Biostatistika

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. Základy popisné statistiky

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

Typy proměnných

Kvalitativní (kategoriální) proměnná

- lze ji řadit do kategorií, ale nelze ji kvantifikovat

  Příklad: ??

Kvantitativní (numerická) proměnná

- můžeme ji přiřadit číselnou hodnotu

  Příklad: ??

Typy proměnných

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 ...*

Kvalitativní znaky

- **Binární znaky:** dvě kategorie, obvykle se kódují pomocí číslic 1 (přítomnost sledovaného znaku) a 0 (nepřítomnost sledovaného znaku).

  Příklad: ??

- **Nominální znaky**: několik kategorií (A, B, C), které nelze uspořádat.

  Příklad: ??

- **Ordinální znaky:** několik kategorií, které lze vzájemně seřadit, tedy můžeme se ptát, která je větší/menší (1<2<3).

  Příklad: ??

Kvalitativní znaky

- **Binární znaky:** dvě kategorie, obvykle se kódují pomocí číslic 1 (přítomnost sledovaného znaku) a 0 (nepřítomnost sledovaného znaku).

  Příklady: Diabetes (1-ano, 0-ne),

  Pohlaví (1-muž, 0-žena).

- **Nominální znaky**: několik kategorií (A, B, C), které nelze uspořádat.

  Příklad: krevní skupiny (A/B/AB/0).

- **Ordinální znaky:** několik kategorií, které lze vzájemně seřadit, tedy můžeme se ptát, která je větší/menší (1<2<3).

  Příklady: stupeň bolesti (mírná/střední/velká),

  stadium maligního onemocnění (I/II/III/IV).

Kvantitativní znaky

- **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 $\alpha \in (0,1)$, pak α-kvantil x_α 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.
- $x_{0,50}$- **medián**, $x_{0,25}$- **dolní kvartil**, $x_{0,75}$-**horní kvartil**, $x_{0,1}$.... $x_{0,9}$-**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,

$$E(x) = \bar{x} = \sum_{i=1}^{n} x_i / n$$

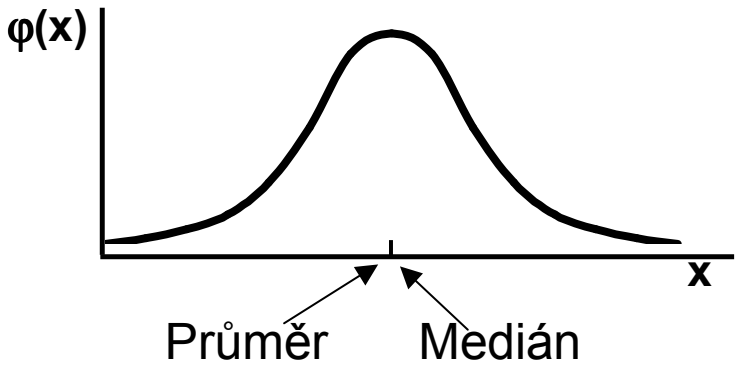
kde x_i jsou jednotlivé hodnoty a n jejich počet

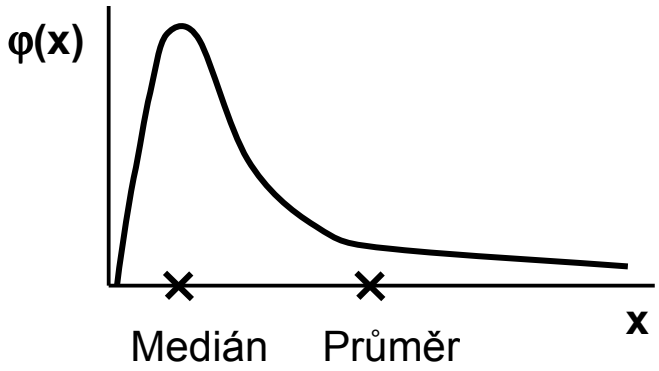
- **Geometrický průměr**: n kladných hodnot x_i, $\sqrt[n]{x_1 * \ldots\ldots * x_n}$, má smysl všude, kde má nějaký informační smysl součin hodnot proměnné. Z praktického hlediska platí, že logaritmus geometrického průměru je roven aritmetickému průměru logaritmovaných hodnot souboru.

Průměr vs medián

PAMATUJ:

- **Průměr je silně ovlivněn extrémními hodnotami (tzv. odlehlá pozorování), medián není ovlivněn vybočujícími pozorováními**
- **Průměr je vhodný ukazatel středu u normálního/symetrického rozložení, medián je vhodnou charakteristikou středu souboru i v případě veličin s neznámým rozdělením**
- V případě symetrického rozložení jsou jejich hodnoty v podstatě shodné, v případě asymetrického rozložení však nikoliv!





Charakteristiky variability

Charakteristiky variability u ordinálních znaků

- **Kvartilové rozpětí (odchylka):** $q = x_{0,75} - x_{0,25}$

Charakteristiky variability u intervalových a poměrových znaků

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

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

  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 %)

Další parametry rozložení

- **Počet hodnot** – důležitý ukazatel, znamená jak moc lze na data spoléhat
- **Suma hodnot**
- **Minimum, maximum**
- **Variační rozpětí (rozsah)** – rozdíl mezi největší a nejmenší hodnotou řady
- **Střední chyba průměru (SE)** – měří rozptýlenost vypočítaného aritmetického průměru v různých výběrových souborech vybraných z jednoho základního souboru

Ukázka popisu a vizualizace kvalitativních dat

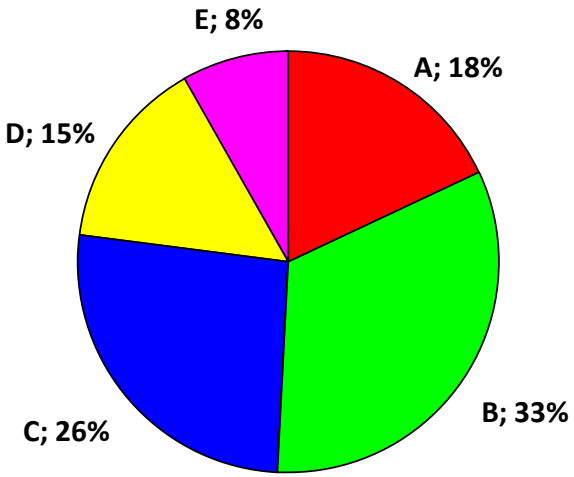
- **Popis kvalitativních dat:** frekvence jednotlivých kategorií
- **Vizualizace kvalitativních dat:** nejčastěji koláčový nebo sloupcový graf

Příklad: Známka z biostatistiky (podzim 2014)

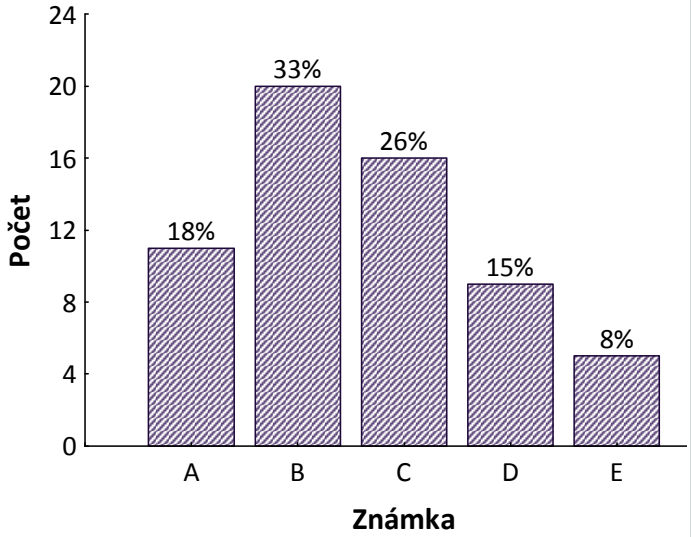
Frekvenční tabulka

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



Ukázka popisu kvantitativních dat

- **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 (cm^2)	22,2
min-max (cm)	144,1 - 169,2
dolní-horní kvartil (cm)	158,1 - 164,2

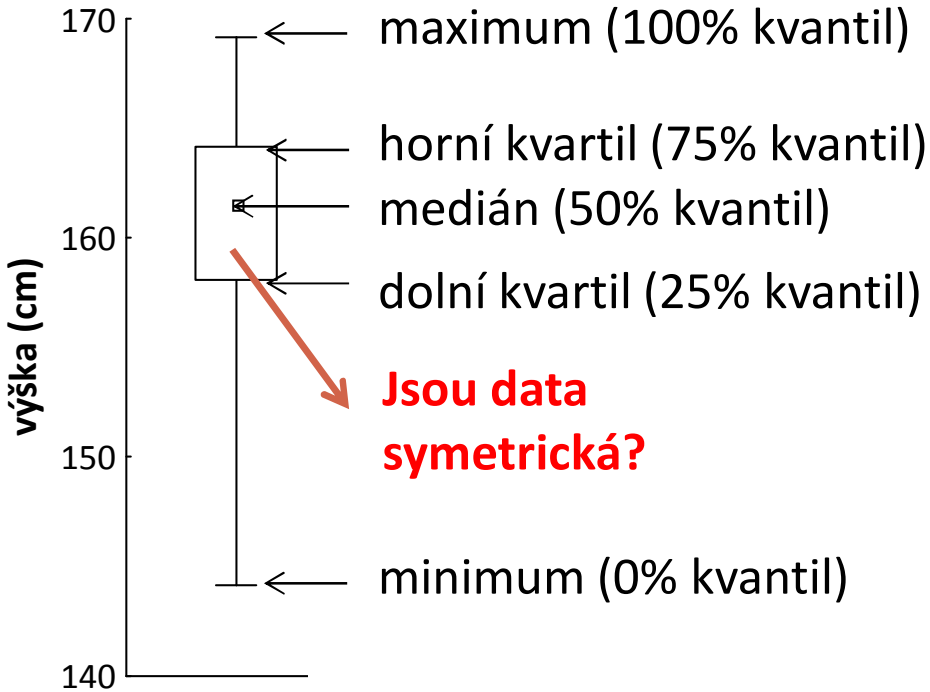
Průměr a medián se téměř shodují. Co nám to říká?

Ukázka vizualizace kvantitativních dat

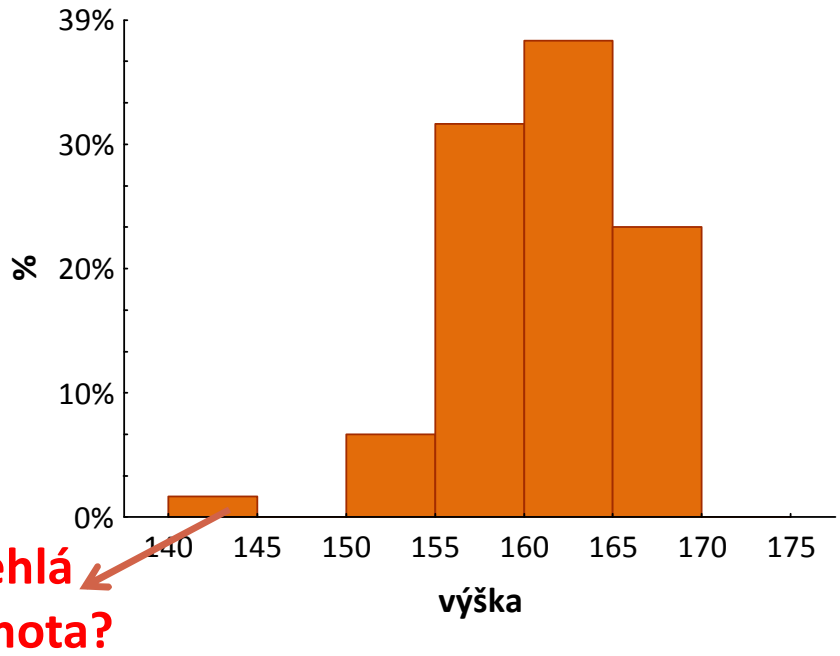
- **Vizualizace kvantitativních dat:** nejčastěji pomocí krabicového grafu nebo histogramu

Příklad: Popis výšky (cm) pacientů

Krabicový graf



Histogram



II. 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**

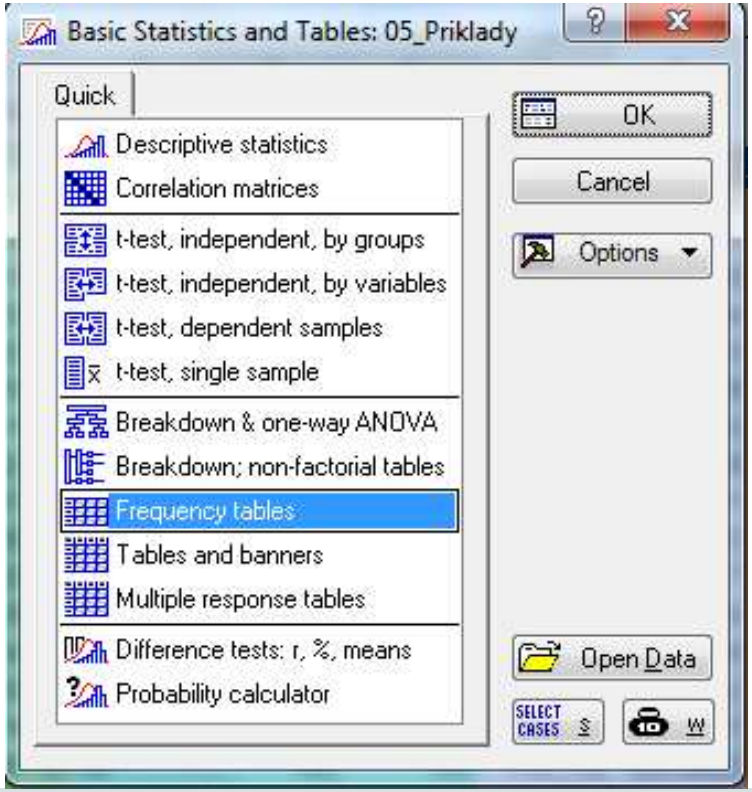
Nalézt: **Statistica 13** – kliknout **Získat**

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***

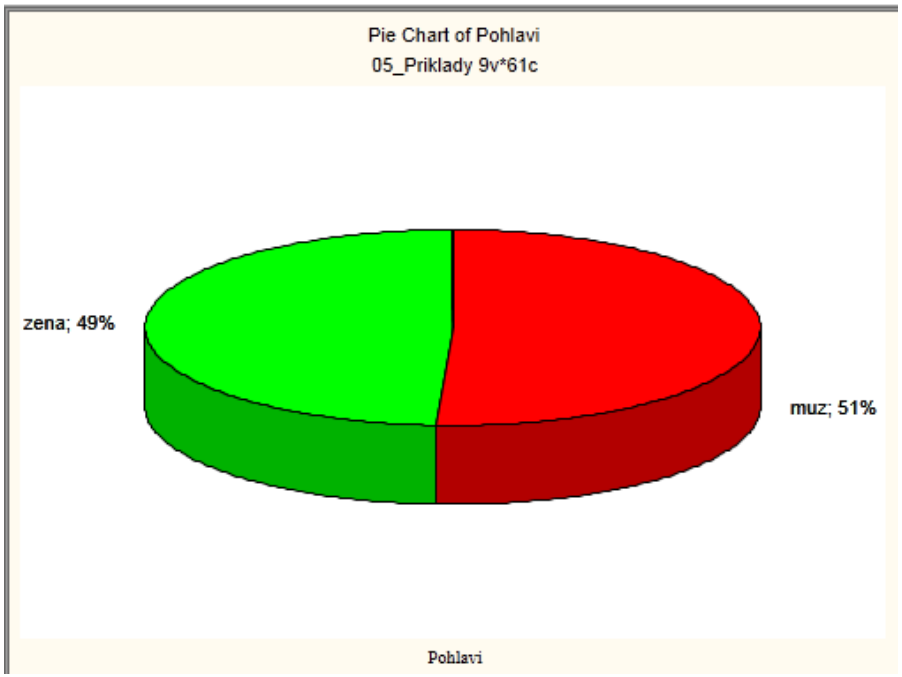


Frequency table: Pohlavi: =Round(Rnd(1);0) (05_F

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

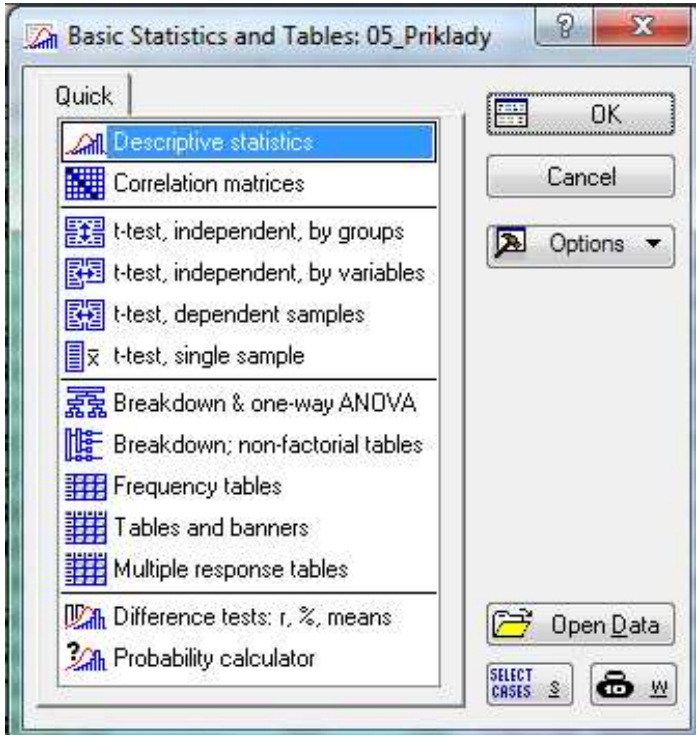
Základy popisné statistiky: soubor pacienti.sta

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



Základy popisné statistiky: soubor pacienti.sta

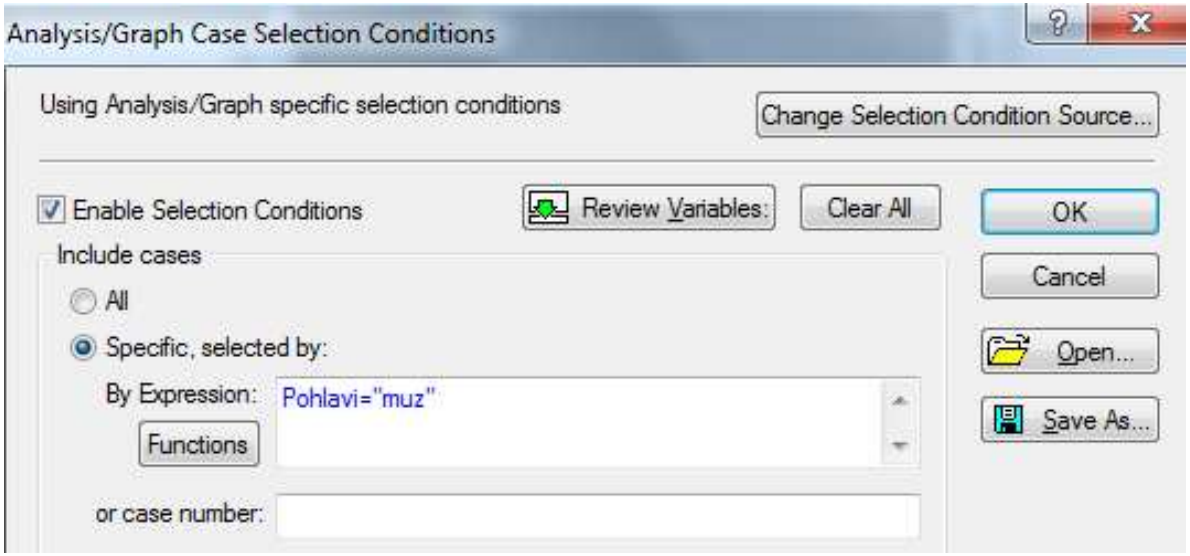
- ***Nyní budeme pracovat se spojitou proměnnou.***
- ***Pro proměnnou váha zjistěte: průměr, medián, minimum a maximum***



Variable	Descriptive Statistics (05_Priklady)					
	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)***

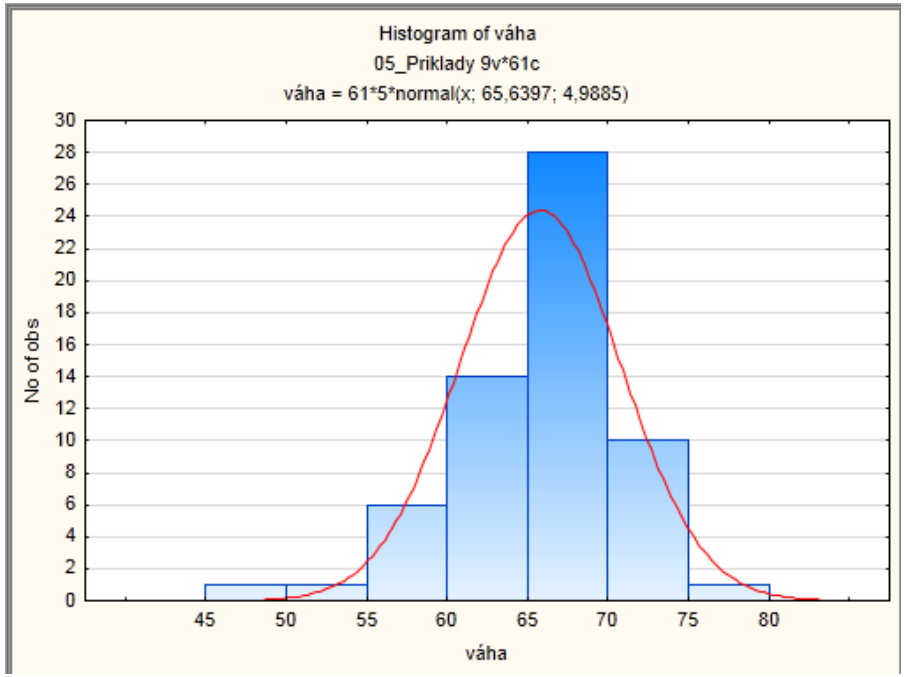


Variable	Descriptive Statistics (05_Priklady) Include condition: Pohlavi="muz" Valid N	Mean	Minimum	Maximum	Std.Dev.
váha	31	65,37337	49,80155	72,14984	5,034108

Základy popisné statistiky: soubor pacienti.sta

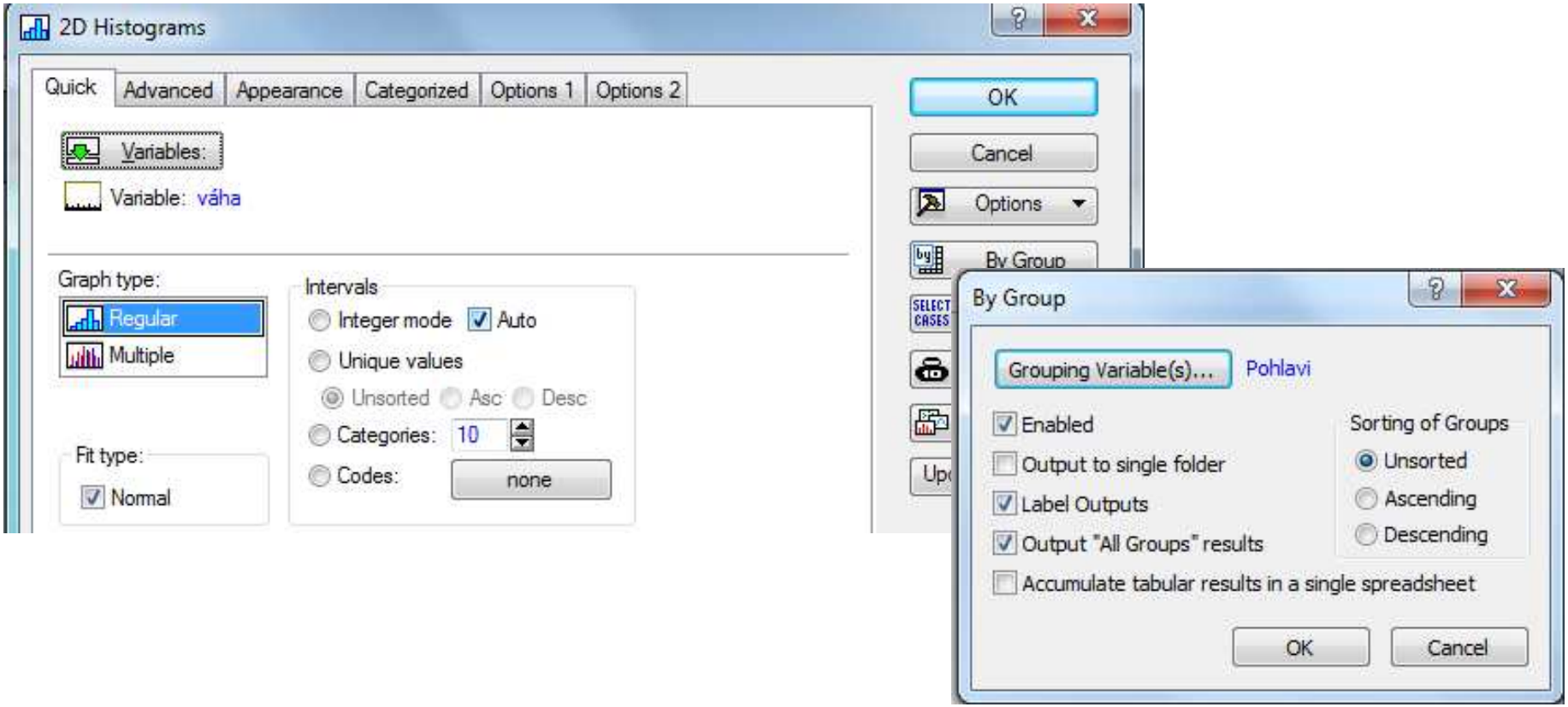
- ***Vytvořte histogram s rozpětím hodnot po pěti, poté zkuste to samé pro muže a ženy.***

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



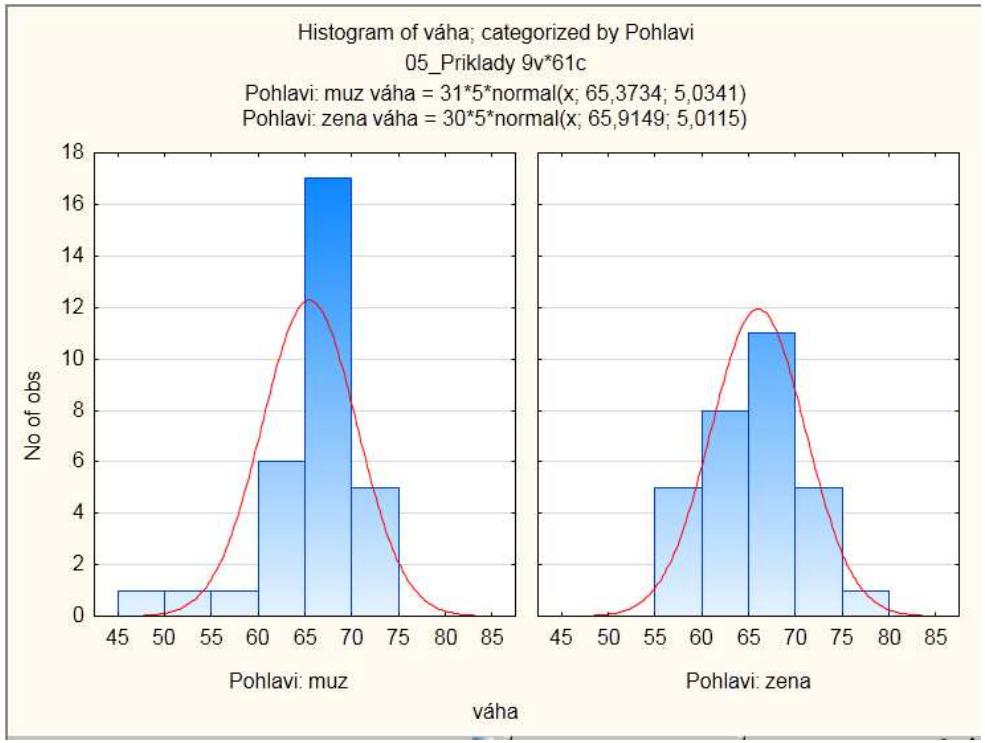
Základy popisné statistiky: soubor pacienti.sta

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



Základy popisné statistiky: soubor pacienti.sta

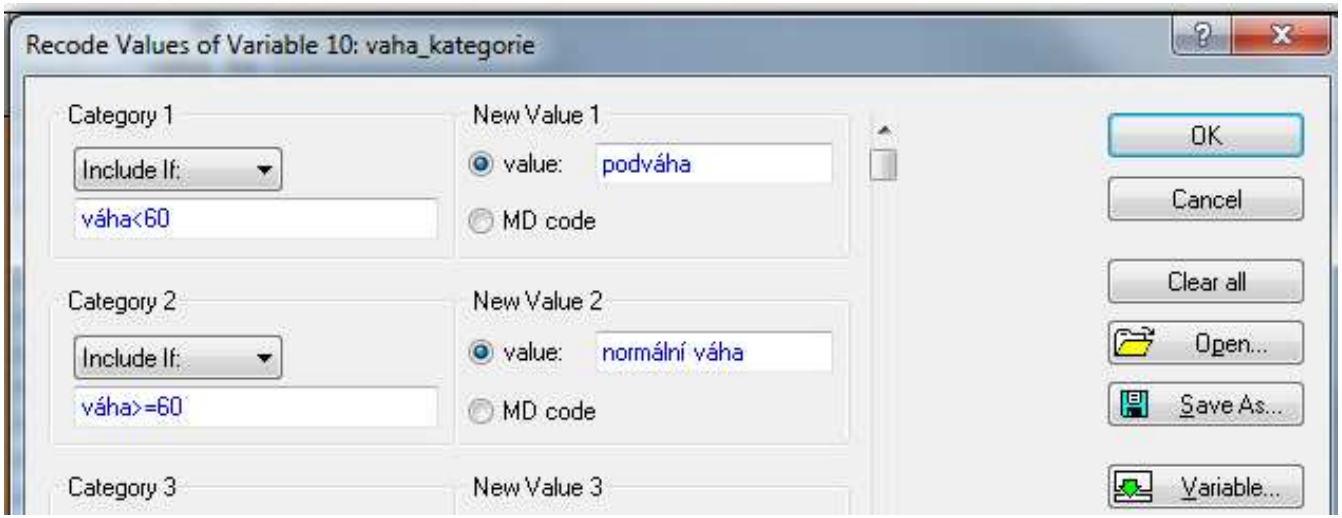
- ***Pokud chceme histogram váhy pro muže i ženy mít v jenom grafu: vybereme záložku Categorized, zapneme kategorii X a změníme proměnnou na pohlaví.***



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

Pomocí sloupcového grafu znázorněte věk pouze pro muže.
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

Opakování

Základy popisné statistiky
Vizualizace dat

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

Opakování

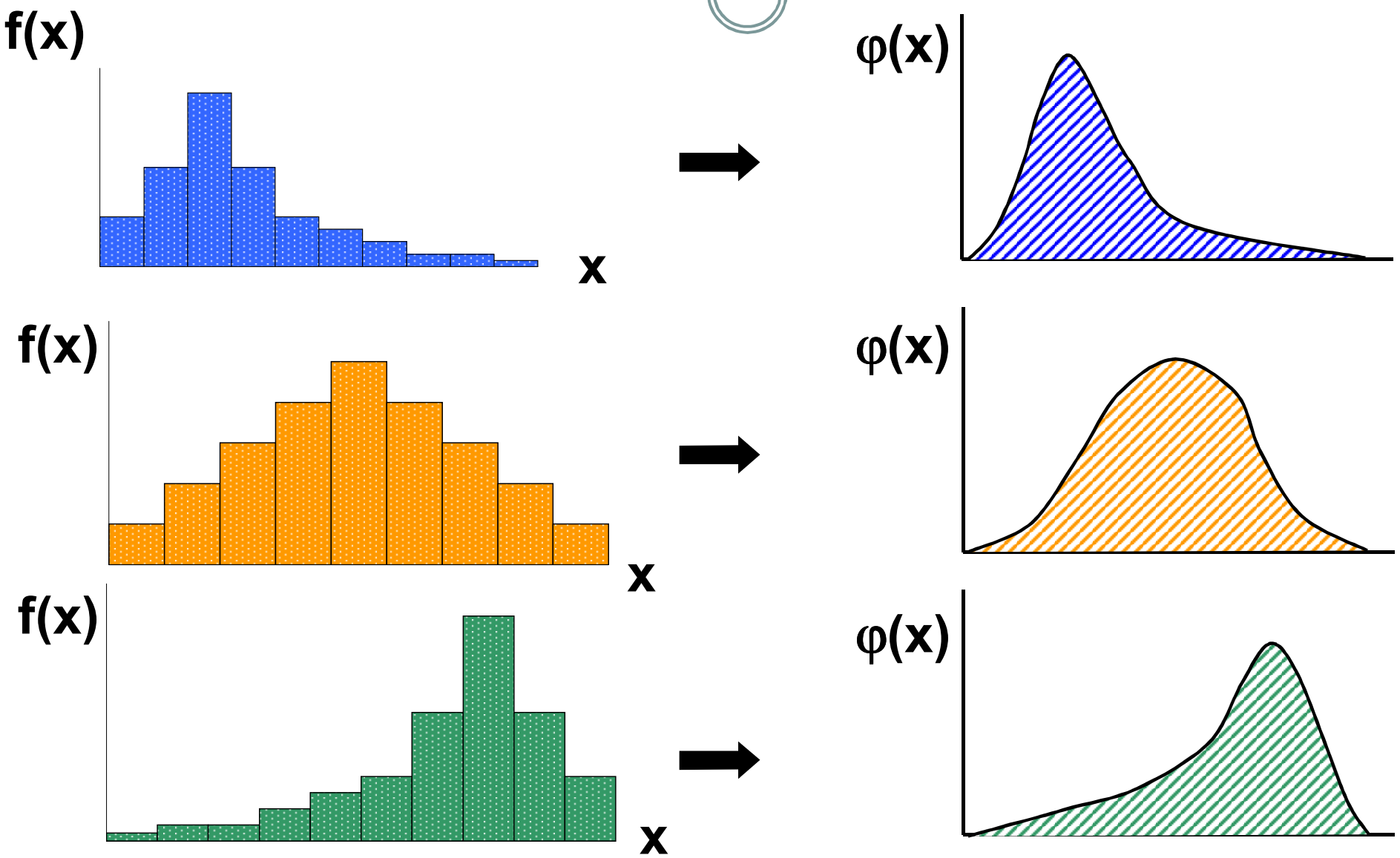
1. Co jsou **kvalitativní** a **kvantitativní data**?
2. Jaký je rozdíl mezi **spojitými** a **diskrétními daty**?
3. Uveďte **příklady binárních / nominálních / ordinálních dat**.
4. Uveďte **příklady spojitých** a **diskrétních dat**.
5. Jakými charakteristikami **popisujeme kvalitativní data**?
6. Jakými charakteristikami **popisujeme kvantitativní data**?
7. Jak správně **vizualizujeme kvalitativní data**?
8. Jak správně **vizualizujeme kvantitativní data**?

Modelová rozložení

Parametry rozložení
Přehled modelových rozložení
Logaritmicko-normální rozložení

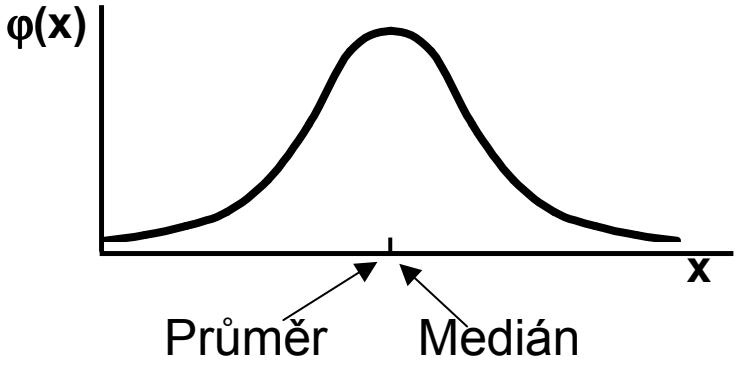
Vytvořil Institut biostatistiky a analýz, Masarykova univerzita
J. Jarkovský, L. Dušek

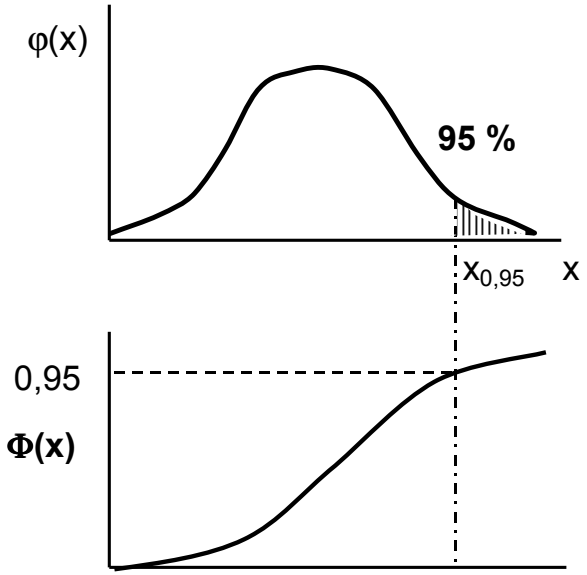
Výběrové rozložení hodnot lze modelově popsat a definovat tak pravděpodobnost výskytu X



Parametry rozložení

- Soubor dat (řada čísel) můžeme charakterizovat parametry jeho rozložení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
  - **Středu** (medián, průměr, geometrický průměr)
  - **Šířky rozložení** (rozsah hodnot, rozptyl, směrodatná odchylka)
  - **Tvaru rozložení** (skewness, kurtosis)
  - **Kvantily rozložení** – kolik % řady dat leží nad a pod kvantilem





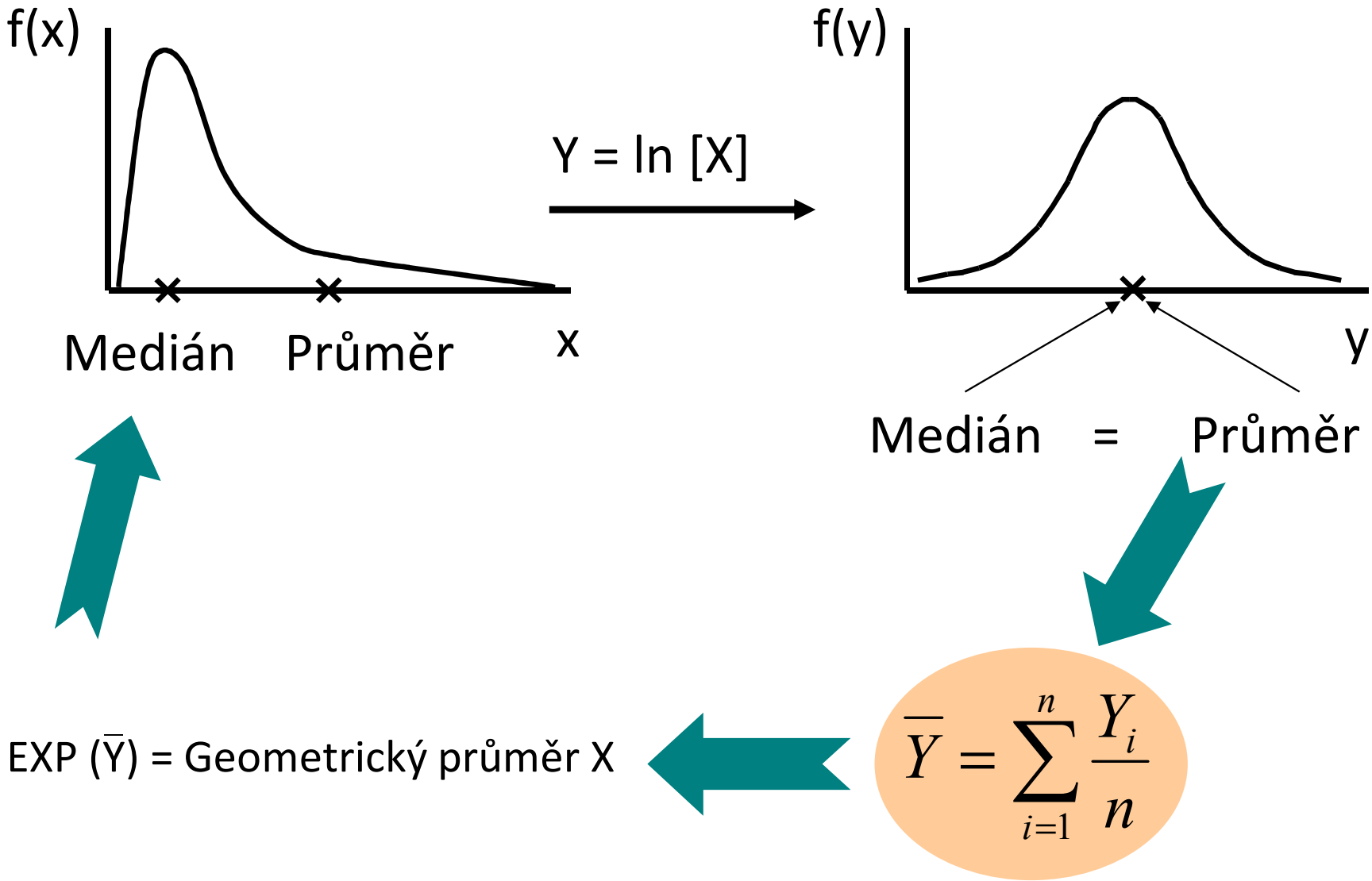
Stručný přehled modelových rozložení I.

Rozložení	Parametry	Stručný popis
Normální	Průměr (μ) Rozptyl (σ^2)	Symetrická funkce popisující intervalovou hustotu četnosti; nejpravděpodobnější jsou průměrné hodnoty znaku v populaci.
Log-normální	Medián Geometrický průměr Rozptyl (σ^2)	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
Weibullovo	α - parametr tvaru β - parametr rozsahu hodnot	Změnou parametru a lze modelovat distribuci doby přežití, např. stresovaného organismu. Rozložení využívané i jako model k odhahu LC_{50} nebo EC_{50} u testů toxicity.
Rovnoměrné	Medián Geometrický průměr Rozptyl (σ^2)	Funkce intervalové hustoty četnosti, která po logaritmické transformaci nabude tvaru normálního rozložení.
Triangulární	$f(x) = [b - ABS(x - a)] / b^2$ $a - b < x < a + b$	Pravděpodobnostní funkce pro typ rozložení, kdy jsou střední hodnoty výrazně pravděpodobnější než hodnoty okrajové.
Gamma	Parametry distribuční funkce: α - parametr tvaru β - parametr rozsahu hodnot	Umožňuje flexibilně modelování distribučních funkcí nejrůznějších tvarů. Např. χ^2 rozložení je rozložení typu Gamma. Gamma rozložení s a = 1 je známo jako exponenciální rozložení.

Stručný přehled modelových rozložení II.

Rozložení	Parametry	Stručný popis
Beta	Parametry distribuční funkce: α - parametr tvaru β - parametr rozsahu hodnot	Pravděpodobnostní funkce pro proměnnou omezenou rozsahem do intervalu [0; 1]. Je matematicky komplikovanější, ale velmi flexibilní při popisu změn hodnot proměnné v ohraničeném intervalu.
Studentovo	Stupně volnosti - uvažuje velikost vzorku Průměr Rozptyl	Simuluje normální rozložení pro menší vzorky čísel. Pro větší soubory (n > 100) se limitně blíží k normálnímu rozložení.
Pearsonovo (Chí-kvadrát)	Stupně volnosti - uvažuje velikost vzorku	Slouží především k porovnání četností jevů ve dvou a více kategoriích. Používá se k modelování rozložení odhadu rozptylu normálně rozložených dat.
Fisher-Snedecorovo	Dvojí stupně volnosti - uvažuje velikost dvou vzorků	Používá se k testování hodnot průměrů - F test pro porovnání dvou výběrových rozptylů; F test, ANOVA atd.

Log-normální rozložení lze jednoduše transformovat



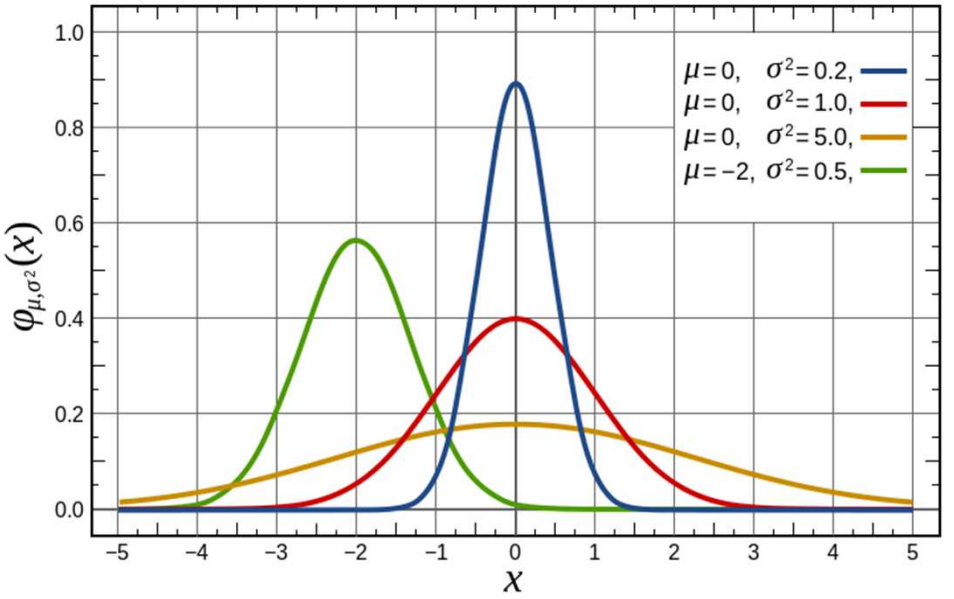
Normální rozložení

Normální rozložení
Pravidlo 3 sigma
Parametry normálního rozložení
Vizuální ověření normality dat

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

Normální rozdělení

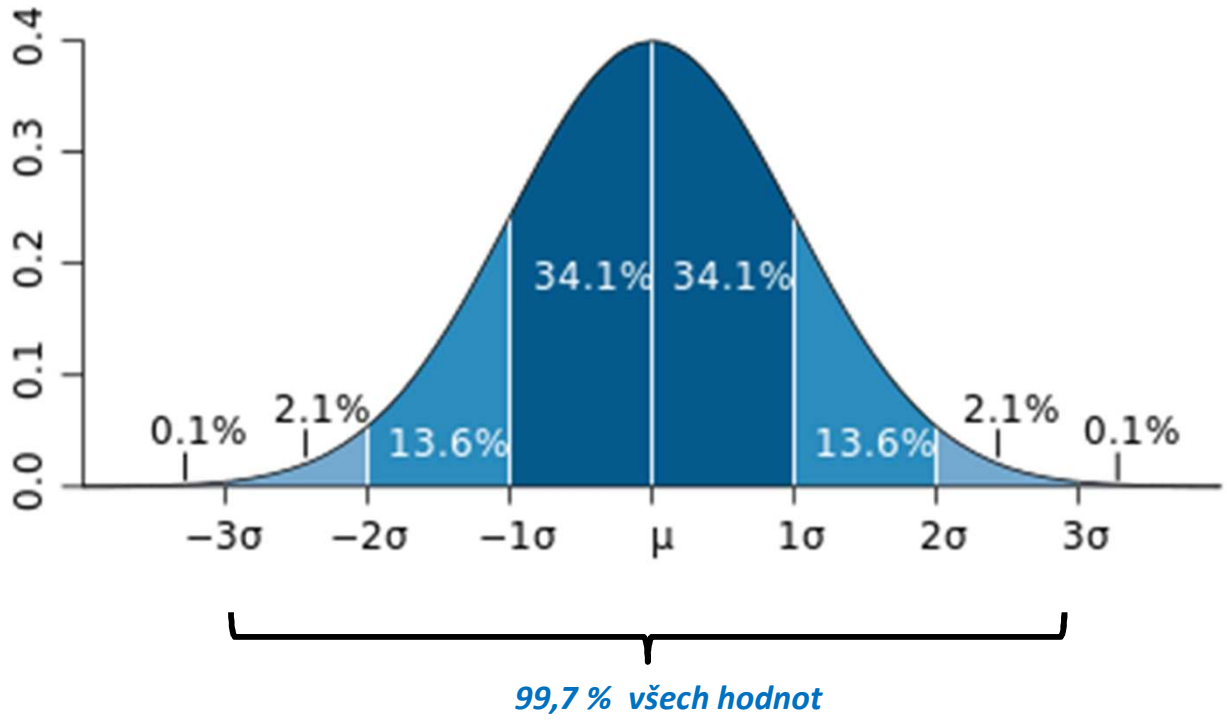
- Nejklasičtějším modelovým rozložením, od něhož je odvozena celá řada statistických analýz je tzv. **normální rozložení**, známé též jako **Gaussova křivka**.
- Popisuje rozdělení pravděpodobnosti spojité náhodné veličiny: např. výška v populaci, chyba měření…
- Je kompletně popsáno dvěma parametry:
  μ – střední hodnota
  σ^2 – rozptyl
  Označení: **N(μ, σ^2)**



- Normalita je klíčovým předpokladem řady statistických metod
- Pro ověření normality existuje řada testů a grafických metod

Pravidlo 3 sigma

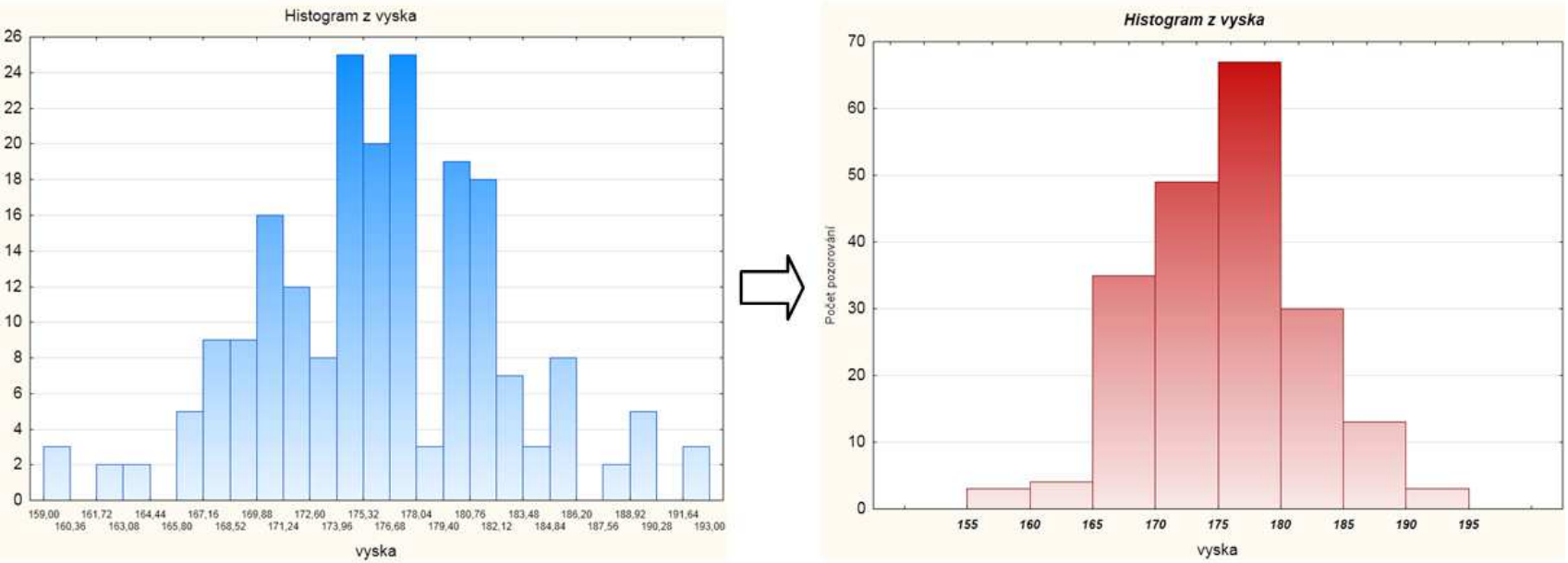
- V rozmezí μ ± 3σ by se mělo vyskytovat 99,7 % všech hodnot



- Použití: zhodnotíme tvar rozdělení (pouze orientačně) a přítomnost odlehlých hodnot

Vizuální ověření normality

- Pro hodnocení tvaru rozložení lze využít histogram
  (nevýhoda: nutné určit „vhodný“ počet sloupců)



- Vhodnější jsou:
  1. **Q-Q graf** (kvantil-kvantilový graf)
  2. **P-P graf** (pravděpodobnostně-pravděpodobnostní graf)
  3. **N-P graf** (normální-pravděpodobnostní graf)

Řešení v softwaru Statistica

• *V menu Graphs zvolíme 2D Graphs*

1

Normal Probability Plots...

Quantile-Quantile Plots...

Probability-Probability Plots...

2

Quantile-Quantile Plots

Quick | Advanced | Appearance | Categorized | Options 1 | Options 2

Variables:

none

Výběr rozdělení

Distribution:

Normal

Beta

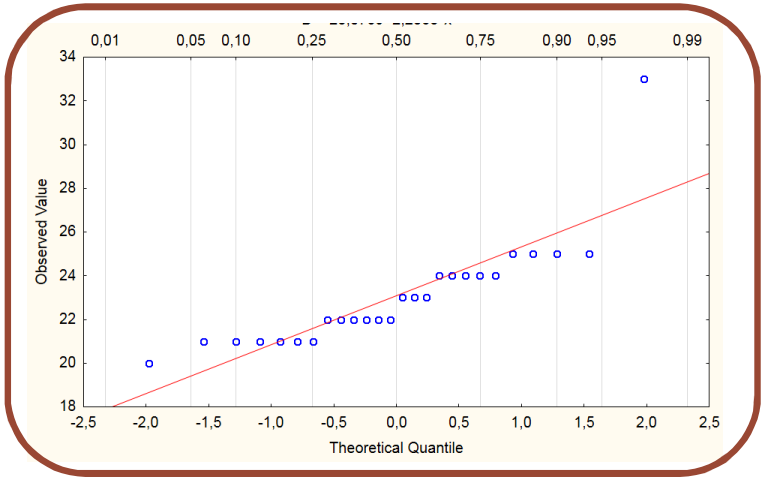
Plot layout

Multiple plots in one graph

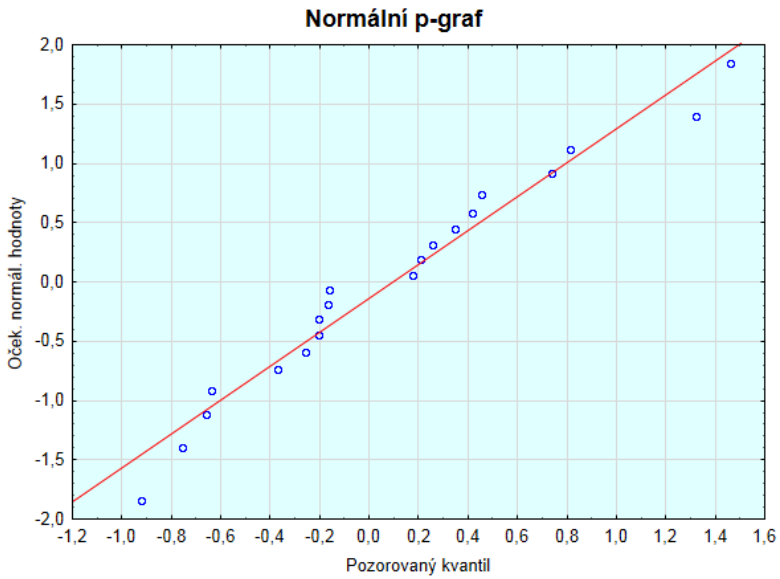
Do not assign average ranks to tied observations

• *V případě, že máme v datech několik stejných hodnot, je vhodné odškrtnout Neurčovat průměrnou pozici svázaných pozorování*

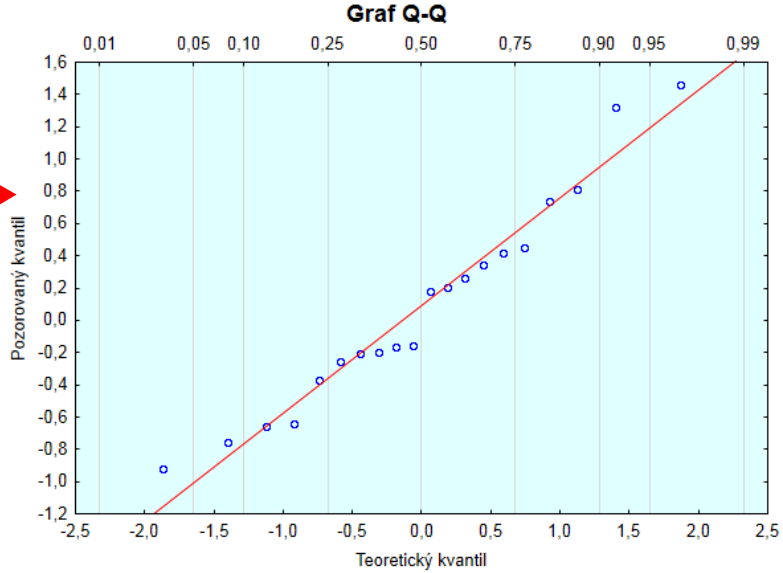
3



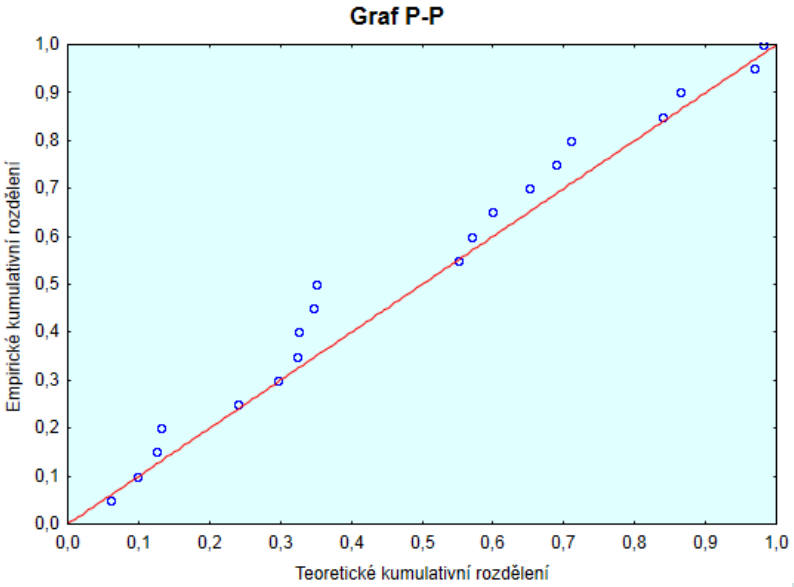
Rozdíl mezi N-P, Q-Q, P-P grafem



???



- *Pouze výměna os*
- *Znázorněn pozorovaný a teoretický kvantil*

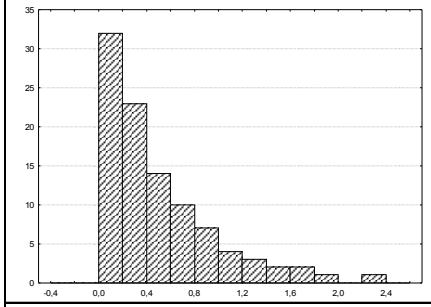
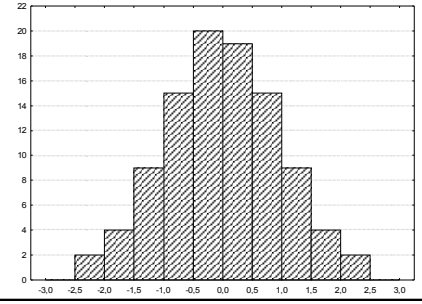
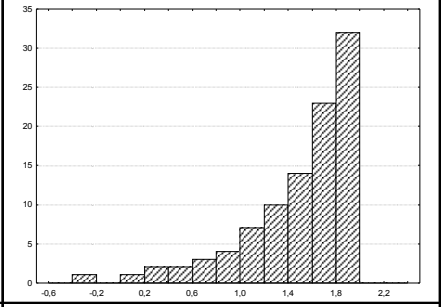
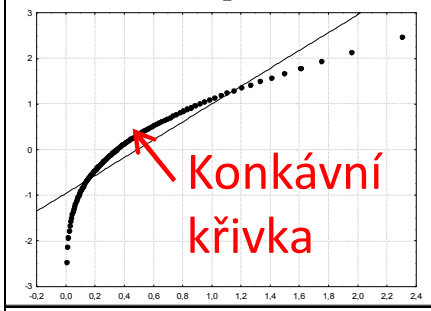
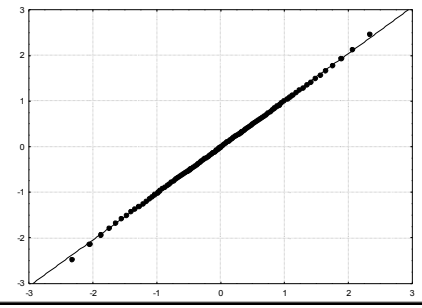
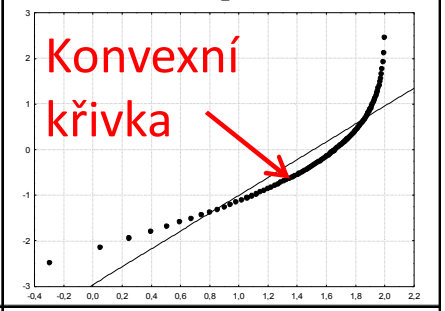
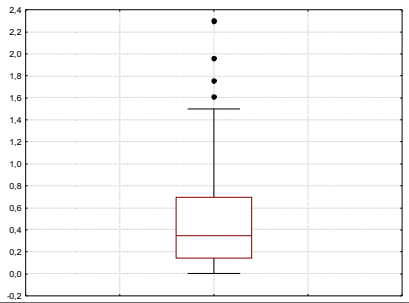
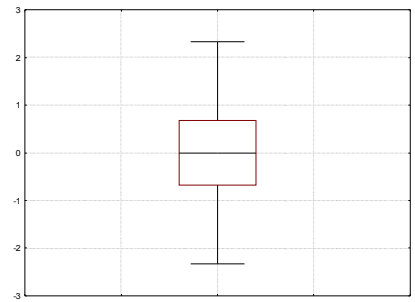
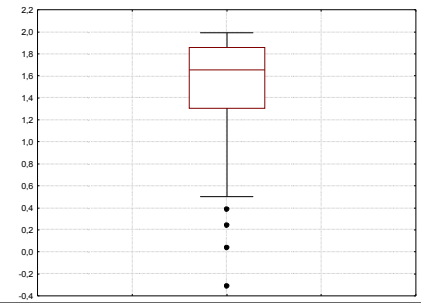


- *Vykresleno kumulativní rozdělení*

PAMATUJ:

Pocházejí-li data z normálního rozložení, pak body budou ležet okolo přímky

Jak se projeví asymetrie dat v diagnostických grafech?

Rozložení s kladnou šikmostí	Normální rozložení	Rozložení se zápornou šikmostí
Histogram	Histogram	Histogram
		
NP plot	NP plot	NP plot
		
Krabicový diagram	Krabicový diagram	Krabicový diagram
		

Výukové materiály: Výpočetní statistika, RNDr. Marie Budíková, Dr., 2011