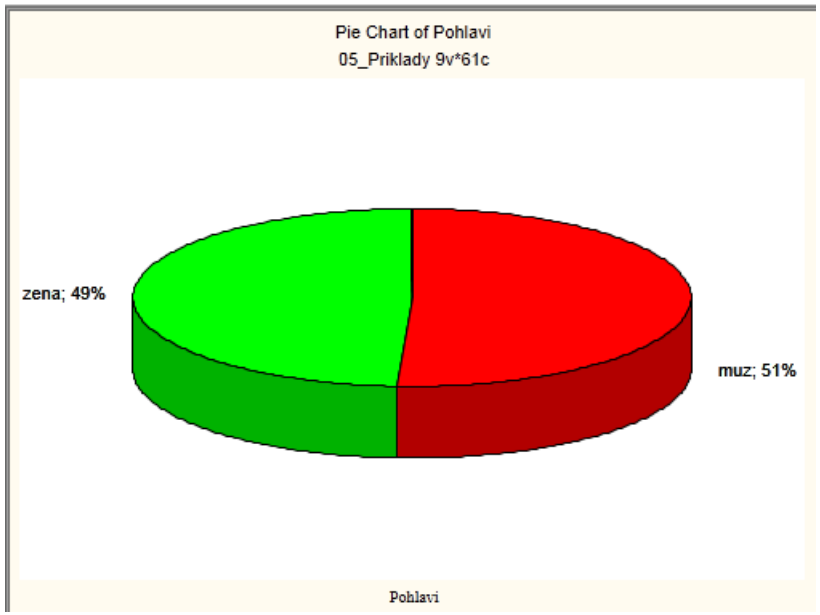


Frequency table: Pohlavi: =Round(Rnd(1);0) (05_F

Category	Count	Cumulative Count	Percent	Cumulative Percent
muz	31	31	50,81967	50,8197
zena	30	61	49,18033	100,0000
Missing	0	61	0,00000	100,0000

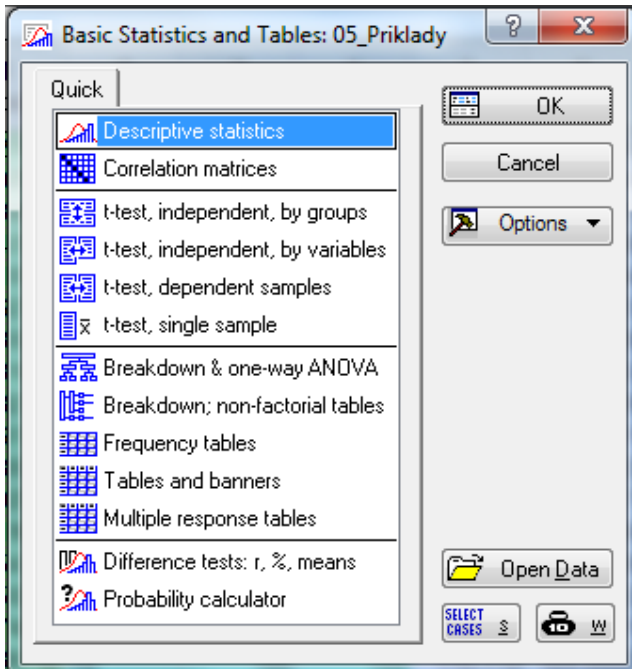
Základy popisné statistiky: soubor pacienti.sta

- ***Pomocí <u>výsečového grafu (koláčového grafu)</u> znázorněte proměnnou Pohlaví, doplňte procenta (relativní četnost).***



Základy popisné statistiky: soubor pacienti.sta

- ***Nyní budeme pracovat se spojitou proměnnou.***
- ***Pro proměnnou váha zjistěte: průměr, medián, minimum a maximum***

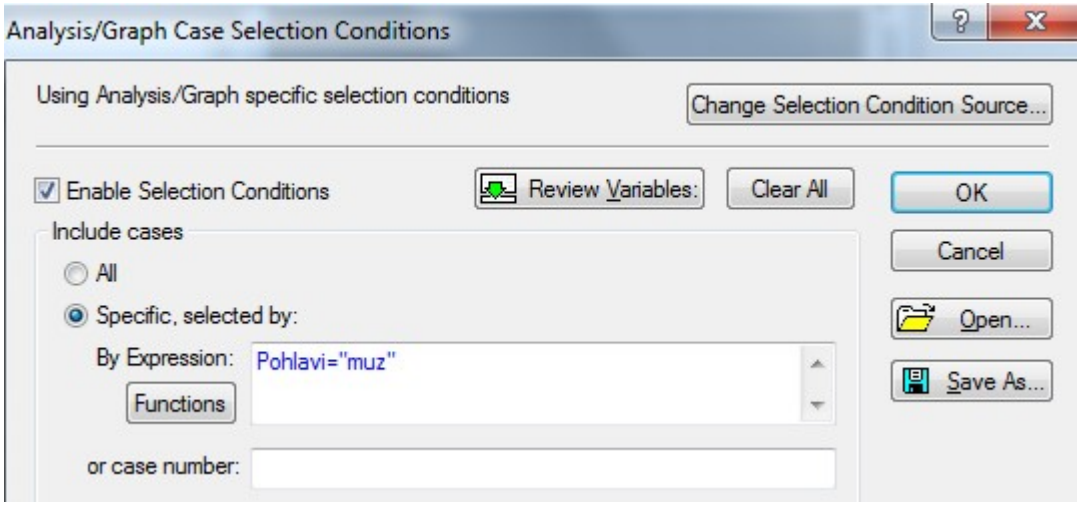


Descriptive Statistics (05_Priklady)

Variable	Valid N	Mean	Median	Minimum	Maximum	Std.Dev.
váha	61	65,63968	66,49219	49,80155	79,20183	4,988461

Základy popisné statistiky: soubor pacienti.sta

- ***Pokud bychom chtěli zjistit průměrnou váhu pouze u mužů, klikneme na tlačítko select cases a zvolíte Pohlaví="muz"(nezapomínejte na uvozovky)***

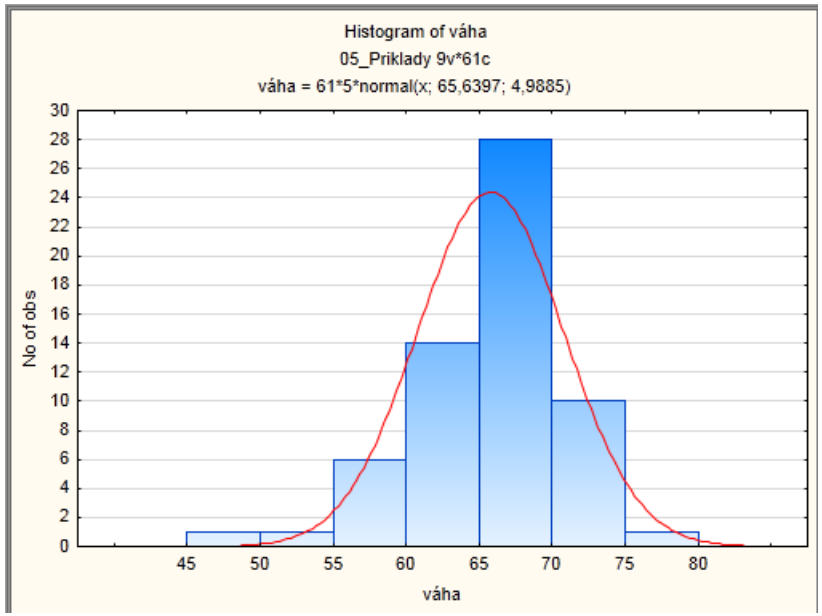


Variable	Descriptive Statistics (05_Priklady) Include condition: Pohlavi="muz" Valid N	Mean	Minimum	Maximum	Std.Dev.
váha	31	65,37337	49,80155	72,14984	5,034108

Základy popisné statistiky: soubor pacienti.sta

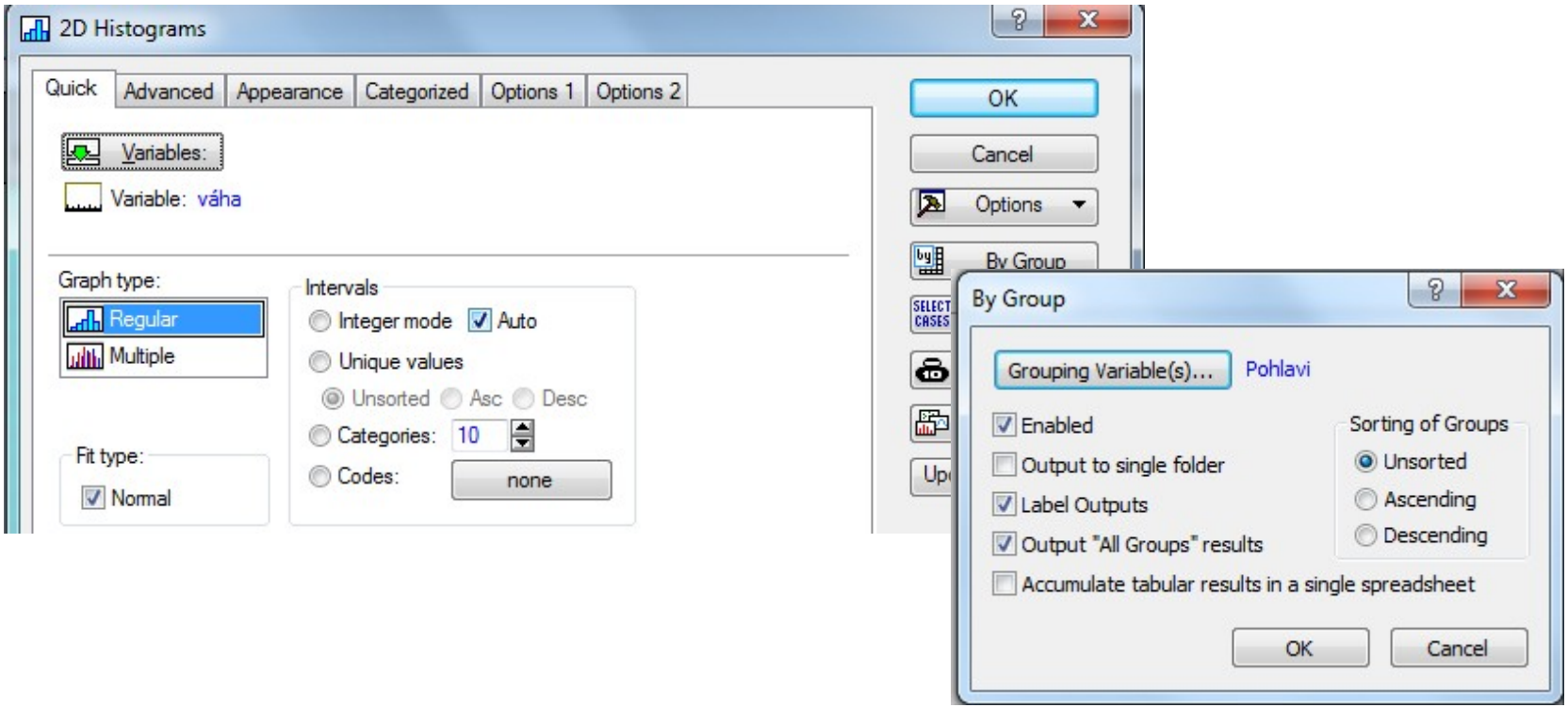
- ***Vytvořte histogram s rozpětím hodnot po pěti, poté zkuste to samé pro muže a ženy.***

Návod: Záložka Graphs->Histogram->proměnná váha, záložka Advanced: Intervals Boundaries, Specifies boundaries



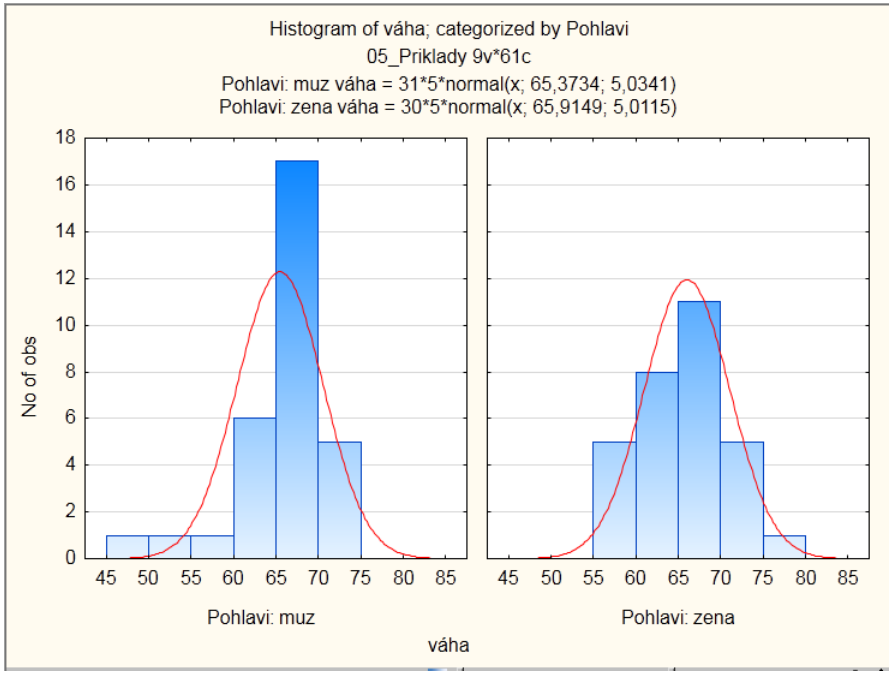
Základy popisné statistiky: soubor pacienti.sta

- ***Pokud chceme váhu odděleně pro pohlaví - po boku vpravo By group: vybereme proměnnou pohlaví .***



Základy popisné statistiky: soubor pacienti.sta

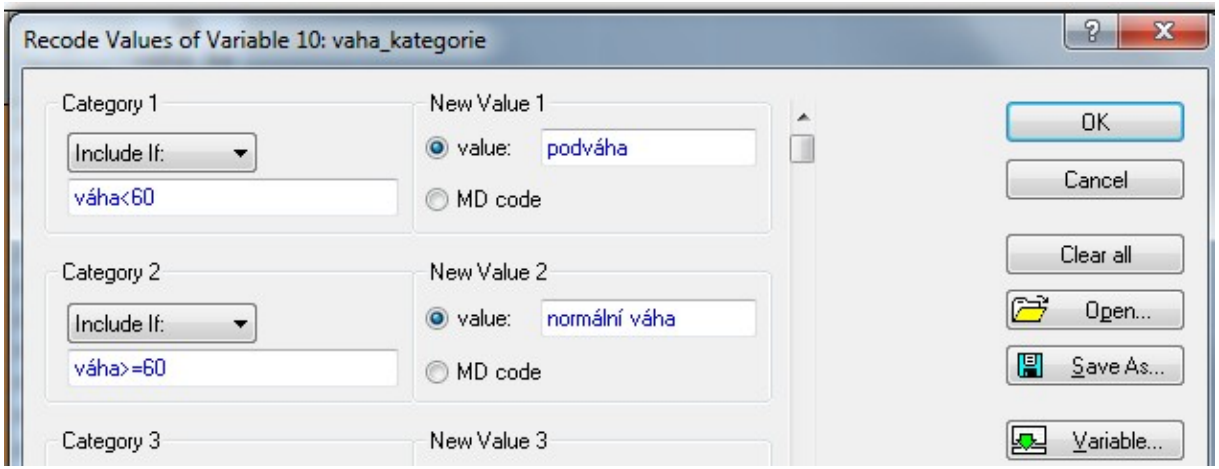
- ***Pokud chceme histogram váhy pro muže i ženy mít v jenom grafu: vybereme záložku Categorized, zapneme kategorii X a změníme proměnnou na pohlaví.***



Základy popisné statistiky: soubor pacienti.sta

- ***Překódovaní proměnné***
- ***Proměnnou váha překódujte do proměnné vaha_kategorie tak, aby pacienti pod 60 kg tvořili jednu skupinu a pacienti 60+ druhou skupinu.***

Návod: Vložíme novou proměnnou vaha_kategorie za proměnnou váha. Označíme novou proměnnou vaha_kategorie, záložka Data -> Recode



- ***Zjistěte, kolik % žen mělo váhu pod 60 kg?***

Samostatné cvičení: soubor studenti.sta

<u>Načtěte soubor studenti.sta</u>, který obsahuje údaje o 26 studentech, získané informace jsou shrnuty v proměnných A,B,C,D.
Návod: Záložka *Home* → *Open* → vybereme soubor studenti.sta.

<u>Změňte názvy proměnných</u>: A-jméno studenta, B-známka z biostatistiky, C-pohlaví, D-věk. U proměnných B a C popište jednotlivé varianty (proměnná B odpovídá známce: 1- výborně, 2- velmi dobře, 3- dobře, 4- nedostatečně; proměnná C odpovídá pohlaví:1 - muž, 2 - žena)
Návod: Vybereme nejprve příslušnou proměnnou A, 2krát klikneme myší → do položky *Name* napíšeme nový název proměnné (*All Specs*… umožní přejmenovat všechny proměnné najednou; *Text Labels* číselným hodnotám přiřadí textový popisek).

<u>Pojmenujte názvy řádků tabulky jmény studentů</u>, poté proměnnou jméno studenta smažte.
Návod: Záložka *Data* → *Names* → *Transfer case names from* → *Variable*: Jméno studenta;
smazání-vybereme proměnnou Jméno studenta, pravé tlačítko myši → *Delete Variable*.

Samostatné cvičení: soubor studenti.sta

U proměnné Známka zjistěte absolutní, relativní četnost, dále absolutní a relativní kumulativní četnost.
Návod: Záložka *Statistics* → *Basic Statistics* → *Frequency tables* → *Variables*: známka z biostatistiky → *Summary*

Zjistěte průměr, medián pro proměnnou Věk. U proměnné pohlaví zjistěte modus. Pro proměnnou známka zjistěte medián, modus.
Návod:
Způsob 1: Označíme proměnnou věk, pravé tlačítko → *Statistics of Block Data* → *Blocks columns* → *All*
Zbůsob 2: Záložka *Statistics* → *Basic Statistics* → *Descriptive statistics* → *Variables:* věk→ záložka *Advanced* → vybereme *Mean*, *Median.*

Samostatné cvičení: soubor studenti.sta

Proměnnou <u>věk překódujte</u> pomocí následujících 5 intervalů: <20,22>, (22,25>, (25,28>, (28,31>, (31,33> do proměnné Věk 2.
Návod: Vložíme novou proměnnou Věk 2 za proměnnou Věk. Označíme novou proměnnou
Věk 2, záložka *Data → Recode → Category 1*: věk>=20 and věk<=22, *New Value*: 1 atd.

Pomocí <u>koláčového grafu</u> znázorněte proměnnou <u>Známku a Pohlaví,</u> doplňte procenta (relativní četnost).
Návod: Záložka *Graphs → 2D → Pie Charts →* Záložka: *Quick: Variables*: Známka, Pohlaví; Záložka:*Advanced → Pie legends* vyber *Text and Percent.*

Pomocí <u>sloupcového grafu</u> znázorněte <u>věk pouze pro muže</u>.
Návod: Záložka *Graphs → 2D → Bar/Column Plots → Variables*: Věk, v tomtéž okně napravo klikneme na *Select Cases* →zaškrtneme možnost *Enable Selection Conditions → Specific→ selected by Expression*: Pohlaví=1.

Samostatné cvičení: soubor studenti.sta

Pro proměnnou Věk vytvořte histogram s intervaly širokými dva roky, poté zkuste to samé zvlášť pro muže a ženy.
Návod:Záložka *Graphs* → *Histogram* → *Variables:* věk, záložka *Advanced*: *Intervals Boundaries* → *Specifies boundaries* po boku vpravo *By group*: vybereme proměnnou pohlaví