

BIOSTATISTIKA

Tato prezentace je autorským dílem vytvořeným zaměstnanci Masarykovy univerzity. Studenti předmětu mají právo pořídit si kopii prezentace pro potřeby vlastního studia. Jakékoliv další šíření prezentace nebo její části bez svolení Masarykovy univerzity je v rozporu se zákonem.

Modelová rozdělení (rozložení)

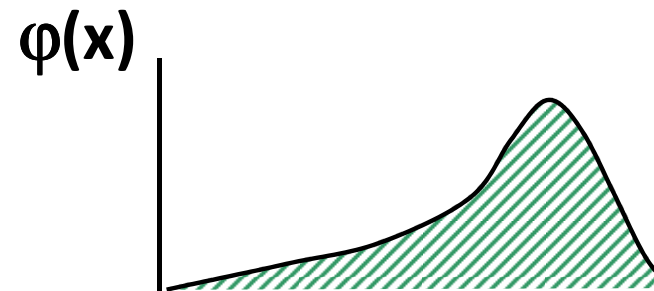
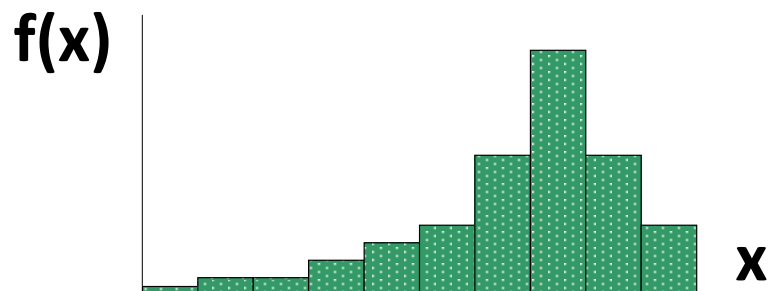
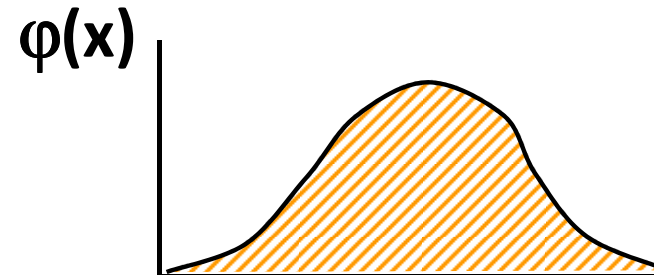
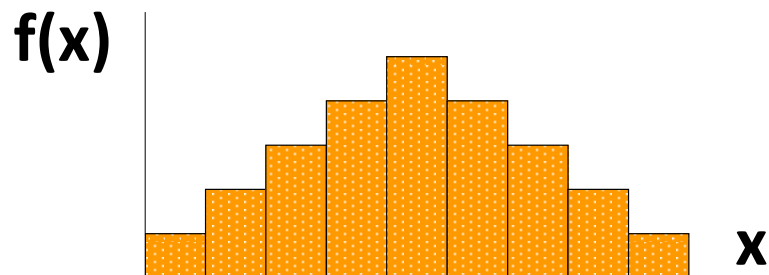
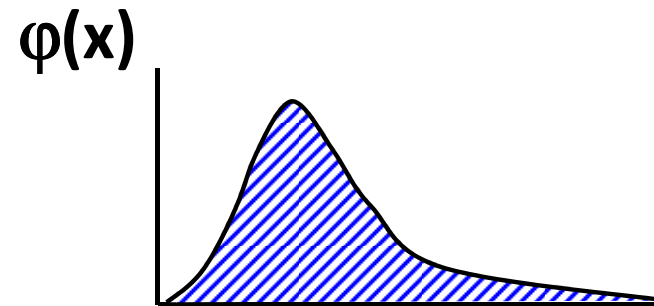
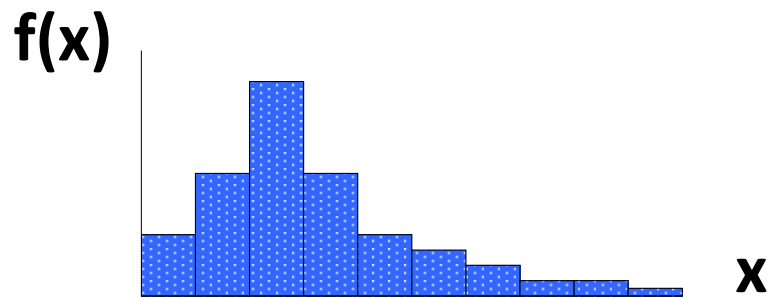
Parametry rozdělení

Přehled modelových rozdělení

Logaritmicko-normální rozdělení

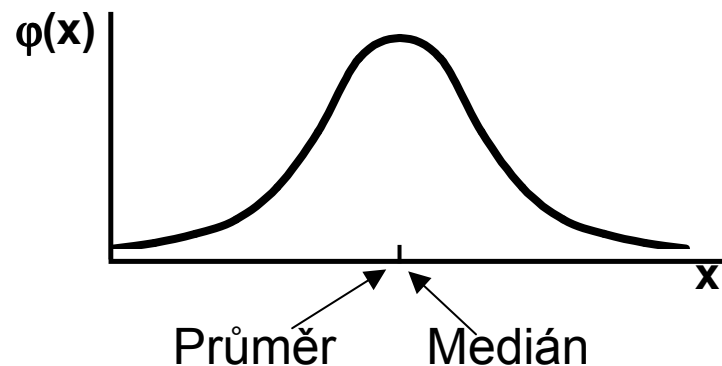
Výběrové rozdělení hodnot

- Lze popsat a definovat pravděpodobnost výskytu X

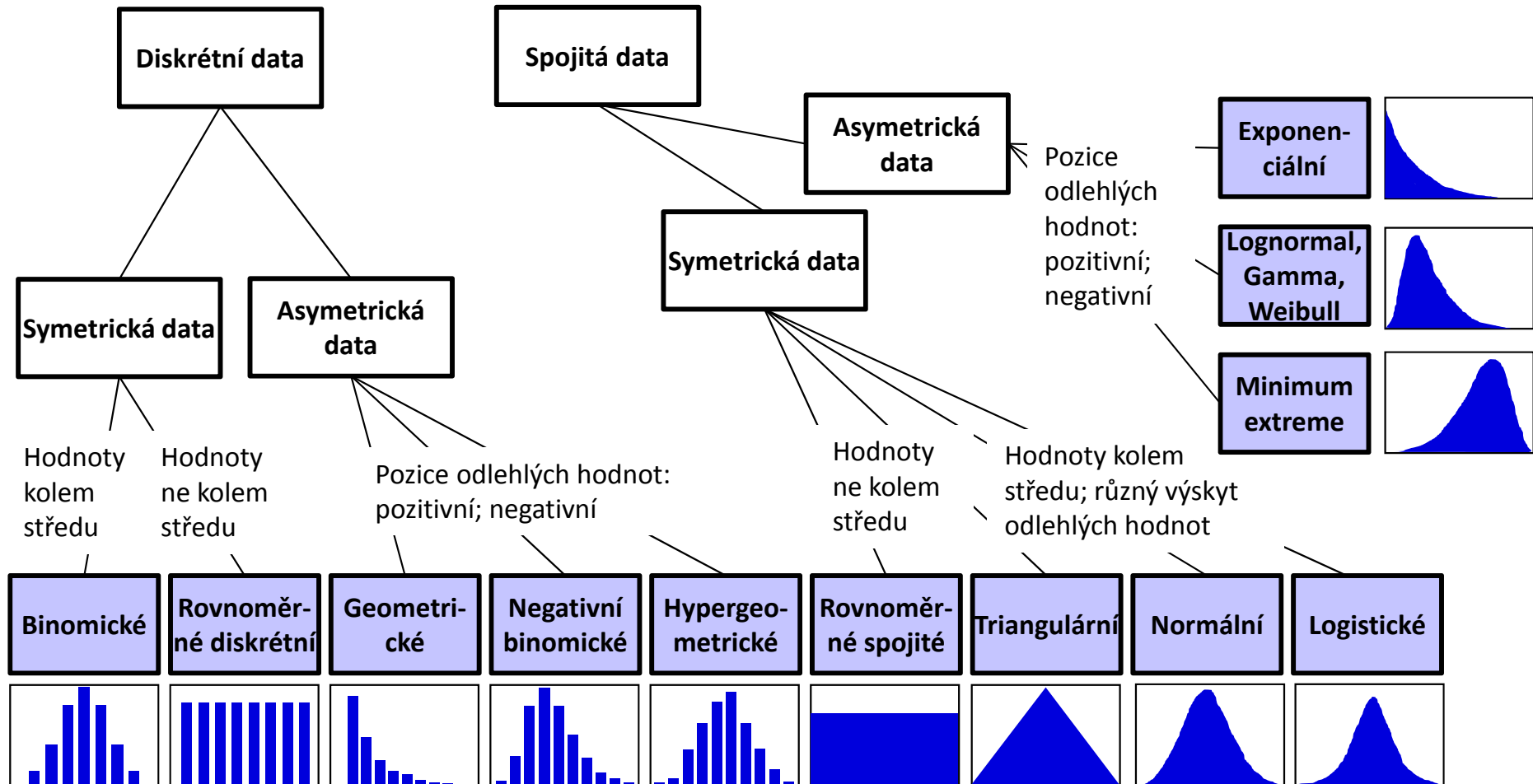


Parametry rozdělení

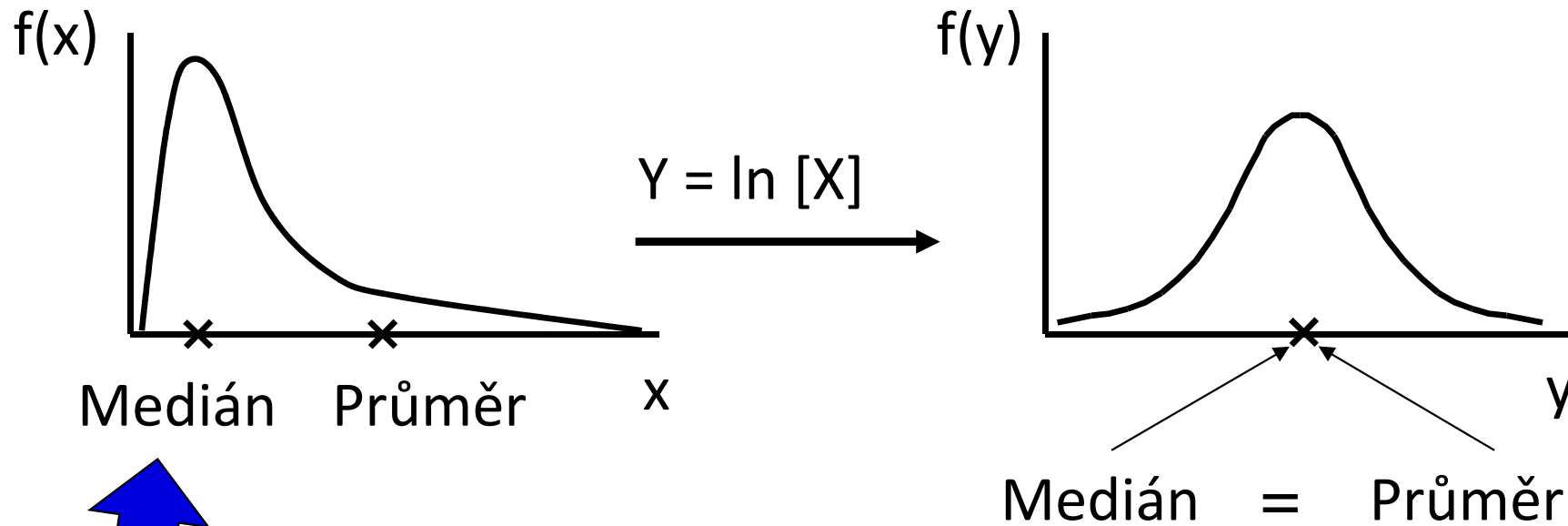
- Proměnné můžeme charakterizovat parametry rozdělení
- Hlavní skupiny těchto parametrů můžeme charakterizovat jako ukazatele:
 - **Středu** (medián, průměr, geometrický průměr)
 - **Šířky rozdělení** (rozsah hodnot, rozptyl, sm. odchylka)
 - **Tvaru rozdělení** (skewness, kurtosis)
 - **Kvantily rozdělení**



Přehled modelových rozdělení



Log-normální a normální rozdělení



EXP (Y) = Geometrický průměr X

$$\bar{Y} = \sum_{i=1}^n \frac{Y_i}{n}$$

Normální rozdělení

Normální rozdělení

Pravidlo 3 sigma

Parametry normálního rozdělení

Vizuální ověření normality dat

Normální rozdělení

- Nejklasičtějším modelovým rozdělením, od něhož je odvozena celá řada statistických analýz je tzv. **normální rozdělení**, známé též jako **Gaussova křivka**.
- Popisuje rozdělení pravděpodobnosti spojité náhodné veličiny, např. výška v populaci, chyba měření ...
- Je kompletně popsáno dvěma parametry:

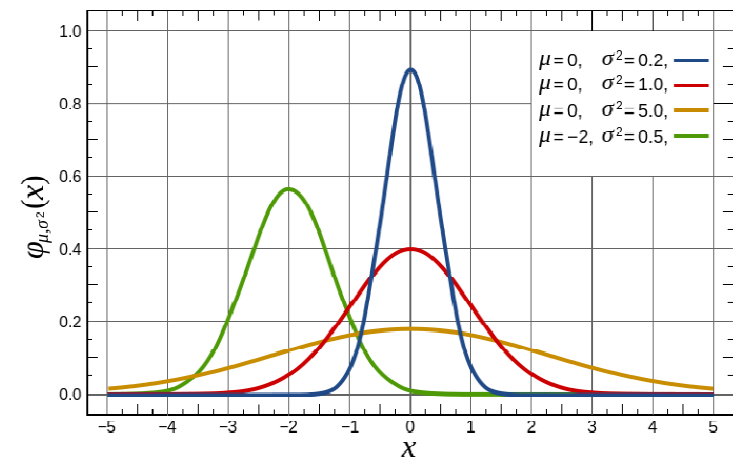
μ – střední hodnota

σ^2 – rozptyl

Označení: **$N(\mu, \sigma^2)$**

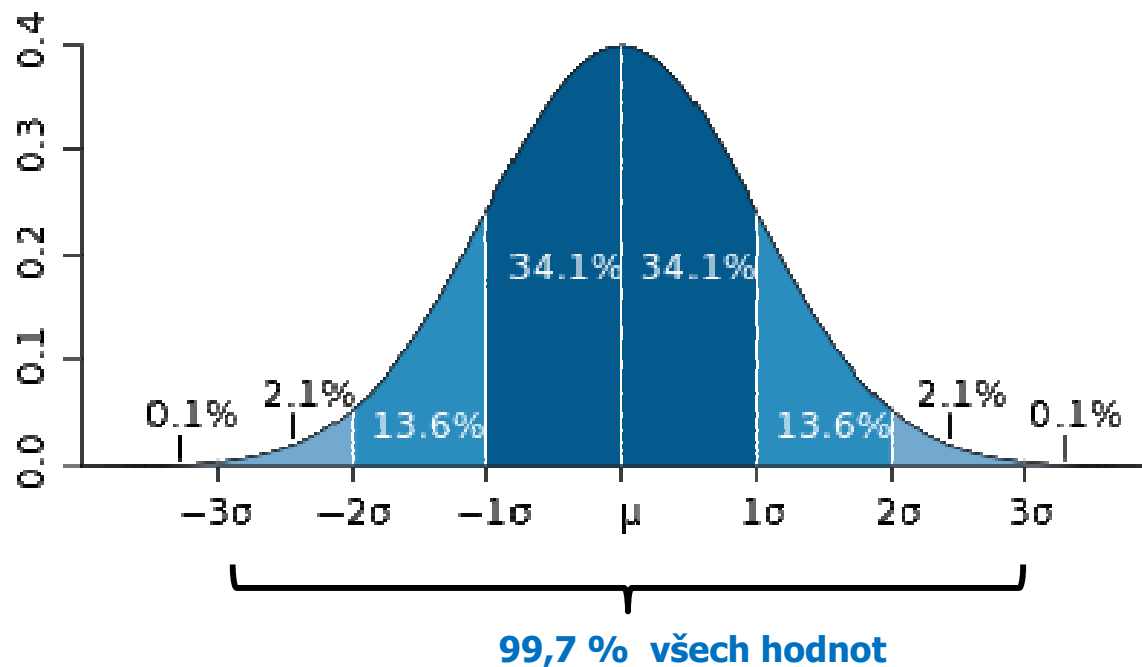


NORMALITA je klíčovým předpokladem řady statistických metod



Pravidlo 3 sigma

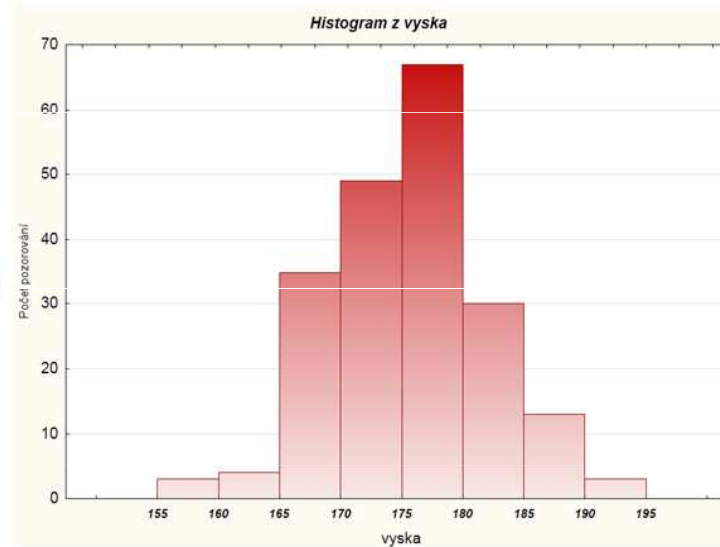
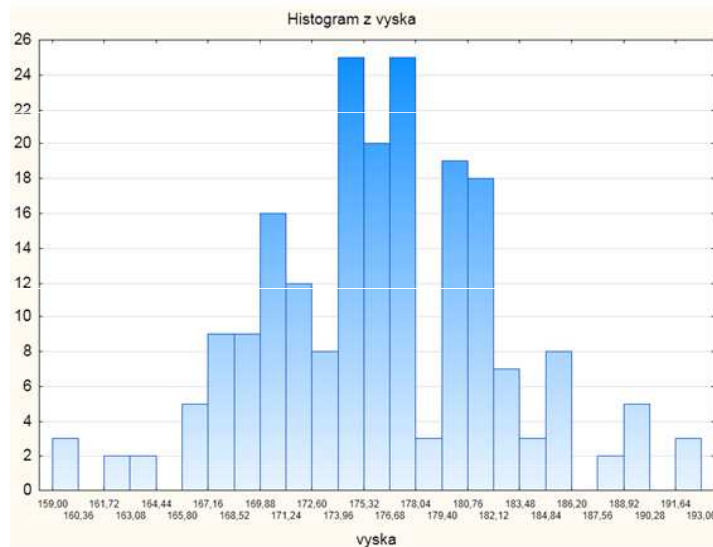
- V rozmezí $\mu \pm 3\sigma$ by se mělo vyskytovat 99,7 % všech hodnot



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

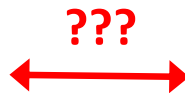
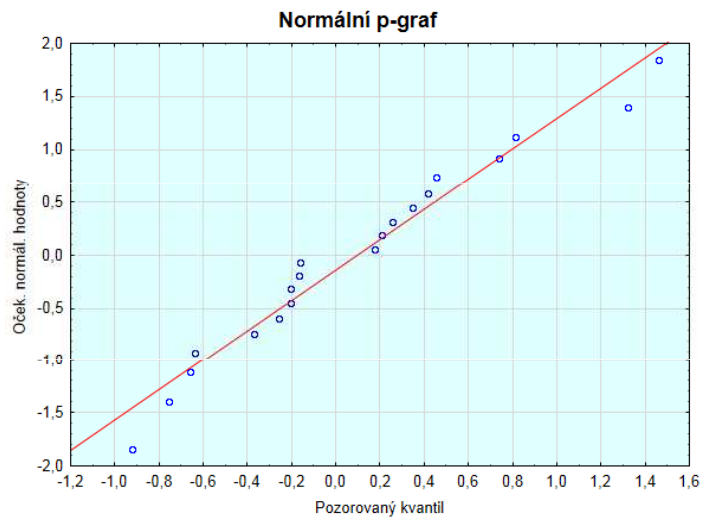
Vizuální ověření normality

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

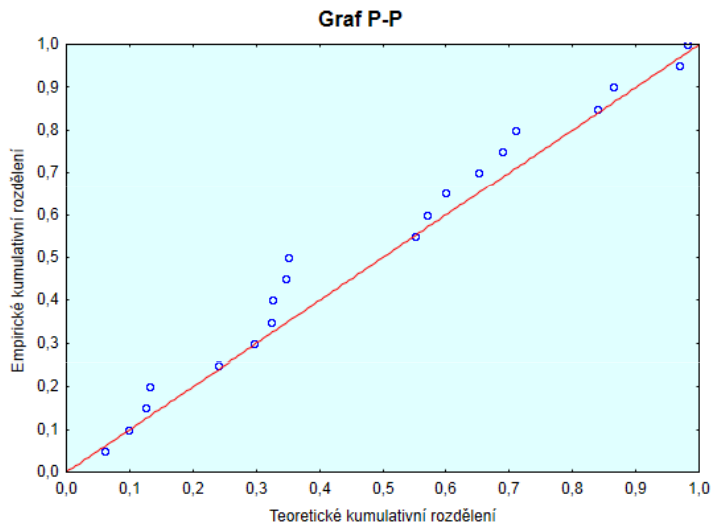
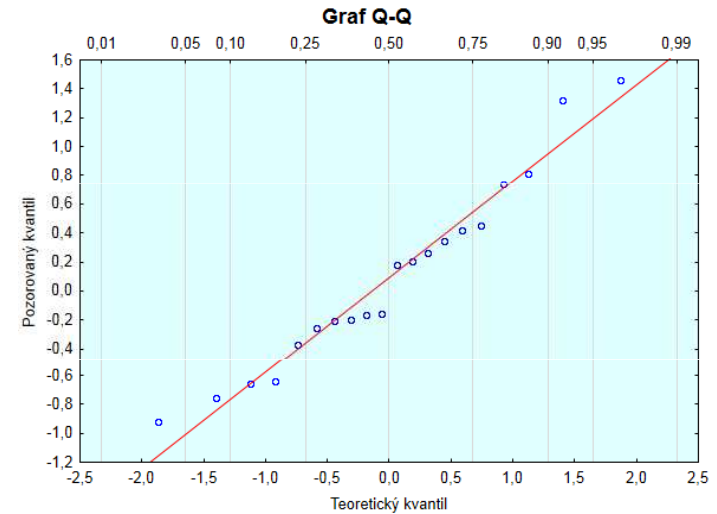


- Vhodnější jsou:
 - Q-Q graf (kvantil-kvantilový graf)
 - P-P graf (pravděpodobnostně-pravděpodobnostný graf)
 - N-P graf (normálně-pravděpodobnostný graf)

Rozdíl mezi N-P, Q-Q, P-P grafem



- Pouze výměna os
- Znázorněn pozorovaný a teoretický kvantil

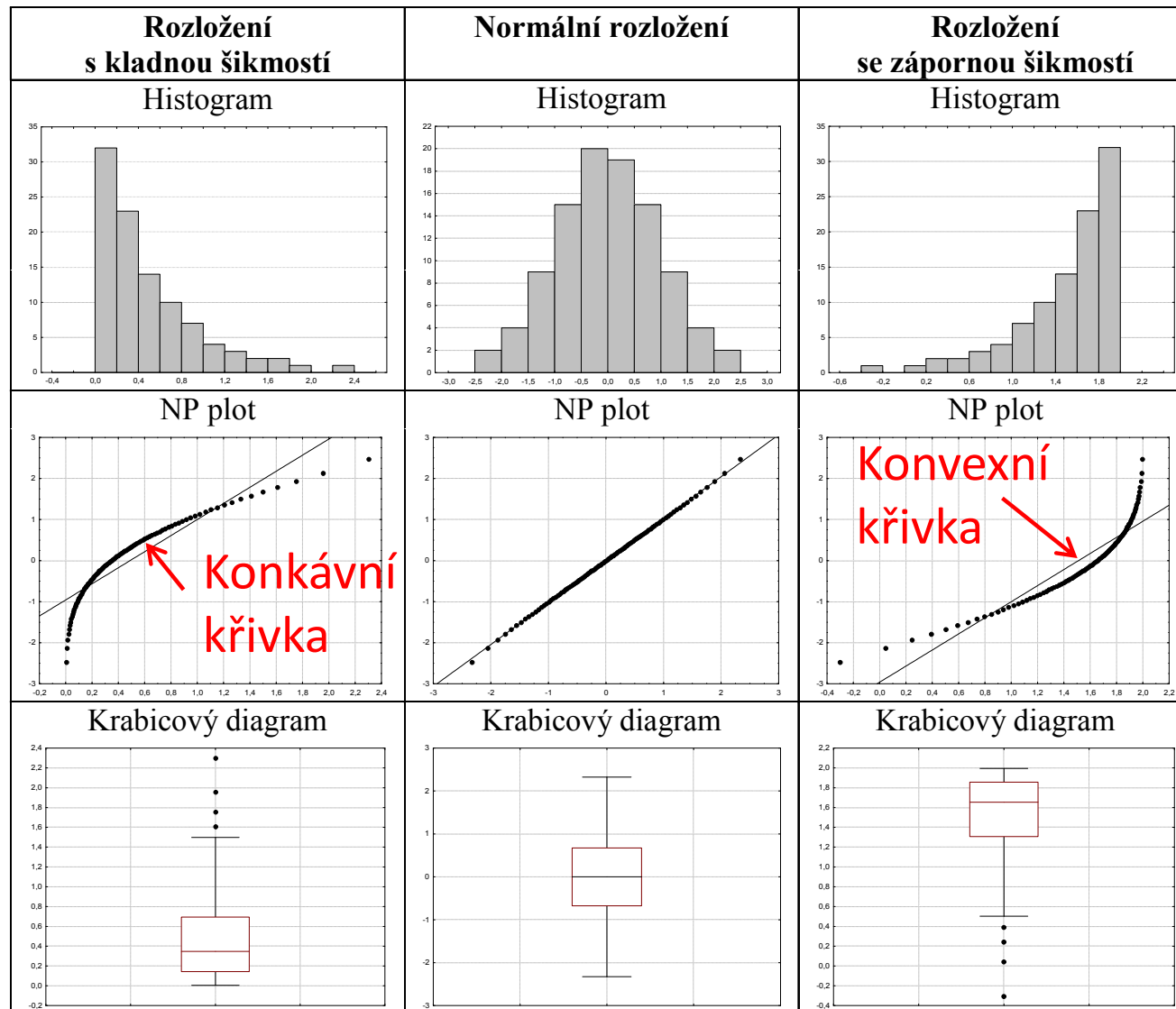


- Vykresleno kumulativní rozdělení



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

Asymetrie v diagnostických grafech



Výukové materiály:
Výpočetní statistika
Dr. Marie Budíková
2011

Základy testování hypotéz

Princip statistického testování hypotéz

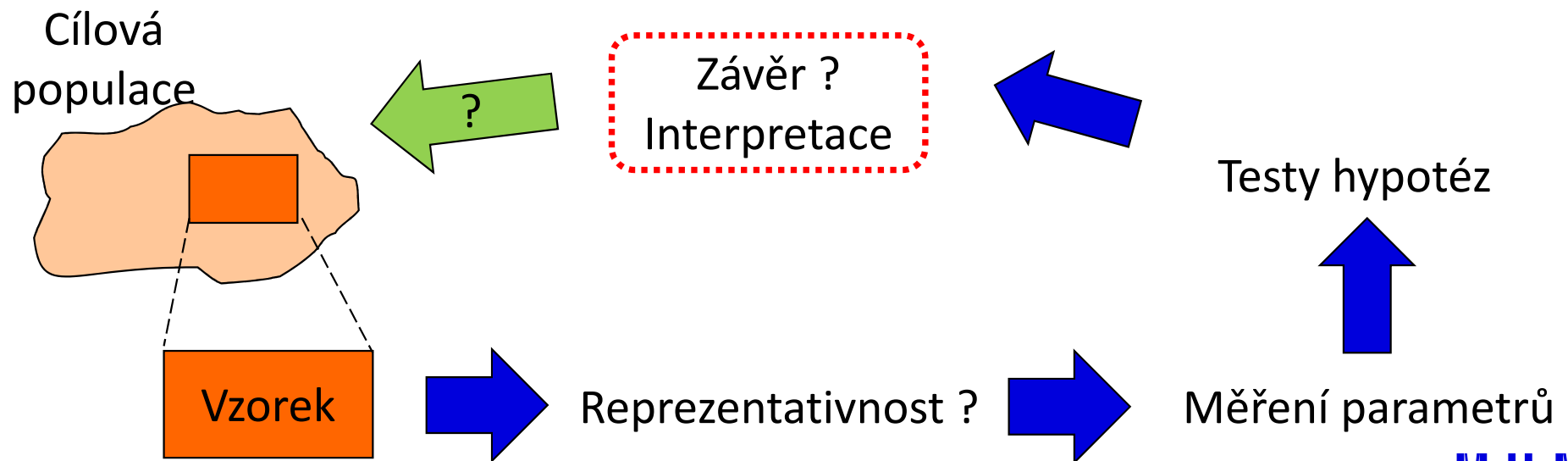
Pojmy statistických testů

Normalita dat a její význam pro testování

Ověření normality dat pomocí testu

Princip testování hypotéz

- Formulace hypotézy
- Výběr cílové populace a z ní reprezentativního vzorku
- Měření sledovaných parametrů
- Použití odpovídajícího testu → závěr testu
- Interpretace výsledků

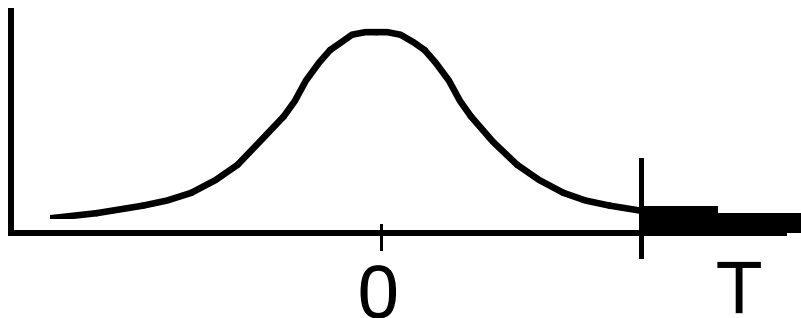


Statistické testování – základní pojmy

- Nulová hypotéza H_0 H_0 : sledovaný efekt je nulový
- Alternativní hypotéza H_A H_A : sledovaný efekt je různý mezi skupinami
- Testová statistika

$$\text{Testová statistika} = \frac{\text{Pozorovaná hodnota} - \text{Očekávaná hodnota}}{\text{Variabilita dat}} * \sqrt{\text{Velikost vzorku}}$$

- Kritický obor testové statistiky



Statistické testování odpovídá na otázku, zda je pozorovaný rozdíl náhodný či nikoliv. K odpovědi na otázku je využit statistický model – testová statistika.

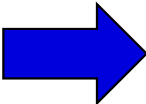
Možné chyby při testování hypotéz

- I přes dostatečnou velikost vzorku a kvalitní design experimentu se můžeme při rozhodnutí o (ne)zamítnutí nulové hypotézy dopustit chyby.

		Závěr testu	
		H_0 nezamítáme	H_0 zamítáme
Skutečnost	H_0 platí	<i>Správně</i> $1 - \alpha$	α Chyba I. druhu Falešně pozitivní závěr testu
	H_0 neplatí	β Chyba II. druhu Falešně negativní závěr testu	$1 - \beta$ <i>Správně</i>

Význam chyb při testování hypotéz

- **Pravděpodobnost chyby 1. druhu**

α  Pravděpodobnost nesprávného zamítnutí nulové hypotézy, **hladina významnosti**

- **Pravděpodobnost chyby 2. druhu**

β  Pravděpodobnost nerozpoznání neplatné nulové hypotézy

- **Síla testu**

$1-\beta$  Pravděpodobnostně vyjádřená schopnost rozpoznat neplatnost nulové hypotézy

Způsoby testování

- Testování H_0 proti H_A na hladině významnosti α můžeme provést třemi různými způsoby:
 - 1. Kritický obor** neboli obor zamítnutí H_0 ,
 - 2. Interval spolehlivosti,**
 - 3. P-hodnota** (vyjadřuje pravděpodobnost za platnosti H_0 , s níž bychom získali stejnou nebo extrémnější hodnotu testové statistiky).

Způsoby testování: P-hodnota

- Významnost hypotézy hodnotíme dle získané **p-hodnoty**, která vyjadřuje pravděpodobnost, s jakou číselné realizace výběru podporují H_0 , je-li pravdivá.
- P-hodnotu porovnáme s hladinou významnosti α (stanovujeme ji na 0,05, tzn. připouštíme 5% chybu testu, tedy, že zamítneme H_0 , ačkoliv ve skutečnosti platí).
- P-hodnotu získáme při testování hypotéz ve statistickém softwaru.

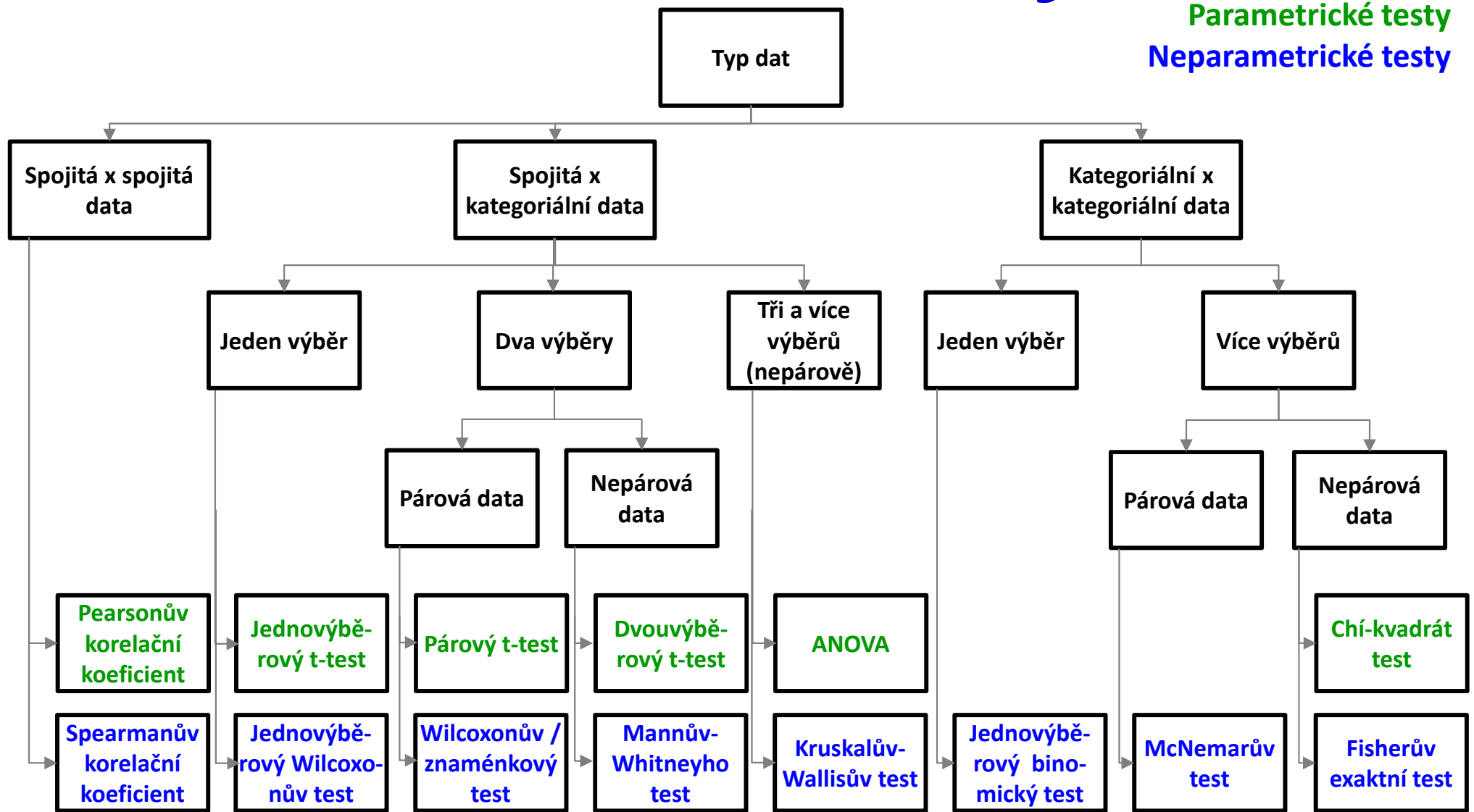
Je-li $p \leq \alpha$, pak H_0 zamítáme na hladině významnosti α a přijímáme H_A .

Je-li $p > \alpha$, pak H_0 nezamítáme na hladině významnosti α .

Poznámky k testování hypotéz

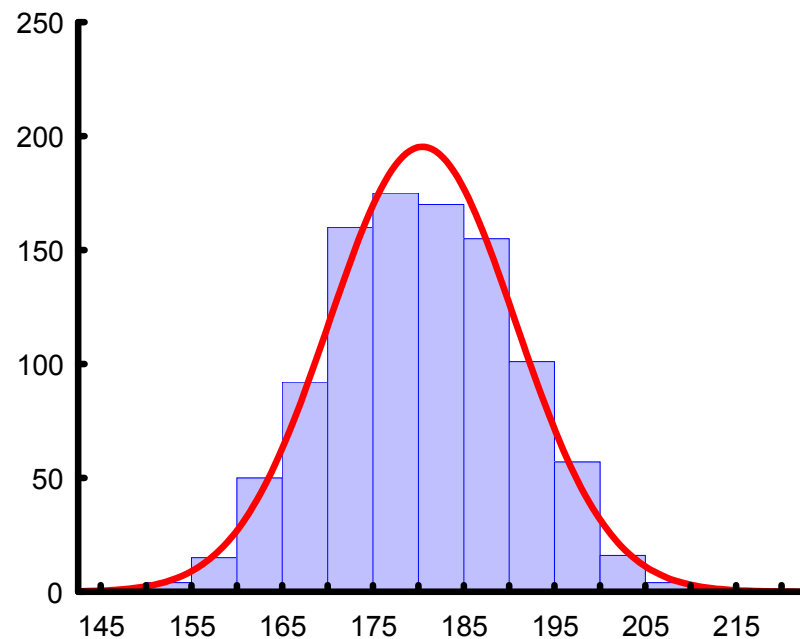
- **Nezamítnutí nulové hypotézy neznamena automaticky její přijetí!** Může se jednat o situaci, kdy pro zamítnutí nulové hypotézy nemáme dostatečné množství informace.
- **Dosažená hladina významnosti testu** (ať už 5 %, 1 % nebo 10 %) **nesmí být slepě brána jako hranice pro (ne)existenci testovaného efektu.**
- **Malá p-hodnota nemusí znamenat velký efekt.** Hodnota testové statistiky a p-hodnota mohou být ovlivněny velkou velikostí vzorku a malou variabilitou pozorovaných dat.
- **Na výsledky testování musí být nahlíženo kriticky** – jedná se o závěr založený „pouze“ na jednom výběrovém souboru.
- **Statistická významnost** indikuje, že pozorovaný rozdíl není náhodný, ale nemusí znamenat, že je významný i ve skutečnosti. Důležitá je i **praktická (klinická) významnost.**

Základní statistické testy



Testy normality

- Testy normality testují nulovou hypotézu, že není rozdíl mezi zpracovávaným rozložením a normálním rozložením. Vždy je ovšem dobré prohlédnout si i histogram, protože některé odchylky od normality, např. bimodalitu některé testy neodhalí.



Chí-kvadrát test dobré shody

Vhodný pro větší datové soubory. Srovnává pozorované četnosti s očekávanými hodnotami v třídách podobně jako při tvorbě histogramu.

Kolmogorovův - Smirnovův test

Často používaný test, zaměřuje se zejména na distribuční funkci. Častěji se používá v jeho modifikaci – Lilieforsův test.

Shapirův-Wilkův test

Jde o neparametrický test použitelný i při velmi malých n (10) s dobrou silou testu. Je zaměřen na testování symetrie.

**M U N I
M E D**

Praktické cvičení v programu Statistica



Datový soubor

Rehabilitace po mozkovém infarktu

Data: 02_Biostatistika_Data02.sta* (24v by 407c)

	Rehabilitace po mozkovém infarktu: data									
	1	2	3	4	5	6	7	8	9	10
	ID	Pohlavi	Vek	Etiologie	Lokalizace	Terapie	Komorbid	Barthel_inc	Kategorie_zavislosti_p	Ukoncen
1	1	muž	82	okluze nek	mozkové tepny	jiná farmakolog	0	25	vysoce závislý	propuště
2	2	žena	81	embolie	mozkové tepny	jiná farmakolog	2	20	vysoce závislý	přeložen
3	3	muž	55	okluze nek	mozkové tepny	jiná farmakolog	0	35	vysoce závislý	propuště
4	4	žena	46	embolie	mozkové tepny	intravenózní trc	0	20	vysoce závislý	propuště
5	5	muž	76	okluze nek	mozkové tepny	jiná farmakolog	0	45	částečně soběstačný	propuště
6	6	muž	72	okluze nek	mozkové tepny	jiná farmakolog	0	25	vysoce závislý	přeložen
7	7	muž	62	trombóza	mozkové tepny	jiná farmakolog	0	40	vysoce závislý	propuště
8	8	muž	64	trombóza	přívodní tepny	jiná farmakolog	0	15	vysoce závislý	propuště
9	9	žena	82	okluze nek	mozkové tepny	jiná farmakolog	0	10	vysoce závislý	přeložen
10	10	muž	58	trombóza	mozkové tepny	jiná farmakolog	0	25	vysoce závislý	propuště
11	11	muž	84	okluze nek	mozkové tepny	jiná farmakolog	0	40	vysoce závislý	propuště
12	12	žena	92	okluze nek	mozkové tepny	jiná farmakolog	0	30	vysoce závislý	propuště
13	13	žena	79	embolie	mozkové tepny	jiná farmakolog	1	40	vysoce závislý	propuště
14	14	muž	69	trombóza	mozkové tepny	jiná farmakolog	3	45	částečně soběstačný	propuště

Rehabilitace po mozkovém infarktu

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

Rehabilitace po mozkovém infarktu

Sbírané informace:

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

Úkol č. 1 – Normálně rozdělená data

Zadání: „Ověřte normalitu věku při mozkovém infarktu.“

Postup:

1. Srovnání průměru a mediánu (*Statistics – Basic Statistics – Descriptive Statistics – Advanced*)
2. Krabicový graf (*Graphs – 2D – Box Plots*)
3. Histogram (*Graphs – Histogram*)
4. Diagnostický N-P graf (*Graphs – 2D – Normal Probability Plots*)
5. Shapirův-Wilkův test nebo Lilieforsovy modifikace Kolmogorovova-Smirnovova testu (*lze provést např. těmito dvěma způsoby: 1) v nastavení histogramu: záložka Advanced → Statistics: vybereme test, 2) v nastavení N-P grafu: záložka: Quick → Statistics: zaškrtneme test*)

Úkol č. 1 – Řešení v programu Statistica

- V menu **Graphs** zvolíme **2D** a vybereme **Box Plots**.
- V menu **Graphs** zvolíme **Histogram**
- V menu **Graphs** zvolíme **2D** a vybereme **Normal Probability Plots**, na záložce **Quick** zaškrtneme test

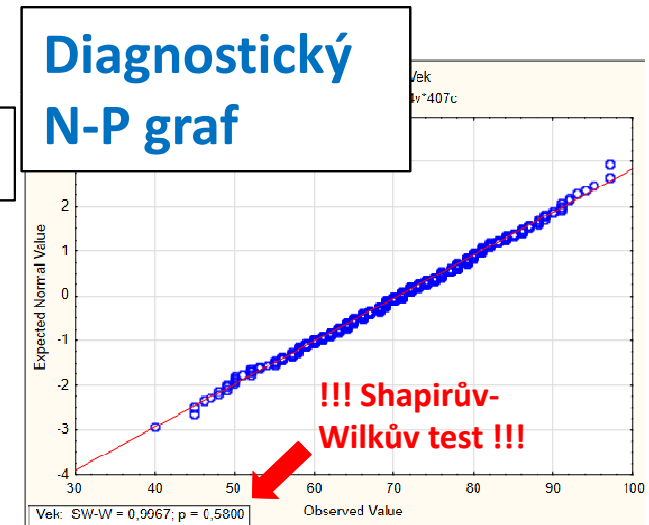
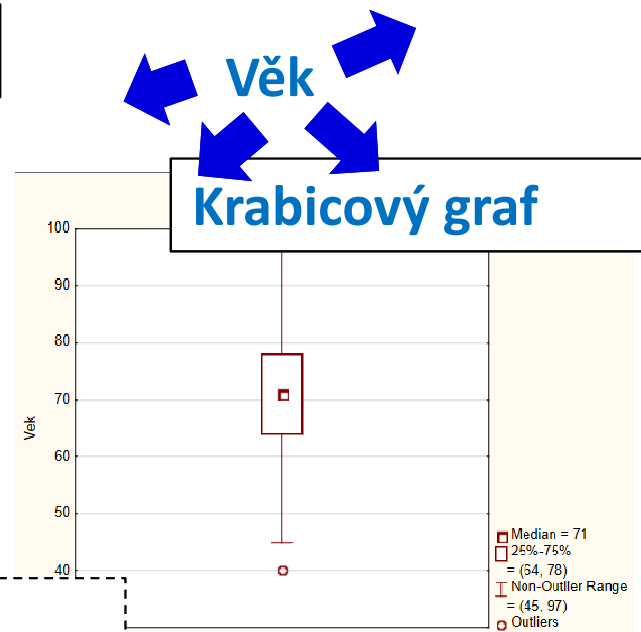
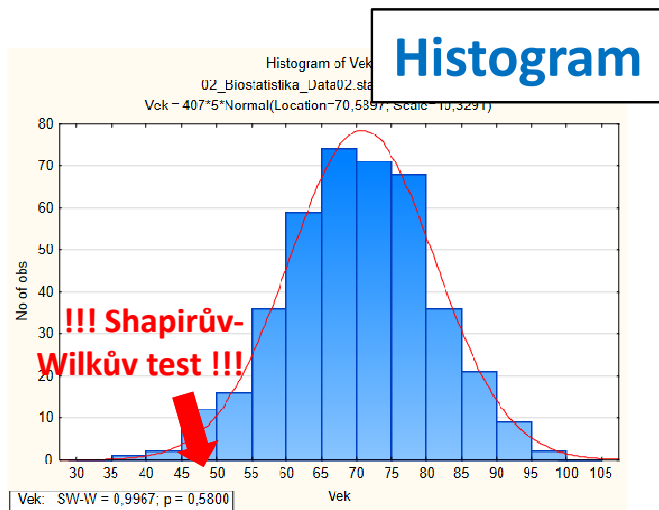
The image shows a screenshot of the Statistica software interface. The 'Graphs' menu is open, showing options like '2D', 'Histograms...', 'Box Plots...', and 'Normal Probability Plots...'. Red arrows with numbers 1, 2, and 3 point to these menu items. Below the menu, the 'Normal Probability Plots' dialog box is shown with the 'Quick' tab selected. A red arrow with the number 4 points to the 'Shapiro-Wilk test' checkbox, which is checked. The dialog box also shows 'Normal' as the selected graph type and 'Do not assign average ranks to tied observations' as checked.

Úkol č. 1 – Výsledky v Statistica

① Průměr a medián jsou téměř shodné (cca 71 let) a data jsou tedy nejspíš alespoň symetrická.

Srovnání průměru a mediánu

Variable	Descriptive Statistics (02_Biostatistik)		
	Valid N	Mean	Median
Věk	407	70,58968	71,00000



② Symetrie je patrná i z krabicového grafu. Navíc histogram naprosto jasně odpovídá průběhu normálního rozdělení. Z N-P grafu také nejsou patrné odchylky od normality.

③ Na základě p-hodnoty 0,580 nezamítáme nulovou hypotézu o normalitě (tj. nezamítáme, že není rozdíl mezi pozorovanými daty a teoretickým normálním rozdělením, ... tj. data jsou normálně rozdělená).

Úkol č. 2 – Odlehlá/chybná hodnota

Zadání: „Ověřte normalitu věku při mozковém infarktu obsahující jeden překlep 40 → 400.“

Postup (*přepište hodnotu 40 na 400 a ke stanovení závěru opět použijte vybrané nástroje vhodné pro ověření normality*):

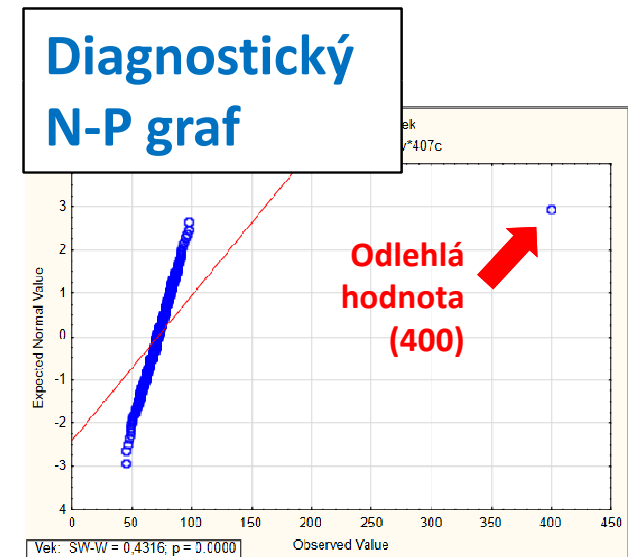
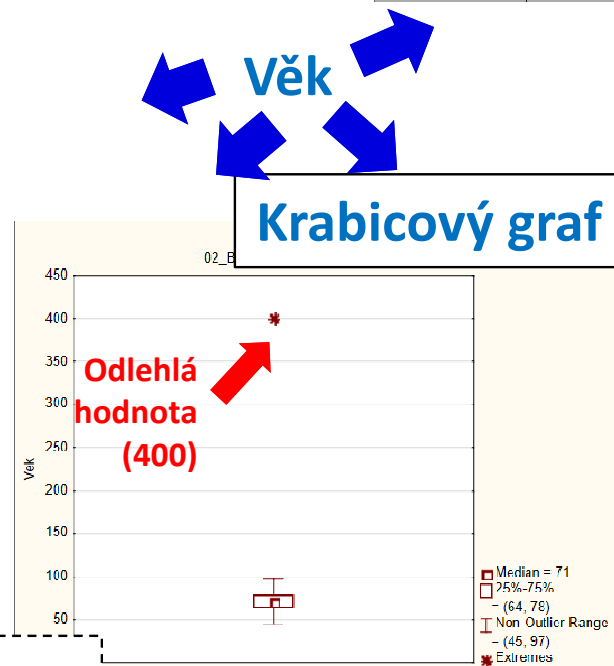
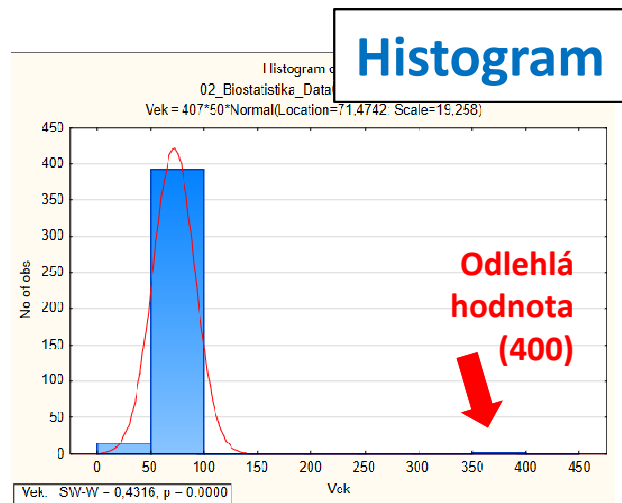
1. Srovnání průměru a mediánu (*Statistics – Basic Statistics – Descriptive Statistics – Advanced*)
2. Krabicový graf (*Graphs – 2D – Box Plots*)
3. Histogram (*Graphs – Histogram*)
4. Diagnostický N-P graf (*Graphs – 2D – Normal Probability Plots*)
5. Shapirův-Wilkův test nebo Lilieforsovy modifikace Kolmogorovova-Smirnovova testu (*lze provést např. těmito dvěma způsoby: 1) v nastavení histogramu: záložka Advanced → Statistics: vybereme test, 2) v nastavení N-P grafu: záložka: Quick → Statistics: zaškrtneme test*)

Úkol č. 2 – Výsledky v Statistica

① Průměr a medián jsou stále podobné (cca 71 let) a data by tedy mohla být alespoň symetrická.

Srovnání průměru a mediánu

Variable	Descriptive Statistics (02_Biostatistik)		
	Valid N	Mean	Median
Věk	407	71,47420	71,00000



② Ze všech tří grafických nástrojů lze identifikovat výskyt odlehlé/chybné hodnoty, jejíž přítomnost zkresluje pohled na zbytek souboru.

③ Na základě p-hodnoty $< 0,001$ zamítáme nulovou hypotézu o normalitě (tj. zamítáme, že není rozdíl mezi pozorovanými daty a teoretickým normálním rozdělením, ... tj. data nejsou normálně rozdělená).

Úkol č. 3 – Asymetrická data

Zadání: „ Ověřte normalitu indexu Barthelové (vyjadřuje stupeň soběstačnosti v základních denních aktivitách) na konci akutní hospitalizační péče o pacienty s mozkovým infarktem.“

Postup:

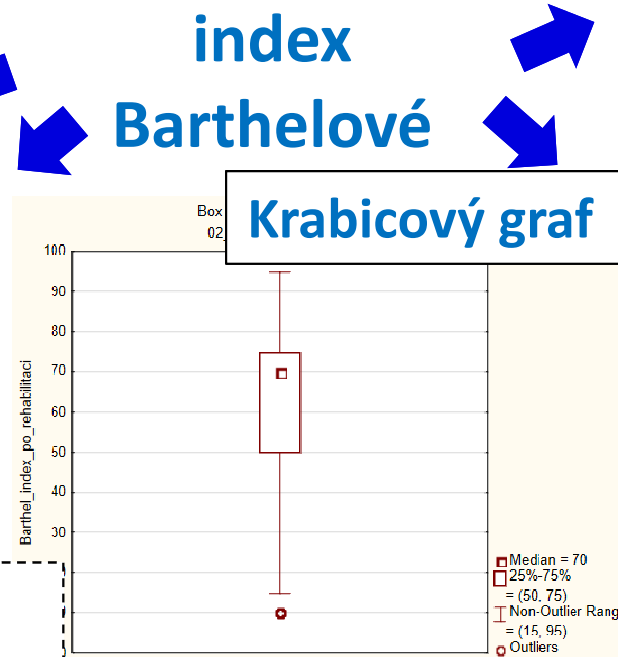
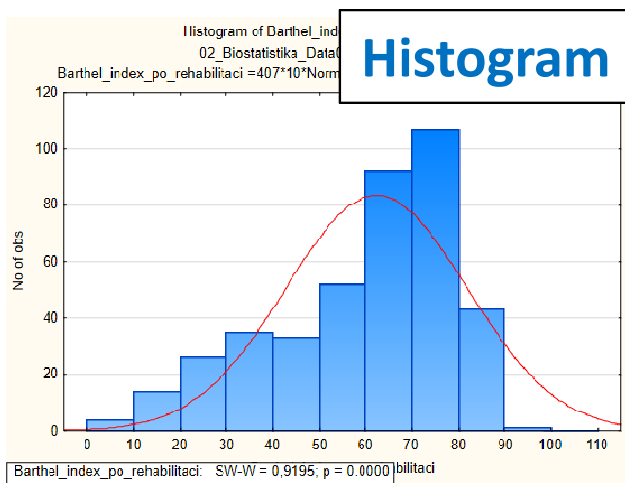
1. Srovnání průměru a mediánu (*Statistics – Basic Statistics – Descriptive Statistics – Advanced*)
2. Krabicový graf (*Graphs – 2D – Box Plots*)
3. Histogram (*Graphs – Histogram*)
4. Diagnostický N-P graf (*Graphs – 2D – Normal Probability Plots*)
5. Shapirův-Wilkův test nebo Lilieforsovy modifikace Kolmogorovova-Smirnovova testu (*Ize provést např. těmito dvěma způsoby: 1) v nastavení histogramu: záložka Advanced → Statistics: vybereme test, 2) v nastavení N-P grafu: záložka: Quick → Statistics: zaškrtneme test*)

Úkol č. 3 – Výsledky v Statistica

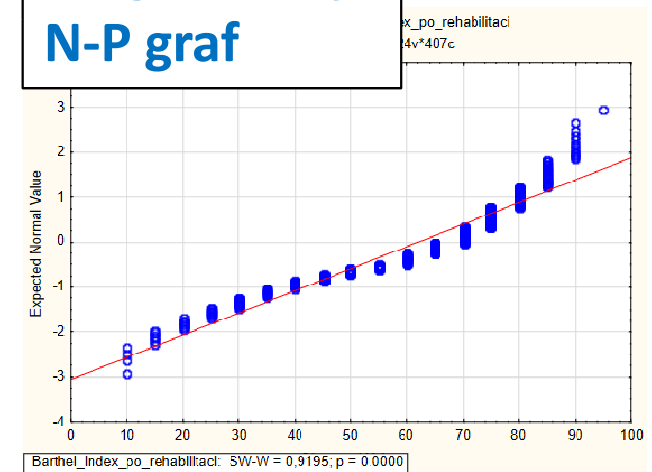
① Průměr a medián se výrazně liší (průměr 62 bodů, medián 70 bodů), což znamená, že data jsou nejspíše asymetrická.

Srovnání průměru a mediánu

Variable	Descriptive Statistics (02_Biostatistik)		
	Valid N	Mean	Median
Barthel_index_po_rehabilitaci	407	62,01474	70,00000



Diagnostický N-P graf



② Asymetrie je patrná i z krabicového grafu a histogramu. Z histogramu je navíc zřetelně vidět odlišnost od normálního rozdělení. Odchytky od normality jsou patrné i z N-P grafu.

③ Na základě p-hodnoty $< 0,001$ zamítáme nulovou hypotézu o normalitě (tj. zamítáme, že není rozdíl mezi pozorovanými daty a teoretickým normálním rozdělením, ... tj. data nejsou normálně rozdělená).