evropský sociální fond v ČR · EVROPSKÁ UNIE · MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY · OP Vzdělávání pro konkurenceschopnost

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

PRŮZKUMOVÁ ANALÝZA DAT (EDA)

Vytvořeno s podporou projektu Průřezová inovace studijních programů Lesnické a dřevařské fakulty MENDELU v Brně (LDF) s ohledem na discipliny společného základu (reg. č. CZ.1.07/2.2.00/28.0021) za přispění finančních prostředků EU a státního rozpočtu České republiky.

POSTUP STATISTICKÉ ANALÝZY JEDNOROZMĚRNÝCH DAT

1. Průzkumová analýza dat (EDA)

- posouzení stupně symetrie a špičatosti dat
- nalezení „podezřelých" (odlehlých) dat
- ověření normality rozdělení
- ověření nezávislosti prvků výběru (autokorelace)

2. Odhady parametrů základního souboru

- výpočet výběrových statistik (momentových nebo robustních)
- (výpočet statistik pomocí transformace pro nenormální rozdělení - podle potřeby)
- výpočet bodových odhadů parametrů
- výpočet intervalových odhadů parametrů

3. Testování statistických hypotéz

- formulace hypotéz
- rozhodnutí o zamítnutí nebo nezamítnutí posuzované hypotézy
- (analýza síly testu - podle potřeby)

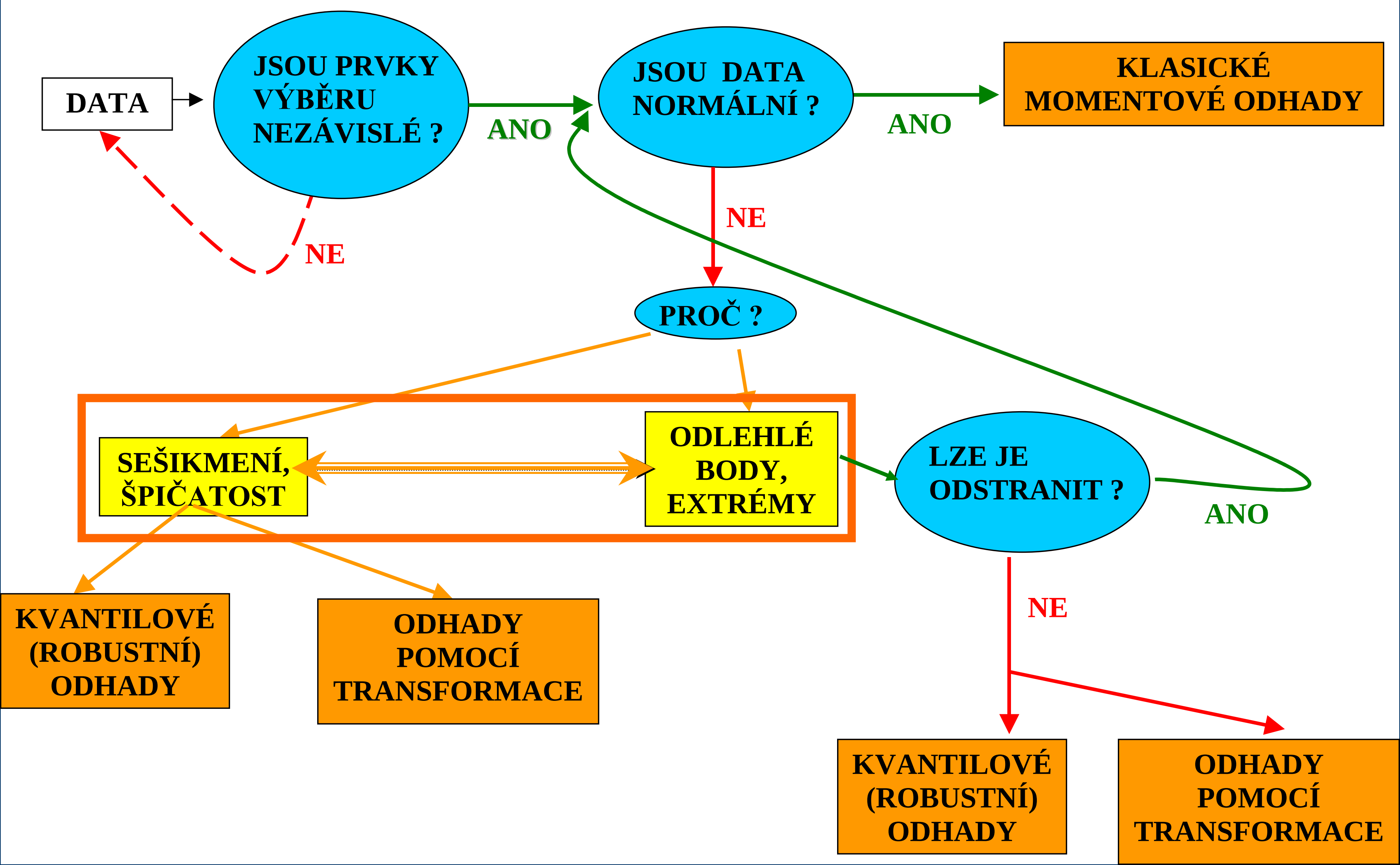
POSTUP ODHADU PARAMETRŮ JEDNOROZMĚRNÝCH DAT

Následující schéma ukazuje nejdůležitější uzlové body analýzy dat. Abychom mohli použít „klasickou“ metodu odhadu parametrů, tj. vypočítat aritmetický průměr, směrodatnou odchylku a další charakteristiky z nich odvozené, musí data splňovat následující předpoklady:

- **data musí být vzájemně nezávislá**
- **musí pocházet ze základního souboru s normálním rozdělením**
- **neměla by obsahovat extrémní body (hodnoty velmi vzdálené od ostatních)**

Proto **nejdříve** (než začneme „cokoliv“ počítat) **musíme ověřit, zda jsou nebo nejsou tyto podmínky splněny** a podle toho zvolit vhodnou metodu odhadu – **použít průzkumovou analýzu dat**.

POSTUP ODHADU PARAMETRŮ JEDNOROZMĚRNÝCH DAT



POSTUP ODHADU PARAMETRŮ JEDNOROZMĚRNÝCH DAT

Nezávislost znamená, že v datech (v tom pořadí jak byla naměřena) není žádný trend (např. data stále stoupající nebo klesající nebo vykazující jinou závislost). Pokud tomu tak není (trend existuje), znamená to, že nebyly splněny podmínky náhodného výběru (jedna ze základních podmínek matematické statistiky). Přísně vzato, taková data by se neměla používat k další analýze a měla by bát naměřena jiná. Vzhledem k tomu, že data jsou často „drahá“ a vzácná, tak se takováto data obvykle používají (proto je na obrázku zpětná šipka „NE“ čárkovaná), ale s vědomím, že jejich získání nebylo ideální, což se musí zohlednit při interpretaci výsledků analýzy a především by se měla odhalit příčina trendu v datech a způsob výběru podle toho korigovat.

POSTUP ODHADU PARAMETRŮ JEDNOROZMĚRNÝCH DAT

Normalita dat

Základní momentové statistické charakteristiky jsou konstruovány na základě předpokladu normálního rozdělení dat. Pokud je tato podmínka splněna, můžeme použít klasické momentové odhady (aritmetický průměr a veličiny z něho odvozené, např. směrodatnou odchylku).

Pokud tomu tak není, musíme nejprve analyzovat hlavní příčinu nenormálního rozdělení (odpověď na otázku „PROČ?").

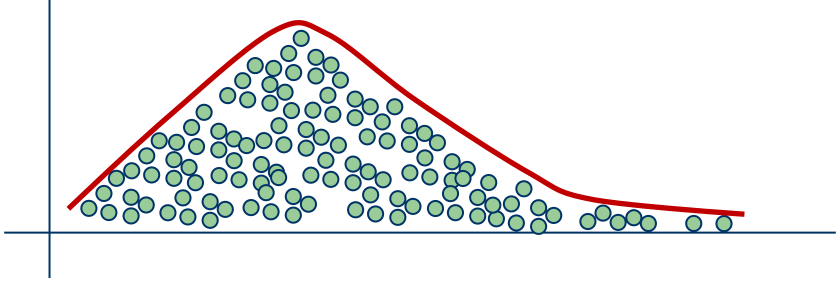
Nejčastěji jsou možné příčiny dvě (a různé stupně jejich kombinace):

- **sešikmení dat (levostranné nebo pravostranné rozdělení nebo špičaté nebo ploché rozdělení)**
- **extrémní hodnoty**

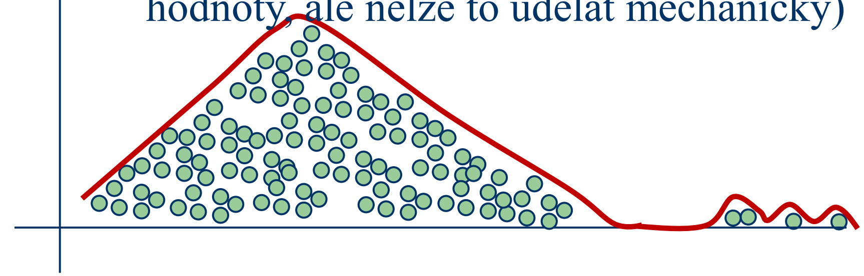
POSTUP ODHADU PARAMETRŮ JEDNOROZMĚRNÝCH DAT

Dva hlavní typy dat nepocházejících z normálního rozdělení

sešikmený soubor (v tomto případě **levostranný)**
většina hodnot je koncentrována nalevo, žádný bod ale není výrazně vzdálený od ostatních, žádný bod není možné vypustit)



sešikmený soubor (v tomto případě **levostranný) s extrémy**
zde je levostrannost způsobena vzdálenými extrémními body napravo, hlavní skupina bodů nalevo je v podstatě symetrická. Musíme uvažovat o možnosti vypustit z analýzy extrémní hodnoty, ale nelze to udělat mechanicky)



Mezi těmito dvěma možnostmi existuje celá řada přechodů!! Proto je vždy nutné pečlivě zvážit, co je hlavní příčinou nenormálního rozdělení dat. K tomu slouží hlavně grafické metody průzkumové analýzy dat

POSTUP ODHADU PARAMETRŮ JEDNOROZMĚRNÝCH DAT

Jak „zacházet“ s extrémními hodnotami

Extrémní hodnoty musíme posuzovat výhradně v kontextu jejich vypovídací hodnoty a správnosti jejich stanovení. **Okamžitě můžeme vyloučit pouze zjevné hrubé chyby** (způsobené např. chybným měřením, zápisem apod. – např. místo výšky stromu 20 m máme v souboru 200 m). **Pokud jsou hodnoty naměřené spolehlivě** (nepřijdeme na žádnou příčinu hrubé chyby a daná hodnota je „možná“), takové hodnoty **nemůžeme vylučovat z analýzy a naopak mohou mít vysokou vypovídací schopnost** (mohou být „cennější“ než „běžná“ data – např. záznam o extrémních hodnotách v souboru, který zachycuje znečistění ovzduší). Potom musíme použít jiné metody odhadu parametrů – kvantily nebo transformace – viz schéma na snímku 2).

Pokud vyloučíme extrémní hodnoty jako hrubé chyby, znovu musíme testovat, zda „zbylý“ soubor pochází z normálního rozdělení.

POSTUP ODHADU PARAMETRŮ JEDNOROZMĚRNÝCH DAT

Ke zjištění důležitých vlastností analyzovaných souborů využijeme metod průzkumové analýzy dat.
Na základě jejích výsledků rozhodneme, zda použijeme momentové odhady (v případě potvrzení základních podmínek) nebo kvantilové (resp. transformační) odhady (v případě jejich nedodržení).

PRŮZKUMOVÁ ANALÝZA DAT (EDA)

EDA – Exploratory Data Analysis (Tuckey, Chambers)

Cílem **průzkumové analýzy dat je nalezení zvláštností statistického chování dat a ověření jejich předpokladů pro následné statistické zpracování „klasickými“ statistickými metodami.**

Hlavní zvláštnosti chování dat	Základní předpoklady
nesymetrie (levostranné – pravostranné)	shoda s teoretickým rozdělením (obvykle normálním)
lokální koncentrace dat (špičatost – plochost)	potřebná velikost výběru
extrémní data	nezávislost dat

METODY EDA

Grafické:

- **graf rozptýlení hodnot**
- **krabicový graf**
- **vrubový krabicový graf**
- **kvantil-kvantilový graf**
- **histogram**
- **graf hustoty pravděpodobnosti**

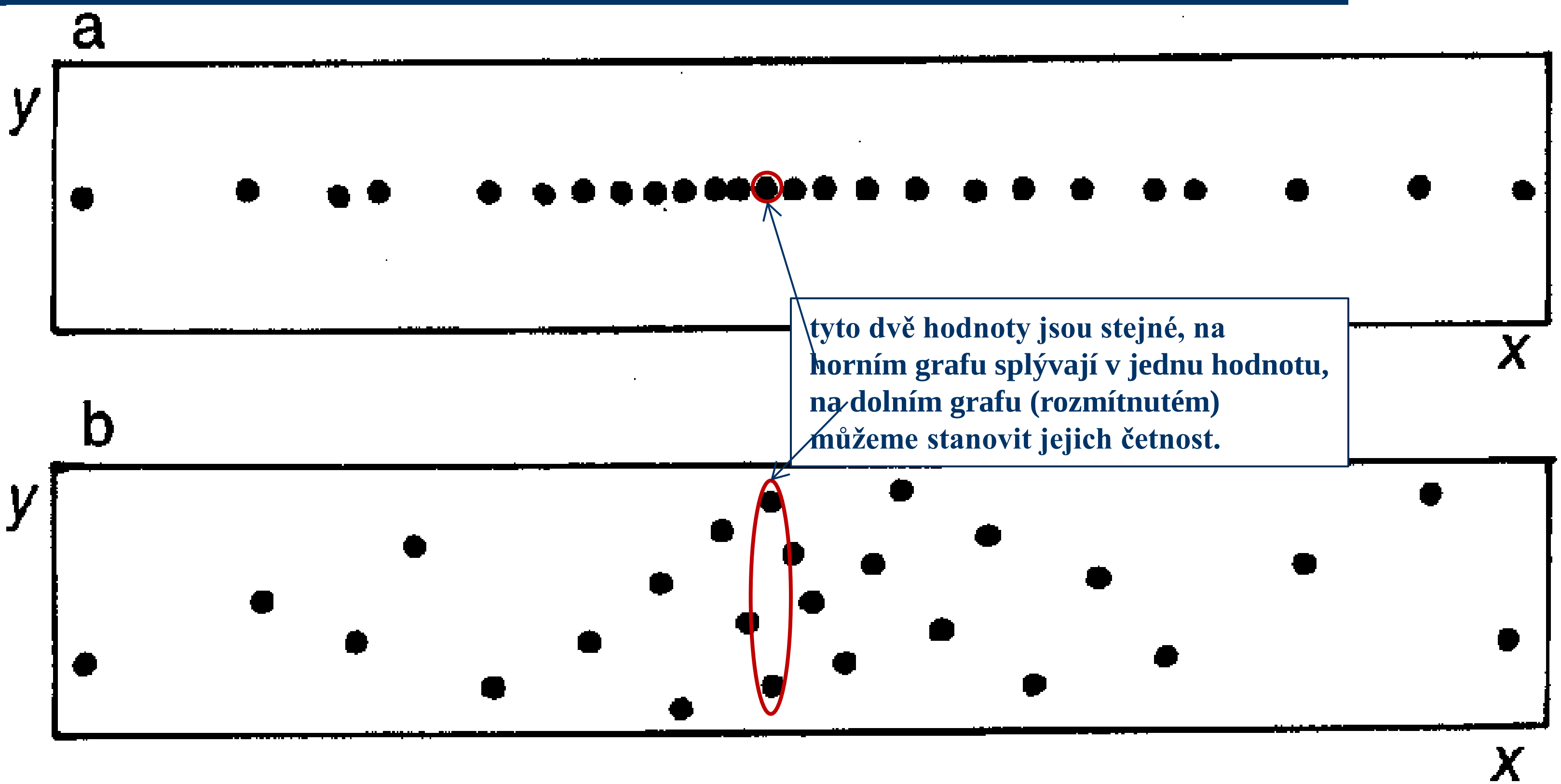
Testy a početní metody:

- **testy shody (normality)**
- **test nezávislosti dat**
- **stanovení minimální velikosti výběru**

Grafické a testové metody se doplňují, **proto by měly být používány společně**:

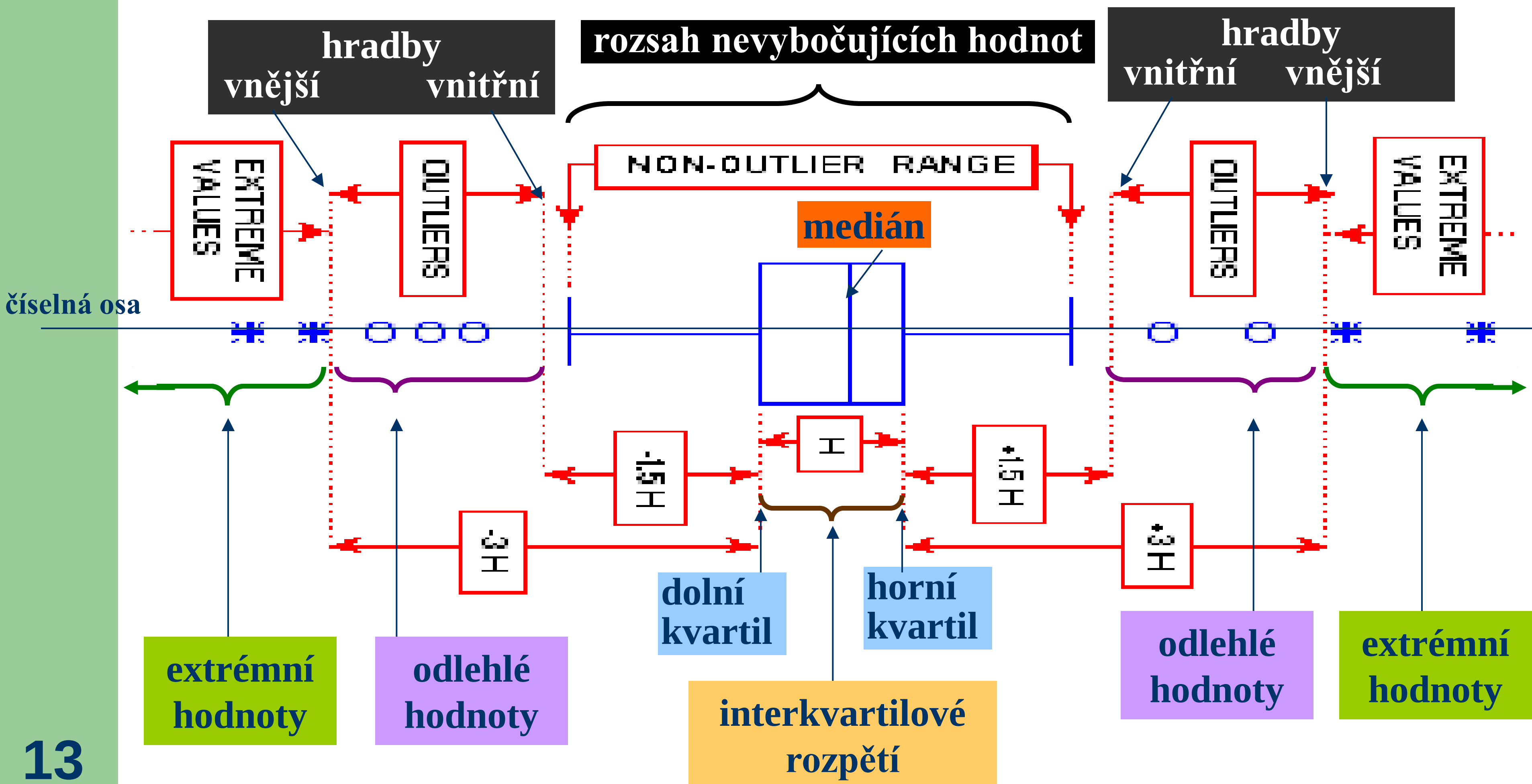
- **testové metody odpovídají na otázku zda je splněna daná podmínka (např. pochází nebo nepochází daný výběr ze základního souboru s normálním rozdělením – ano nebo ne?**
- **grafické metody odpovídají na otázku, proč tato podmínka není splněna (např. co je příčinou, že rozdělení dat není normální)**

GRAF ROZPTÝLENÍ HODNOT



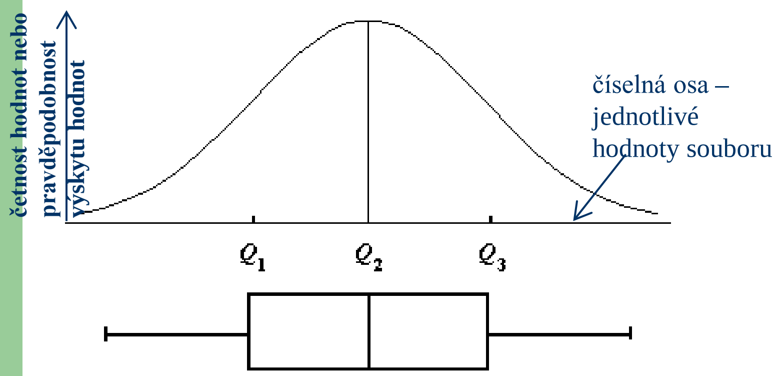
Prosté vynesení bodů na číselnou osu. Dolní variantě říkáme rozmítnutý graf rozptýlení a je výhodný v tom, že stejné hodnoty se nepřekrývají a můžeme stanovit jejich četnost.

KRABICOVÝ GRAF

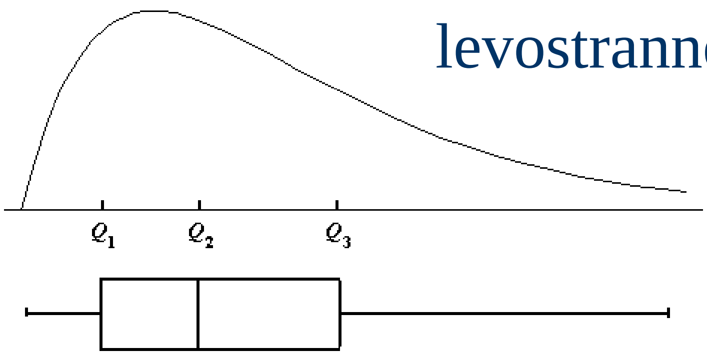


KRABICOVÝ GRAF

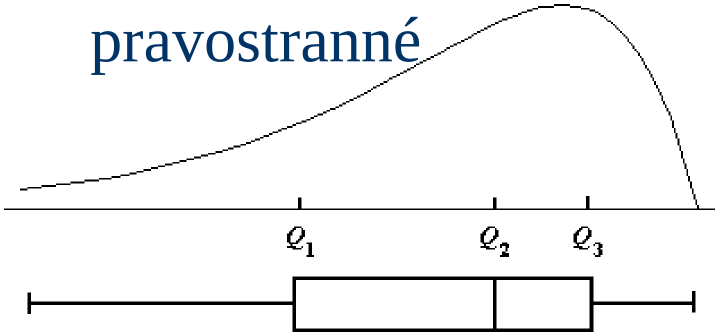
Typické tvary krabicových grafů pro základní tvary rozdělení



normální



levostranné



pravostranné

Q_1 – dolní kvartil

Q_2 – medián

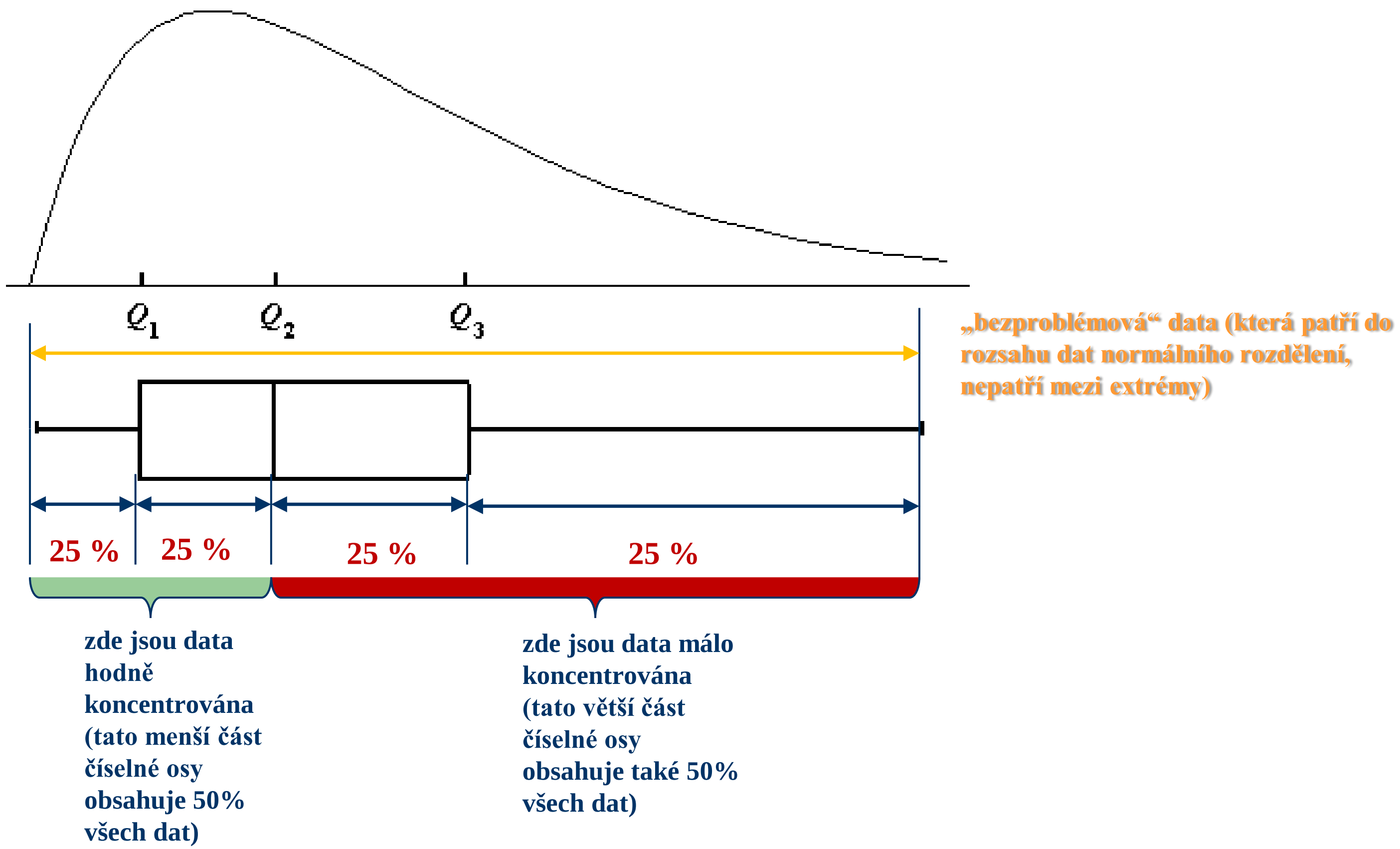
Q_3 – horní kvartil

KRABICOVÝ GRAF

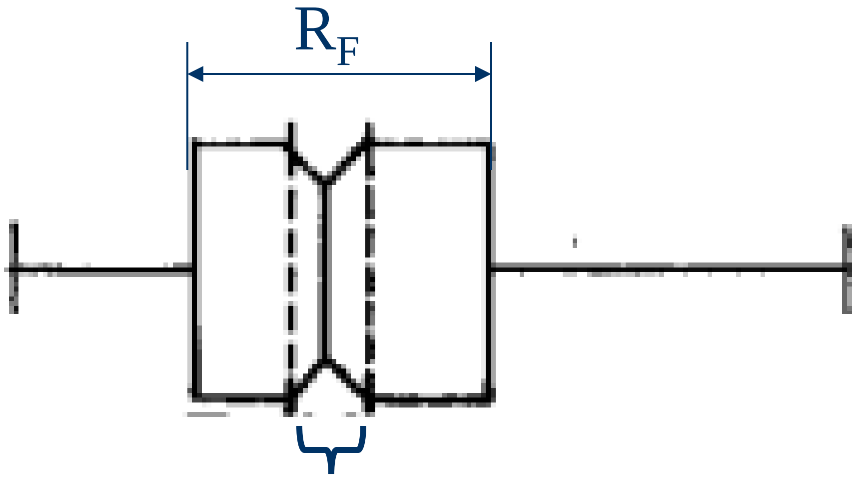
Z krabicového grafu můžeme získat tyto hlavní informace:

- **zda soubor obsahuje extrémní hodnoty (jsou vyznačeny samostatnými značkami)**
- **jaký je rozsah „bezproblémových“ hodnot (rozsah „fousů“ včetně krabičky – žlutá šipka na následujícím obrázku)**
- **jak jsou data v souboru rozložena (každá část krabicového grafu – dolní „fous“,dolní část „krabičky“, horní část „krabičky“, horní „fous“ – ukazuje, v jakém intervalu číselné osy se nachází 25 % dat – tedy čím je příslušná část užší, tím jsou data v daném úseku více koncentrována – viz následující obrázek)**

KRABICOVÝ GRAF



VRUBOVÝ KRABICOVÝ GRAF

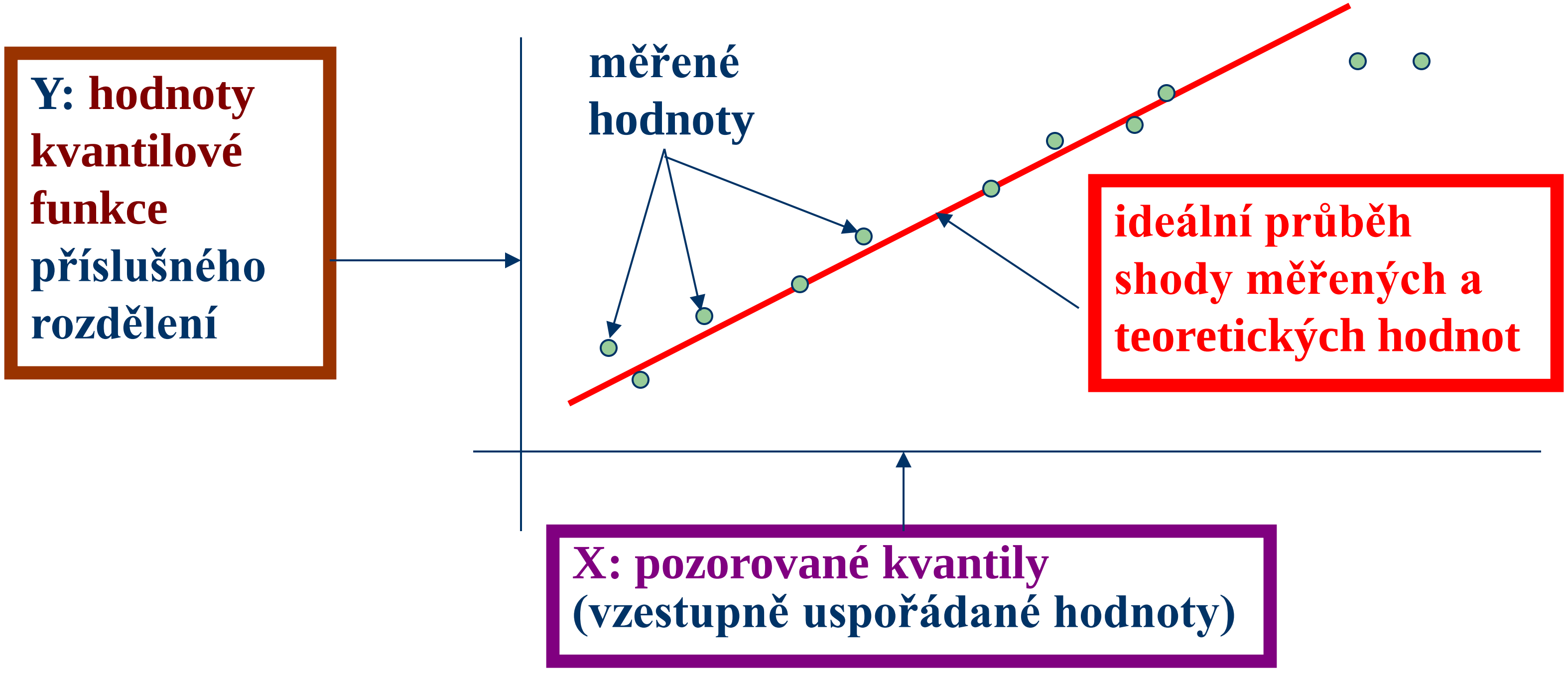


intervalový odhad mediánu

$$I_{D,H} = M \pm \frac{1,57 \cdot R_F}{\sqrt{n}}$$

KVANTIL-KVANTILOVÝ GRAF (Q-Q GRAF)

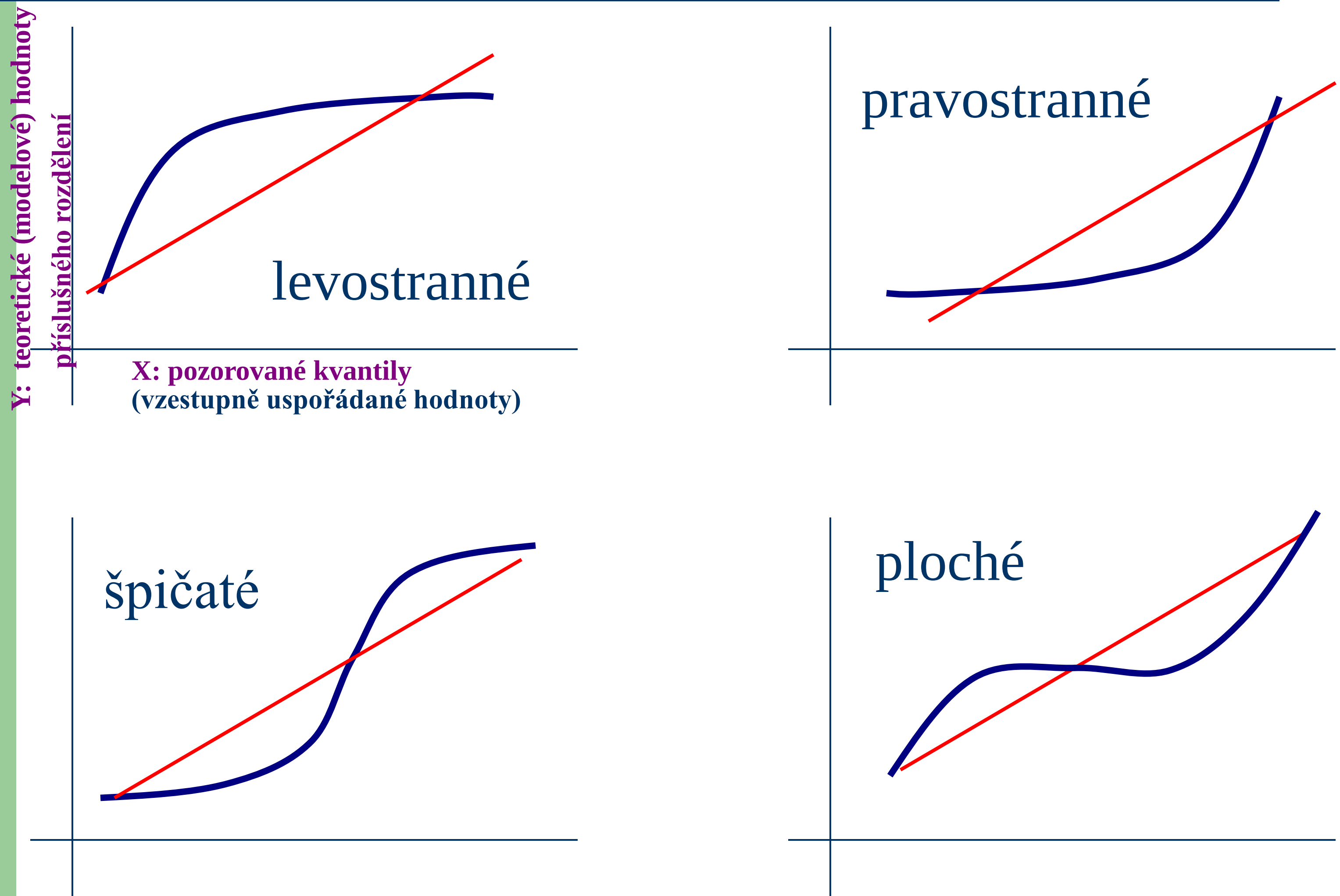
nejlepší grafická metoda pro posouzení shody měřených hodnot s daným rozdělením



KVANTIL-KVANTILOVÝ GRAF (Q-Q GRAF)

QQ grafy se používají pro srovnání měřených hodnot s jakýmkoliv rozdělením, jehož matematický model známe.
Nejčastější porovnání je s rozdělením normálním. V tomto případě se jako teoretické hodnoty vynášejí hodnoty normovaného normálního rozdělení (k tomuto rozdělení viz **teorie text I, 73-78 a prezentace „rozdělení“**).

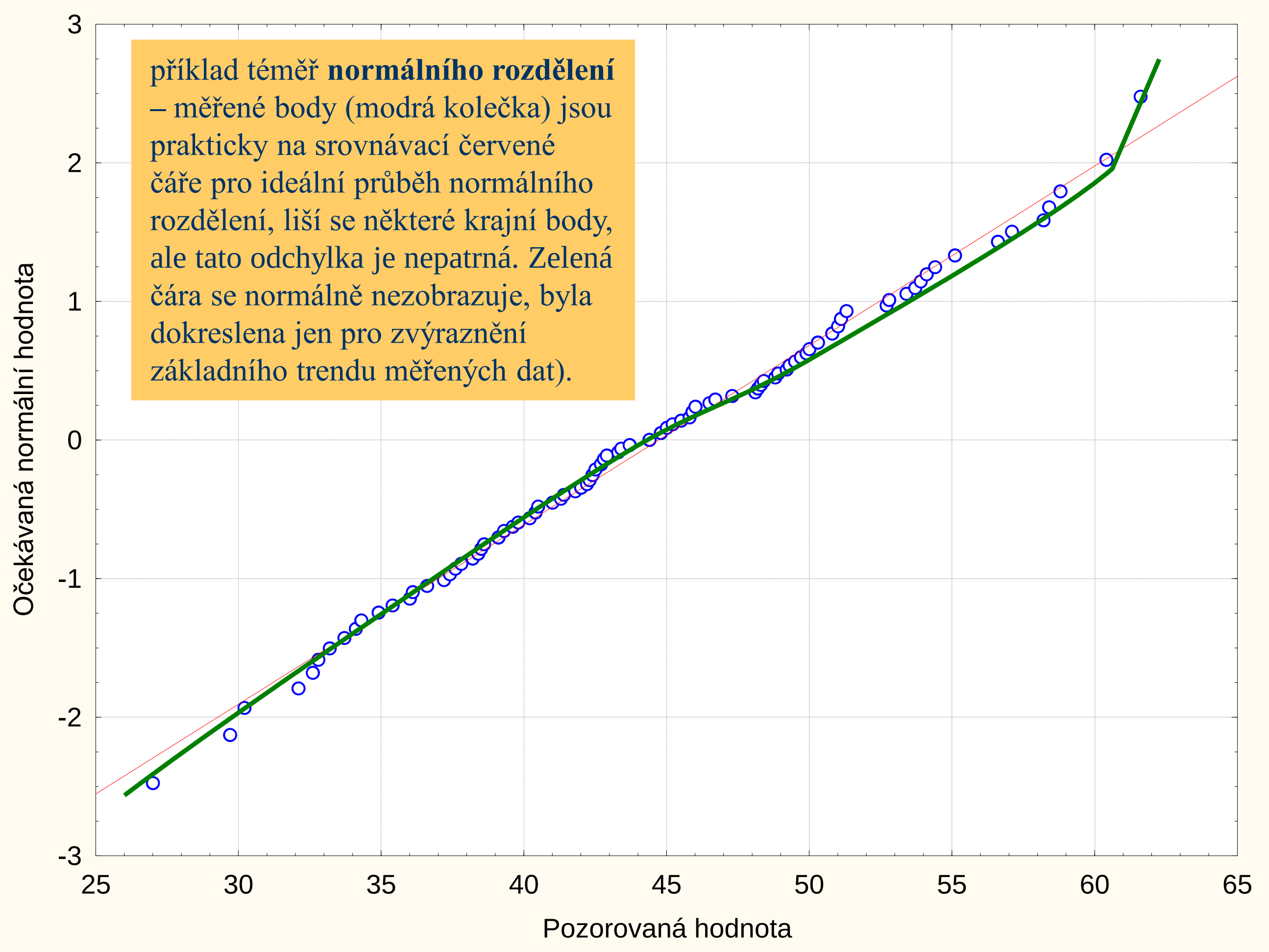
KVANTIL-KVANTILOVÝ GRAF (Q-Q GRAF)



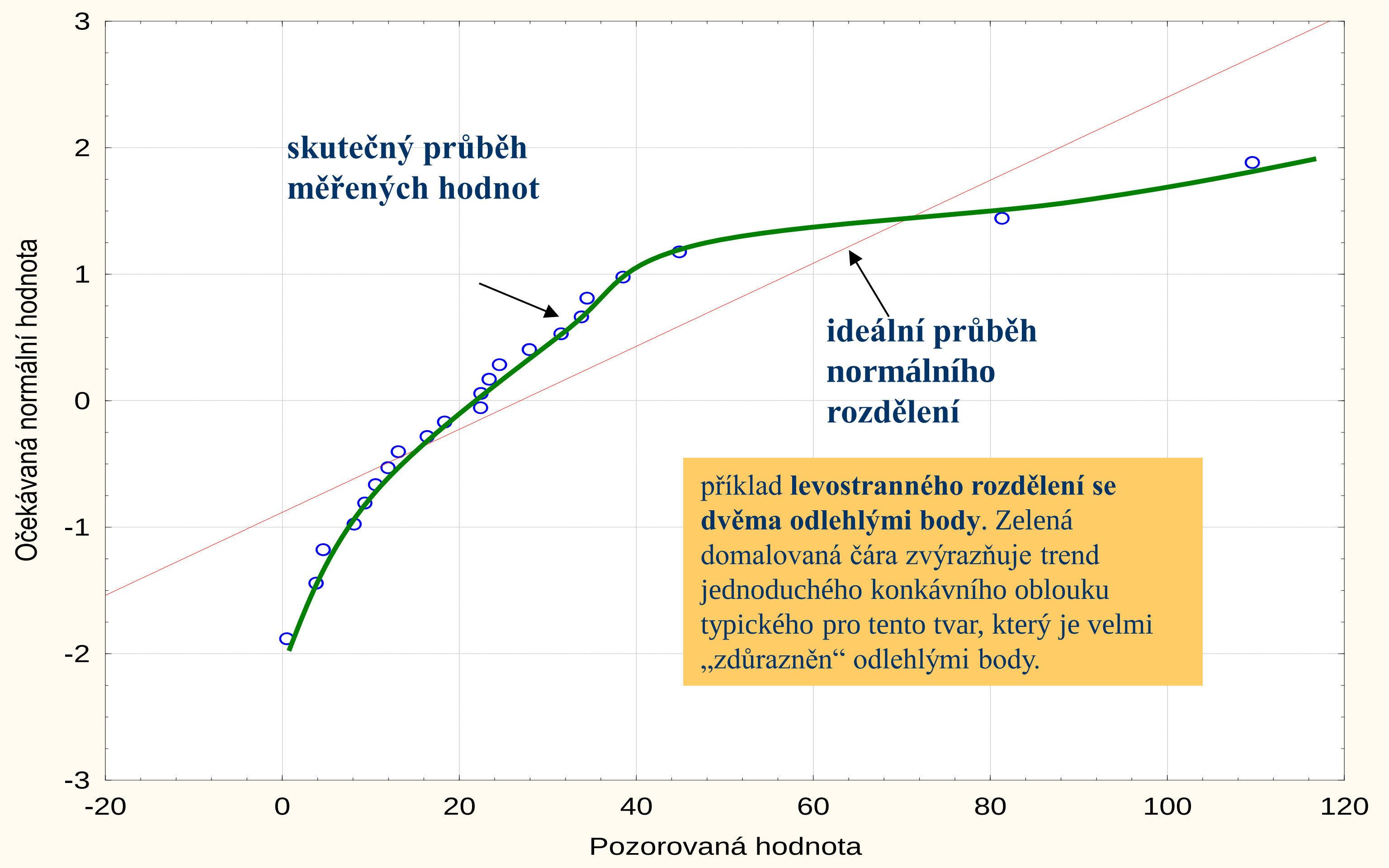
KVANTIL-KVANTILOVÝ GRAF (Q-Q GRAF)

TATO INTERPRETACE PLATÍ POUZE PŘI USPOŘÁDÁNÍ OS, KTERÉ JE UVEDENO NA OBRÁZKU (tj. modelové (teoretické) hodnoty na ose Y, měřené hodnoty na ose X). Pokud by byly osy „přehozeny“, jak tomu bývá v některých statistických programech, byla by interpretace opačná (tento případ je uveden v teorii text, II, str. 8).

KVANTIL-KVANTILOVÝ GRAF (Q-Q GRAF)

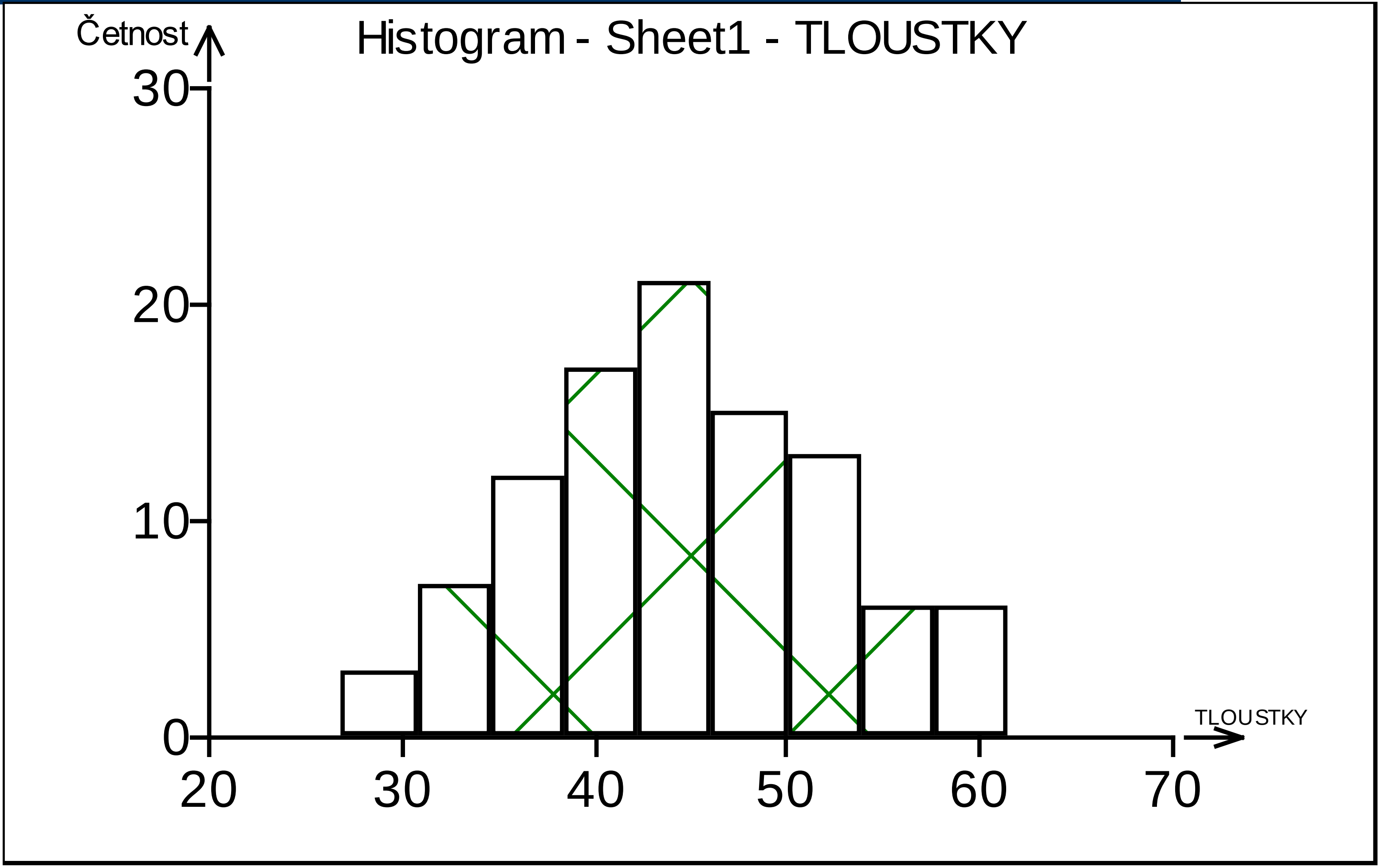


KVANTIL-KVANTILOVÝ GRAF (Q-Q GRAF)



příklad **levostranného rozdělení se dvěma odlehlými body**. Zelená domalovaná čára zvýrazňuje trend jednoduchého konkávního oblouku typického pro tento tvar, který je velmi „zdůrazněn“ odlehlými body.

HISTOGRAM – graf četností

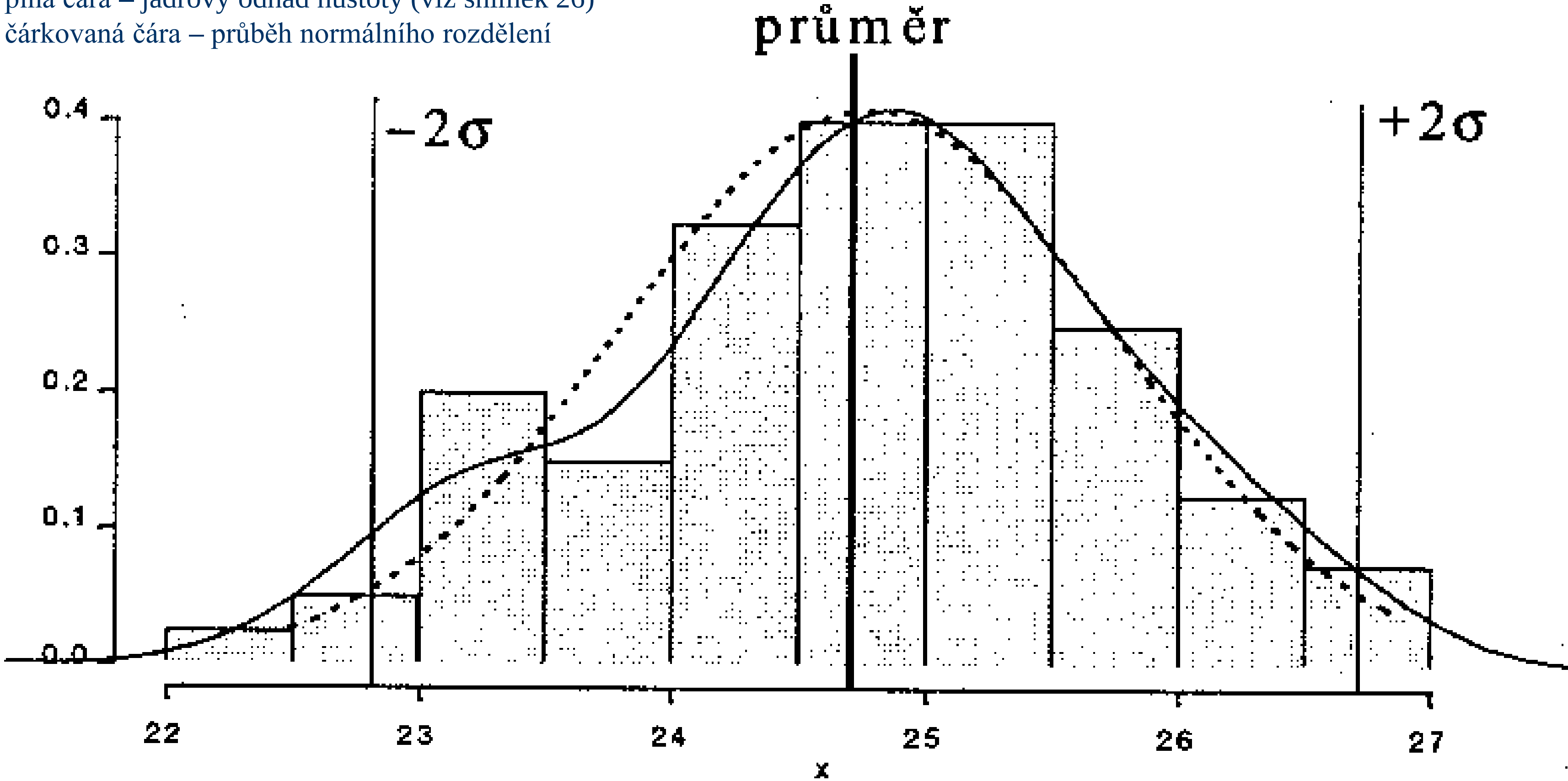


HISTOGRAM – graf četností

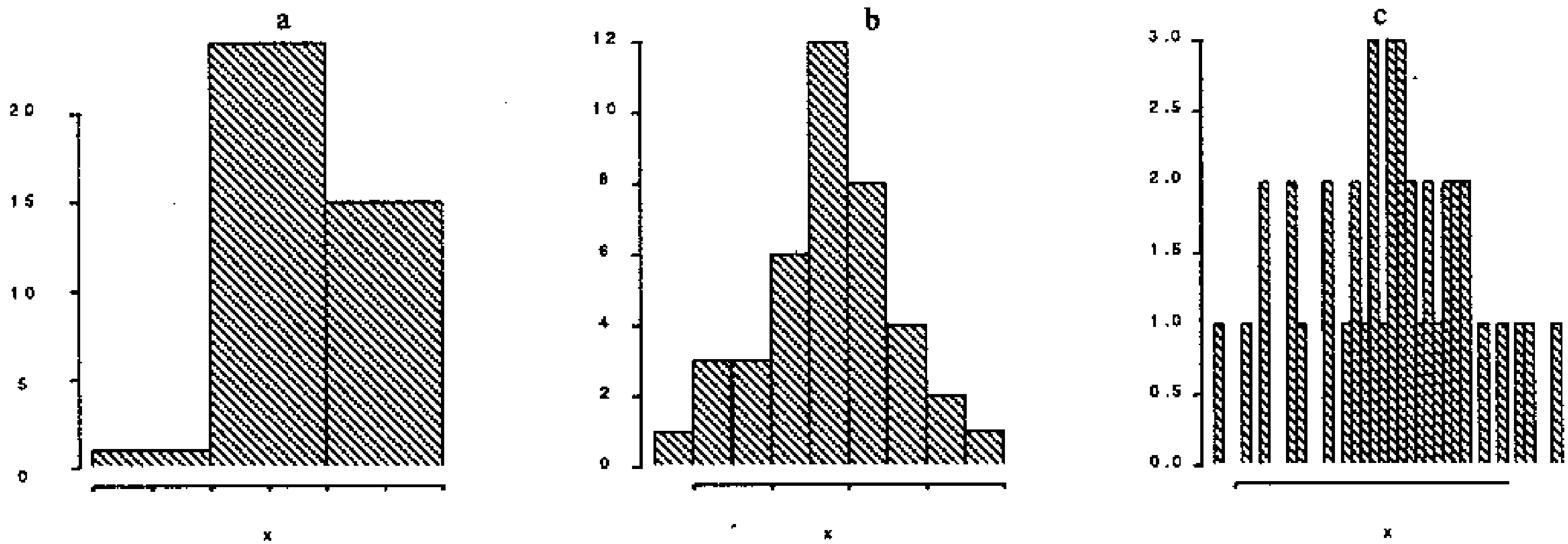
Histogram je graf četností. Na ose X jsou vyneseny intervaly měřených hodnot, na ose Y četnosti hodnot spadajících do těchto intervalů (nebo pravděpodobnosti výskytu těchto hodnot).

HISTOGRAM – graf četností

příklad histogramu – sloupcový graf
plná čára – jádrový odhad hustoty (viz snímek 26)
čárkovaná čára – průběh normálního rozdělení



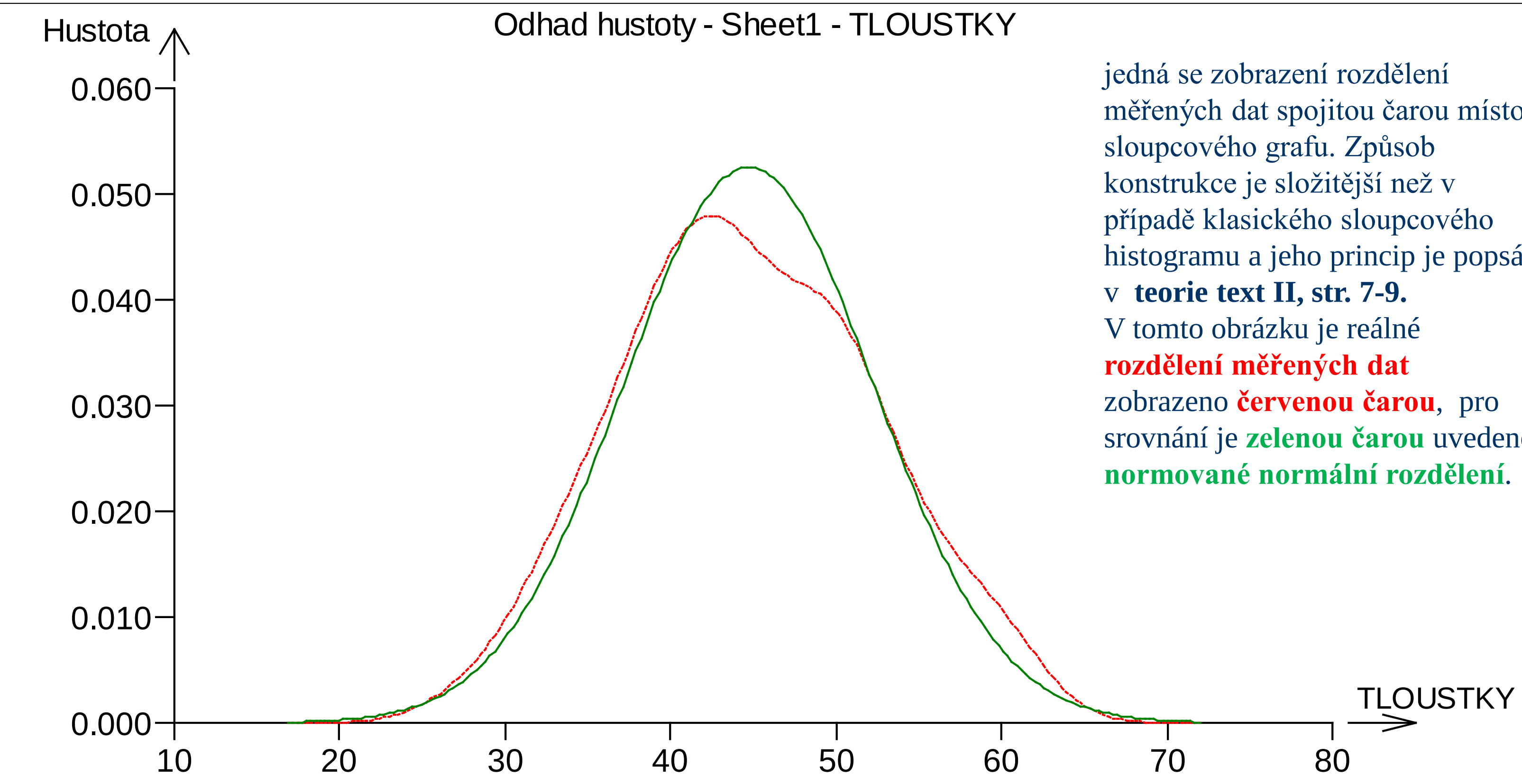
HISTOGRAM – graf četností



Důležitá je správná volba šířky „třídy“ – tj. intervalu na ose X.
Obrázek uvádí tři příklady třídění stejných dat. Nalevo je příliš malý počet tříd, napravo příliš velký, nejvhodnější je třídění na obrázku uprostřed. Možný postup zjištění doporučovaného počtu tříd a šířky intervalu uvádí **teorie text I, str. 18-20**, další možné vzorce pro zjištění doporučovaného počtu tříd (L) jsou uvedeny zde („int“ znamená celočíselnou část čísla v závorce, „n“ je počet měřených hodnot).

$$L = \mathrm{int}\left[2{,}46 \cdot (n-1)^{0{,}4}\right] \qquad \boldsymbol{L} = \mathrm{int}\left(2 \cdot \sqrt{\boldsymbol{n}}\right)$$

HISTOGRAM – jádrový odhad hustoty



jedná se zobrazení rozdělení měřených dat spojitou čarou místo sloupcového grafu. Způsob konstrukce je složitější než v případě klasického sloupcového histogramu a jeho princip je popsán v **teorie text II, str. 7-9.**
V tomto obrázku je reálné **rozdělení měřených dat** zobrazeno **červenou čarou**, pro srovnání je **zelenou čarou** uvedeno **normované normální rozdělení**.

TESTY NORMALITY

testují H_0:

Výběr pochází ze základního souboru s normálním rozdělením

Používané testy:

- Kolmogorov – Smirnovův (KS test)
- d'Agostinův
- Shapiro – Wilkův
- Lillieforsův
- a mnoho dalších ….

Teorie některých z těchto testů (d'Agostinův a Shapiro – Wilkův) je uvedena v **teorie text II, str. 19-22,** Kolmogorov-Smironovova testu v **teorie text I, str. 130-132**
Výpočet je obvykle poměrně složitý (snad v vyjímkou K-S testu), relativně jednoduchý a vhodný např. pro výpočet v Excelu je Lilieforsův test (viz následující snímek).

TESTY NORMALITY – Lilieforsův test

Test šikmosti (A-koef. šikmosti, n – velikost výběru)

$$A_1 = \frac{|A|}{\sqrt{\dfrac{6\cdot(n-2)}{(n+1)\cdot(n+3)}}}$$

Test špičatosti (E-koef. špičatosti)

$$E_1 = \frac{|E| - \dfrac{6}{n+1}}{\sqrt{\dfrac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}}$$

Nulovou hypotézu přijímáme, jestliže platí: A_1 a současně $E_1 \leq z_{\alpha/2}$, kde $z_{\alpha/2}$ je kvantil normovaného normálního rozdělení N(0,1).

Pokud alespoň jedno testové kritérium (buď A_1 nebo E_1) nevyhoví této nerovnosti, nulová hypotéza se zamítá.

Výhodou tohoto testu je jednoduchý výpočet a také skutečnost, že zvlášť testuje šikmost a špičatost. Tím je možné zjistit, zda se rozdělení měřených hodnot odlišuje od normálního rozdělení jen v šikmosti nebo jen ve špičatosti nebo v obojím.

ZÁVISLOST A AUTOKORELACE

Obecná definice závislosti:

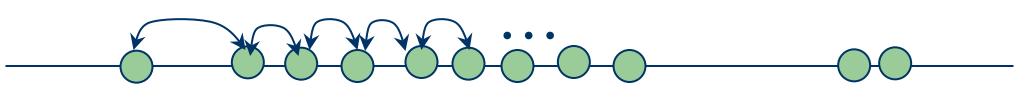
$$\mathbf{x_i = kF(x_1, x_2, \dots, x_{i-1}) + e_i}$$

pokud platí k = 0, jedná se o data nezávislá

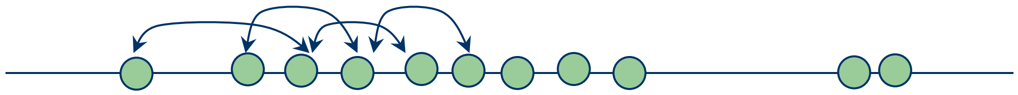
Vzájemná závislost prvků jednoho souboru - **AUTOKORELACE**

$$\mathbf{x_i = \rho_k\, x_{i-k} + e_i}$$

ρ_k autokorelační koeficient k-tého řádu

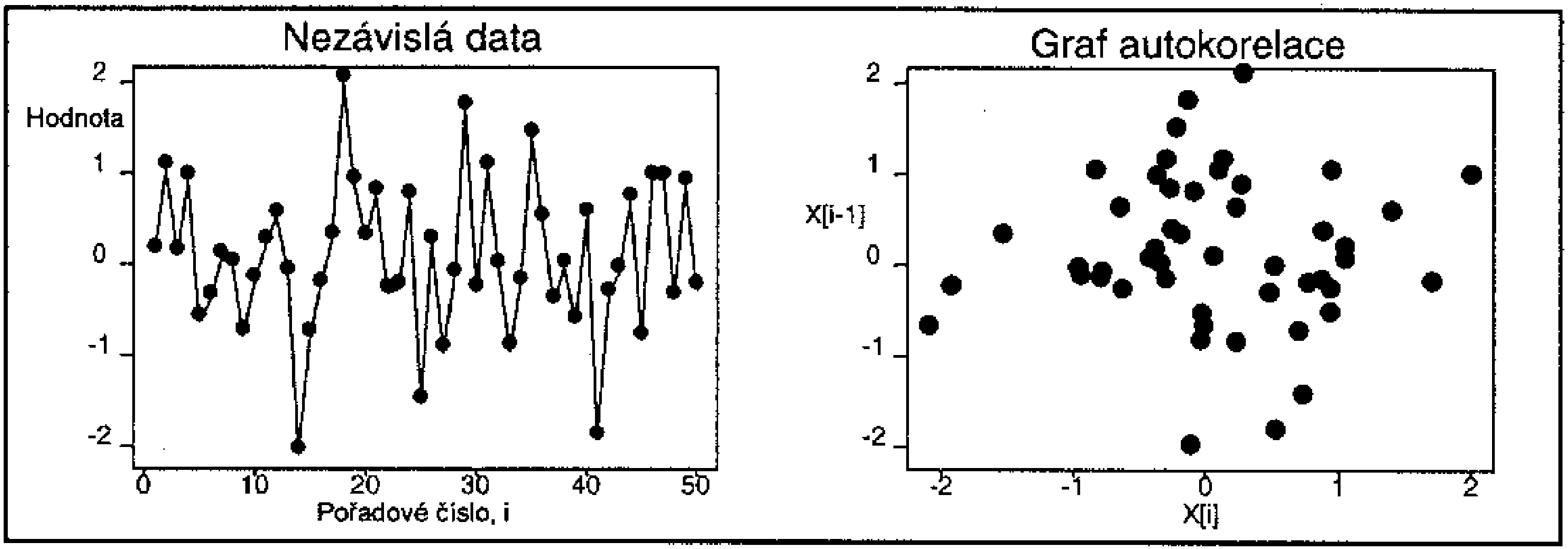


autokorelace I. řádu
sousední hodnoty



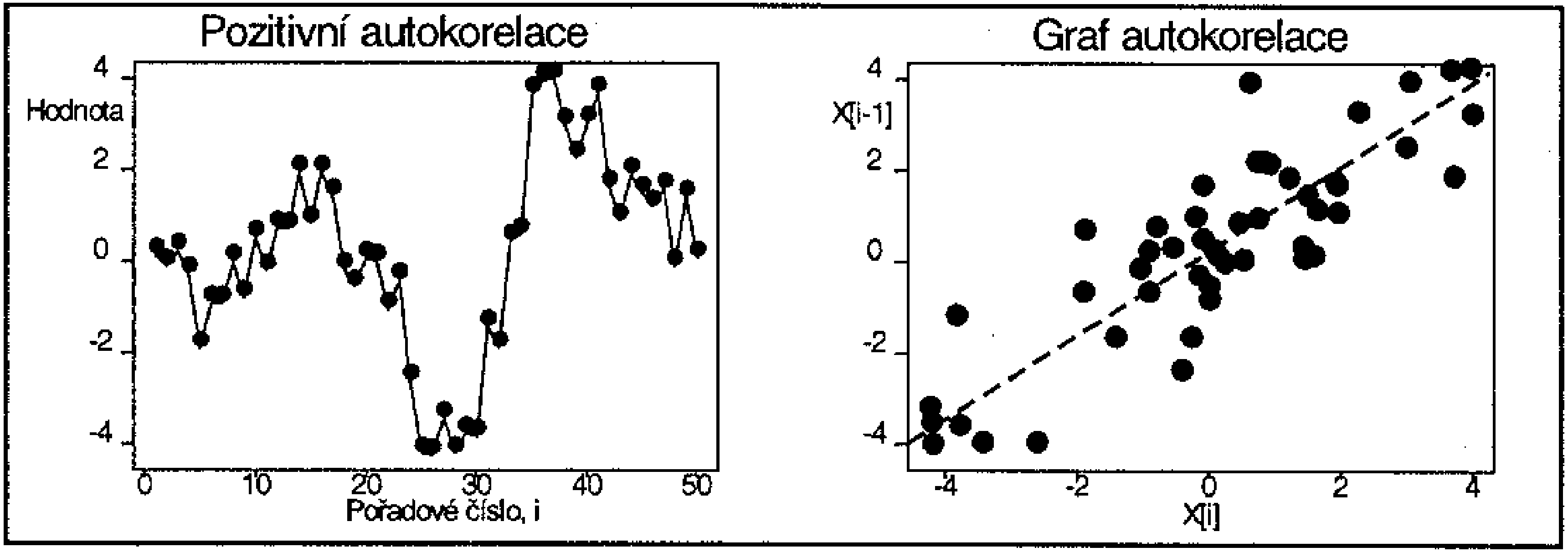
autokorelace II. řádu
hodnoty „přes jednu“

PŘÍKLADY AUTOKORELACE



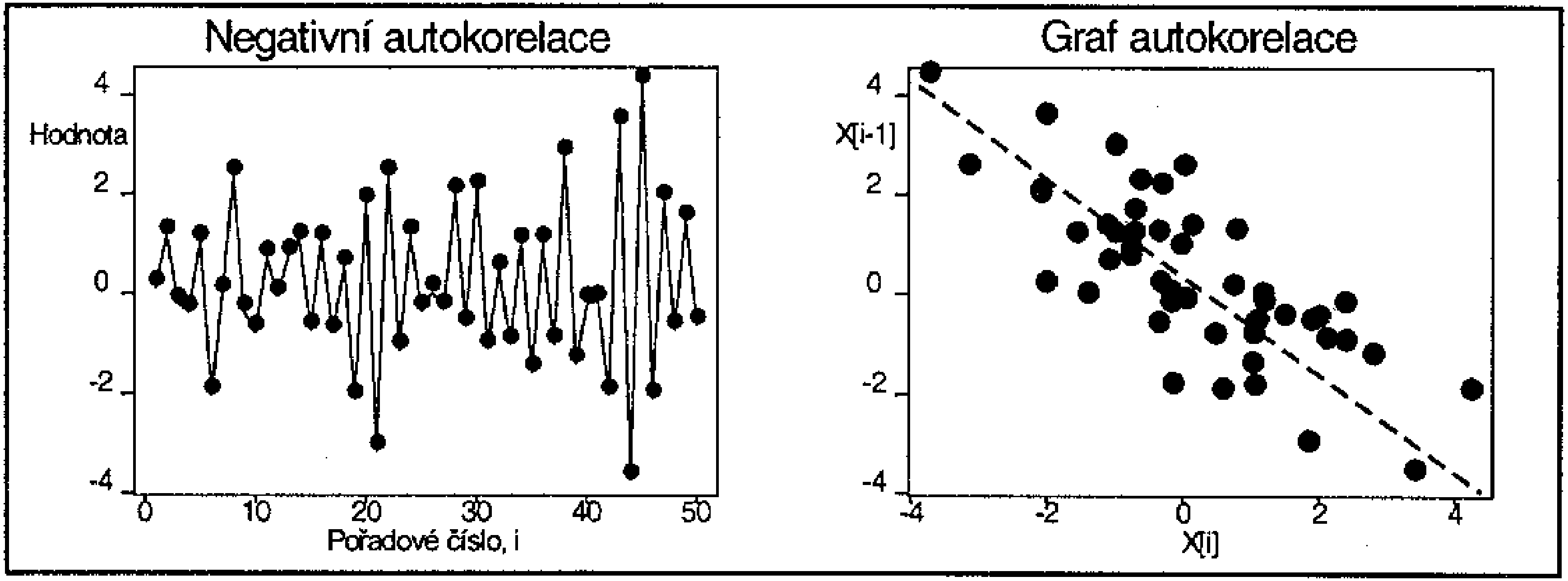
vlevo jsou naměřená data v pořadí tak jak byla změřena, vpravo je graf autokorelace I. řádu (závislost dat jdoucích bezprostředně za sebou (x_i a x_{i-1}). Pokud tento graf vytváří „mrak" bodů bez trendu jako je tomu na tomto obrázku, jsou data navzájem **nezávislá** (což je dobře, je tak dodržena základní podmínka náhodného výběru) a podmínka pro výpočet momentových charakteristik

PŘÍKLADY AUTOKORELACE



Pozitivní autokorelace je vytvářena dlouhými sekvencemi dat stoupajících nebo klesajících, tj. jdoucích ve stejném trendu. Na pravém grafu je vidět výrazný pozitivní trend – **data jsou závislá**, není dodržena podmínka náhodného výběru o nezávislosti dat. Dalším cílem analýzy by mělo být najít příčinu tohoto trendu a odstranit ji.

PŘÍKLADY AUTOKORELACE



Negativní autokorelace vzniká hlavně pravidelným střídáním trendu dat (sekvence vyšší-nižší-vyšší-nižší-…) – **data jsou také vzájemně závislá**)

TESTY NEZÁVISLOSTI

testují H_0:

Všechny prvky výběru jsou NAVZÁJEM nezávislé, ve výběru není autokorelace (autokorelační koeficient se rovná nule).

Používají se:

- testy autokorelace určitého řádu, např. pro autokorelaci I. řádu von Neumannův test (viz **teorie text I, str. 116**)
- je možné použít obecné testy významnosti korelačních koeficientů (viz testování významnosti korelačních koeficientů **teorie text II, str. 112**, kde používáme test shody kor. koef. se zadanou hodnotou). V našem případě zvolíme hodnotu 0 (pokud je autokorelační koef. základního souboru nulový, potom jsou data nezávislá). Tedy pokud nulovou hypotézu tvrdící, že autokorelační koeficient je v základním souboru nulový, nezamítneme (přijmeme jako platnou), potom data považujeme za vzájemně nezávislá. V opačném případě je v datech prokázán trend a data se považují za závislá.

TESTY ODLEHLÝCH HODNOT

H$_0$: *Odchylka extrémní hodnoty je náhodná*

GRUBBSŮV TEST (předpokládá normální rozdělení)

$$T_n = \frac{x_n - \bar{x}}{S} \qquad T_1 = \frac{\bar{x} - x_1}{S}$$

Hypotéza je přijata, když $T_1 < T_{1,\alpha}$, resp. $T_n < T_{n.\alpha}$.

x_n – nejvyšší hodnota souboru,
x_1 – nejmenší hodnota souboru
$\bar{x}$ – průměr dat
S – směrodatná odchylka
x_{n-1} – předposlední hodnota
x_2 - druhá nejmenší hodnota
$T_{n,\alpha}$ – hodnota Studentova T-rozdělení pro hladinu významnosti α a n stupňů volnosti

DIXONŮV TEST (nepředpokládá normální rozdělení)

$$Q_n = \frac{x_n - x_{n-1}}{x_n - x_1} \qquad Q_1 = \frac{x_2 - x_1}{x_n - x_1}$$

Hypotéza je přijata, když $Q_1 < Q_{1,\alpha}$ resp. $Q_n < Q_{n,\alpha}$.

TESTY ODLEHLÝCH HODNOT

metoda **modifikovaných vnitřních hradeb**

$$B_D^* = F_D - K\,R_F \qquad B_H = F_H + K\,R_F$$

$$K \approx 2.25 - \frac{3.6}{n}$$

TRANSFORMACE

Podrobnější teorie transformace a jejího použití k odhadu střední hodnoty v **teorie text II, str. 29 - 33**

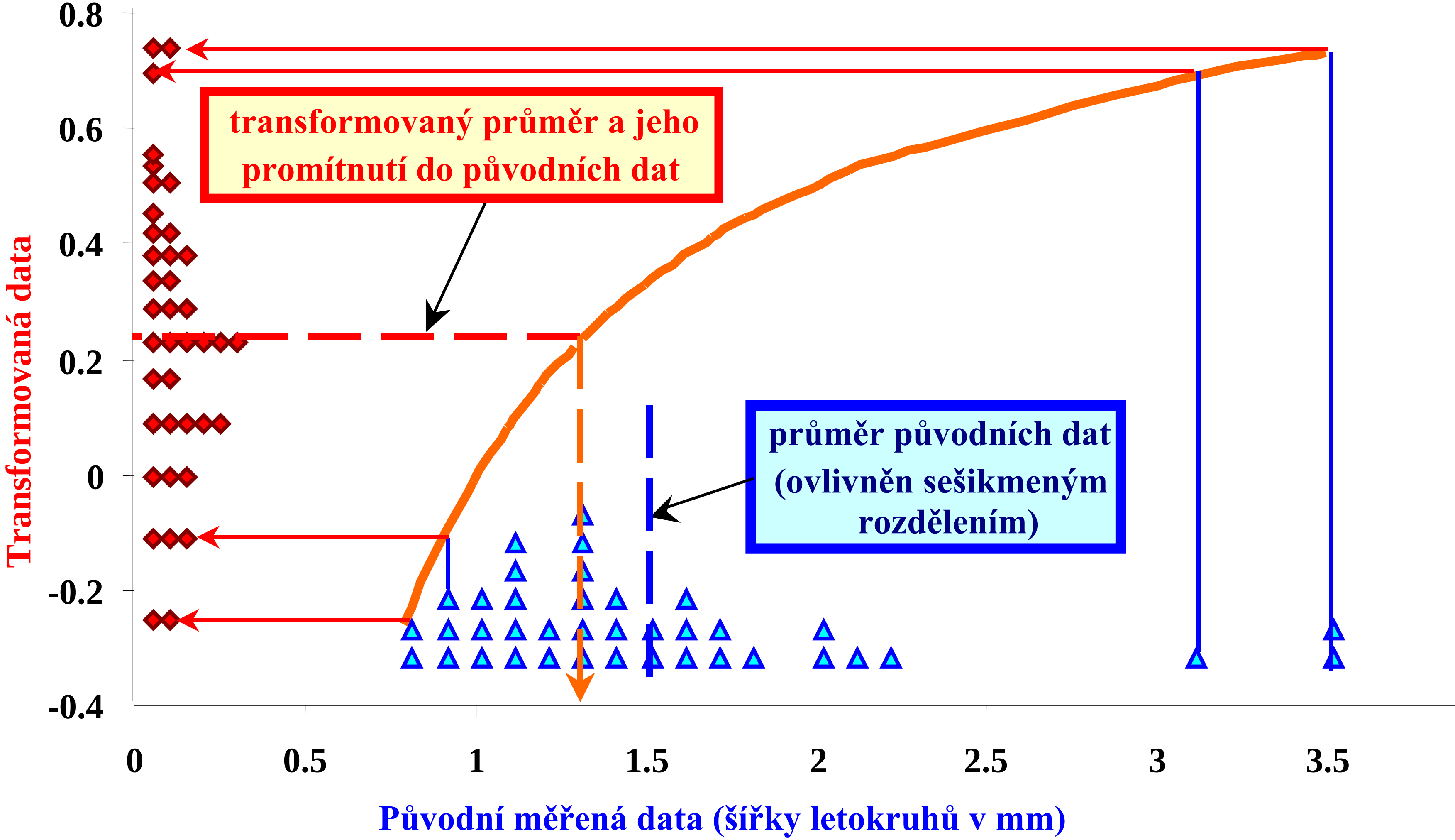
Transformace slouží nejčastěji k:

- **snížení rozptylu** dat
- dosažení **vyššího stupně symetrie** (nejlépe **normality**) dat

Požadavky na transformační funkci:

- **nelineární funkce** (jinak by došlo pouze k posunu dat a změně měřítka)
- **monotónnost** průběhu (aby se nezměnilo pořadí velikosti dat)
- musí směřovat k **maximální symetrii**

TRANSFORMACE -princip



TRANSFORMACE -princip

- Máme výběr, který se vyznačuje silnou asymetrií (data vyznačena modrými trojúhelníčky). Data se vyznačují jednak silnou koncentrací mezi hodnotami 1 a 1,5 mm, jednak odlehlými hodnotami (3-3,5 mm). Proto aritmetický průměr (vyznačen modrou čárkovanou čárou) není vhodný – je ovlivněn nesouměrným souborem a odlehlými hodnotami.
- Ověřili jsme si, že odlehlé hodnoty byly správně stanoveny, nejedná se o hrubé chyby, není tedy možné je vyloučit.
- Potřebujeme nalézt takový odhad střední hodnoty, který bude zahrnovat vliv všech dat, ale nebude negativně ovlivněný výše uvedenými skutečnostmi.
- Řešením je nalezení vhodného tvaru transformační funkce (na obrázku vyznačena oranžovou čarou);
- Pomocí této funkce transformujeme původní data tak, aby „nová" data (na obrázku jsou jejich hodnoty vyznačeny červenými kosočtverci) byla pokud možno co nejsymetričtější (je vidět, že transformace odstranila hlavní odlehlé hodnoty a že „nová data" vykazují podstatně vyšší míru symetrie než původní – transformace pro nejvychýlenější původní hodnoty - 3,5 - je vyznačena pomocí krátce čárkované čáry).

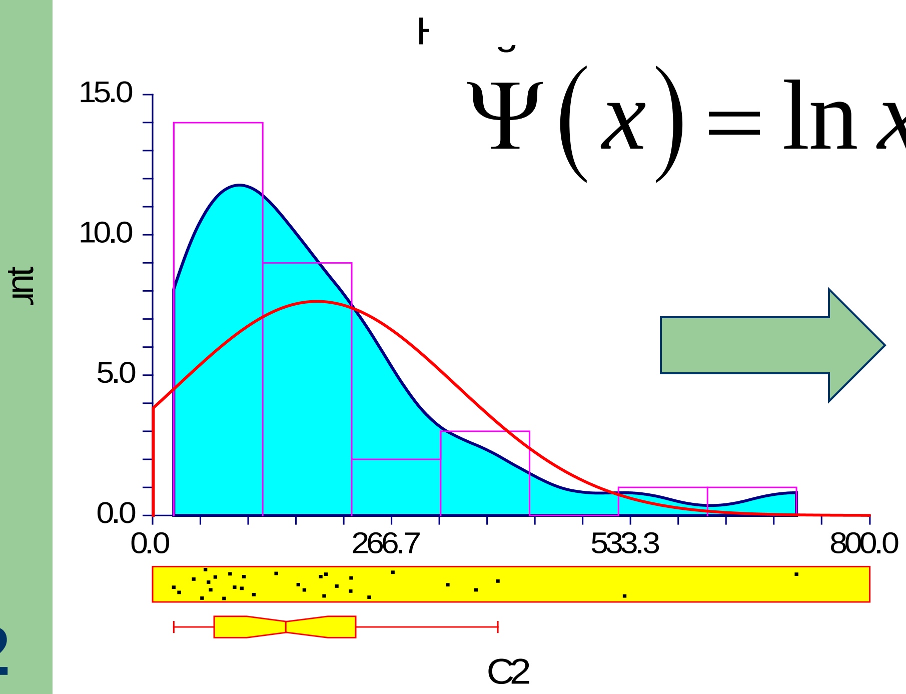
TRANSFORMACE -princip

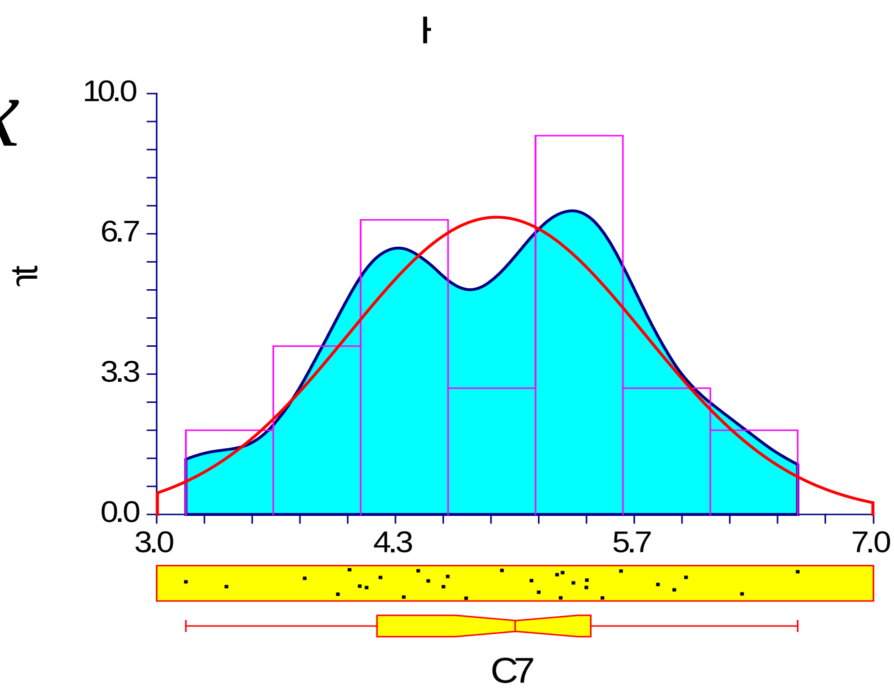
- Je zřejmé, že transformační funkce vhodného tvaru velmi koncentrované hodnoty od sebe „oddaluje“ a velmi vzdálené hodnoty „přibližuje“, nicméně základní vztahy mezi daty zůstávají zachovány (např. pořadí hodnot);
- V souboru „nových dat“ již můžeme vypočítat aritmetický průměr běžným způsobem (tato data jsou minimálně symetrická, pokud máme štěstí, tak i normální), stejně jako interval spolehlivosti, apod.;
- Problémem je, že „nová“ (transformovaná) data mají úplně jiné měřítko než data původní (zde např. původní data byla přibližně v intervalu 0,8 -3,5 mm, transformovaná data jají interval přibližně -0,25 – 0,75. Tedy, i když vypočítáme jejich průměr, neřekne nám to nic o průměru původních dat. Proto musíme výsledné hodnoty (např. průměr a jeho intervalový odhad) retransformovat do původních dat.
- Odhady parametrů vypočítané pro transformované hodnoty promítneme (retransformujeme) do původních souřadnic pomocí funkce inverzní k původní transformační funkci. Tím získáme kvalitnější odhady parametrů a intervaly spolehlivosti než z původních dat.

TRANSFORMACE – logaritmická transformace

Logaritmická transformace (používá se s zpravidla pro veličiny s výrazně levostranným rozdělením) a spočívá v zlogaritmování dat.

$$\breve{\Psi}(x) = \ln x$$





TRANSFORMACE – Box-Coxova

$$\Psi(x) = \begin{cases} \dfrac{x^{\lambda} - 1}{\lambda} & \lambda \neq 0 \\ \ln x & \lambda = 0 \end{cases}$$

V současné době nejčastěji používaná transformace, její „úspěch“ závisí na správném stanovení hodnoty λ. „x“ je původní (měřená) hodnota. Pokud se hodnota λ=0, jedná se o logaritmickou transformaci.

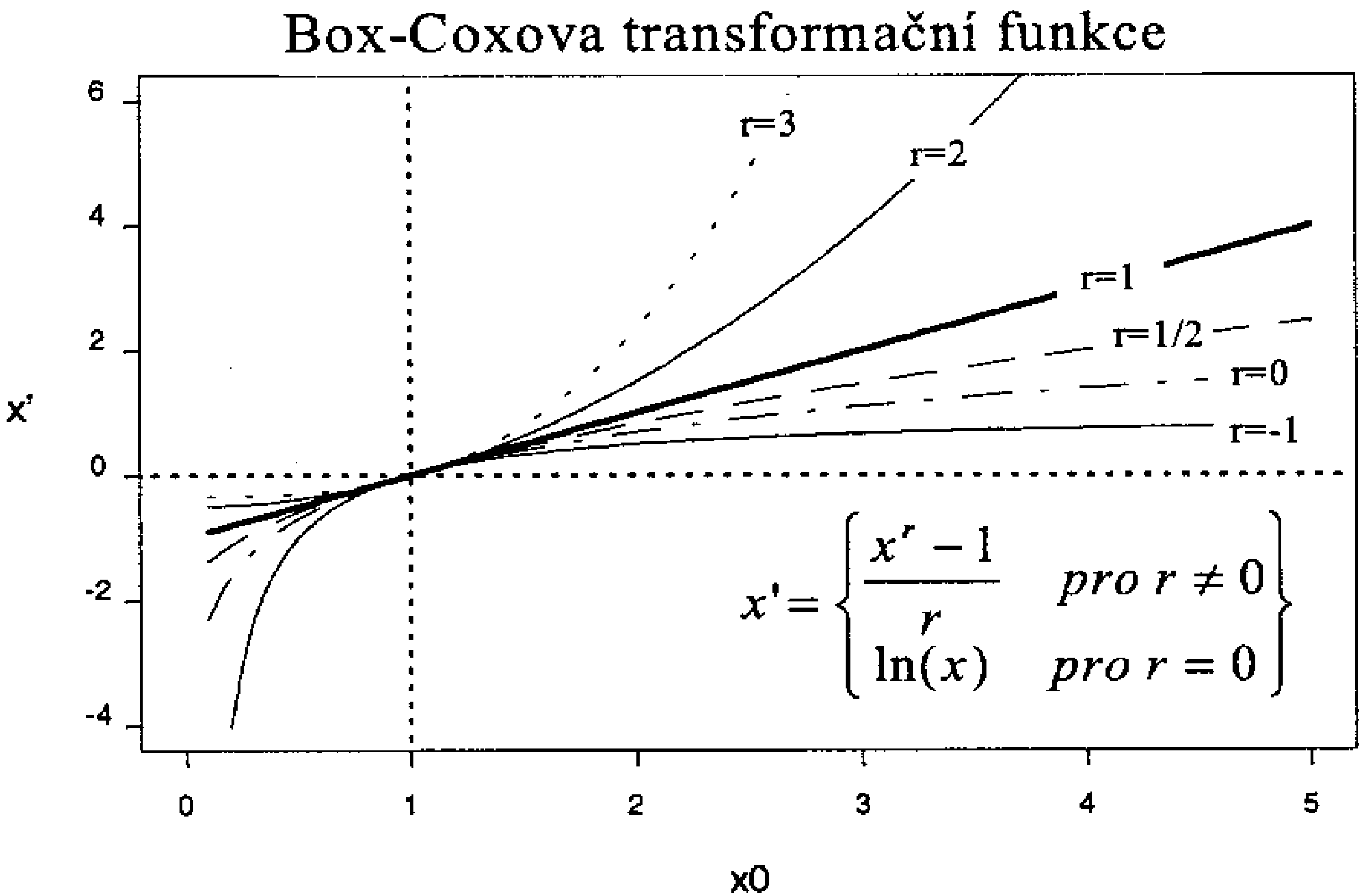
TRANSFORMACE – Box-Coxova

Následující obrázek ukazuje průběh Box Coxovy funkce pro různé hodnoty λ (na obrázku označené jako „r“). Pokud platí, že λ= +1, potom je funkcí jen přímka a k žádné transformaci nedochází. Proto je nutné zjistit, zda vypočítaná hodnota λ není rovna nebo blízká hodnotě +1 (a zda interval spolehlivosti λ neobsahuje tuto hodnotu). Pokud tomu tak je, transformaci nemá cenu provádět.

Pozor! Pro hodnotu λ = -1 toto neplatí. Pokud je λ = -1, potom je transformace účinná.

Nejběžnější hodnoty λ se pohybují v rozmezí (-3;+3), hodnoty záporné a menší než 1 jsou vhodná pro levostranná rozdělení hodnoty vyšší než 1 pro pravostranná.

TRANSFORMACE – Box-Coxova



TRANSFORMACE – Box-Coxova

Stanovení optimální hodnoty λ

„Úspěch“ transformace je závislý na tvaru transformační funkce a tedy na hodnotě λ. Následující obrázek ukazuje princip jejího stanovení.
Optimální hodnota se stanoví jako hodnota na ose X, která odpovídá extrému (v tomto případě maximu) logaritmu věrohodnostní funkce (rovnice viz **teorie text II, str. 31**).
Optimální hodnota je vyznačena zelenou šipkou. Poté se stanoví její interval spolehlivosti (modré čárkované čáry). Pokud tento interval neobsahuje hodnotu +1, povede pravděpodobně transformace ke kvalitnímu odhadu střední hodnoty.

TRANSFORMACE – odhad optimálního λ