Univerzita Palackého v Olomouci

Průzkumová analýza dat

Chemometrie I (ACH/CHEX1)

1

Univerzita Palackého v Olomouci

Průzkumová analýza dat

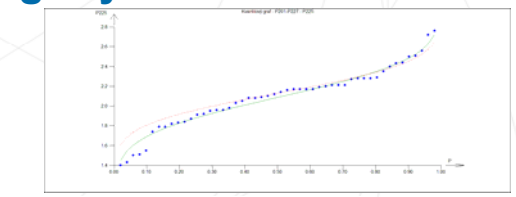
- V aj. *Exploratory Data Analysis*, často používaná zkratka EDA.
- Moderní statistická metoda využívající možnosti statistických SW k podrobnému zkoumání dat, často kombinující grafy a statistické testy k získání komplexních informací o datech.
- Cíl EDA: odhalení případných statistických zvláštností v datech, tj. odchylek od předpokládaného rozdělení dat. V přírodních vědách je nejčastějším předpokladem, že data pocházejí z normálního rozdělení.
- **OBECNÝ POSTUP:**
  1. Diagnostické grafy EDA – identifikace statistických zvláštností.
  2. Ověření základních předpokladů (ZP) o datech s použitím statistických testů.
  3. Nejsou-li splněný ZP → transformace dat → vyčíslení odhadů polohy a rozptýlení.
  4. Jsou-li splněny ZP → vyčíslení odhadů polohy a rozptýlení.
- Pojmy používané v grafech:
  - pořádková statistika: vzestupně setříděný soubor dat,
  - pořadová pravděpodobnost P_i: $P_i = \frac{i}{n+1}$; *i* je pořadí hodnoty v setříděném souboru dat a *n* je počet hodnot.

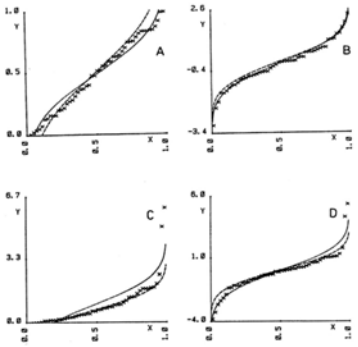
2

Univerzita Palackého v Olomouci

Diagnostické grafy EDA

- KVANTILOVÝ GRAF (osa x: pořadová pravděpodobnost P_i, osa y: pořádková statistika)
  - Zobrazuje empirické kvantily dat proložené kvantilovou funkcí normálního rozdělení. V QC Expertu zelená křivka odpovídá funkci s klasickým průměrem a rozptylem, červená křivka odpovídá mediánu a mediánové odchylce. Podle toho, která z křivek lépe prokládá data, je vhodné zvolit jako odhad střední hodnoty průměr nebo medián.
  - Ukazuje lokální koncentrace dat, odchylky v symetrii (nesigmoidální tvar) a indikuje OB.





A. rovnoměrné r.
B. normální r.
C. exponenciální r.
D. Laplaceovo r.

3

Univerzita Palackého v Olomouci

Diagnostické grafy EDA

- DIAGRAM ROZPTÝLENÍ (osa x: hodnoty x_i, osa y: nemá význam)
  - Zobrazuje data ve skutečném měřítku na ose x.
  - Jednorozměrná projekce kvantilového grafu do osy x.
  - Ukazuje lokální koncentrace dat a indikuje OB.

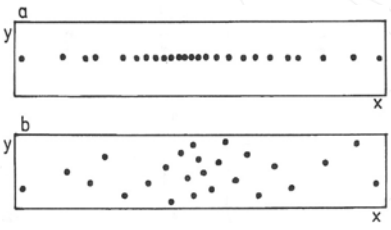
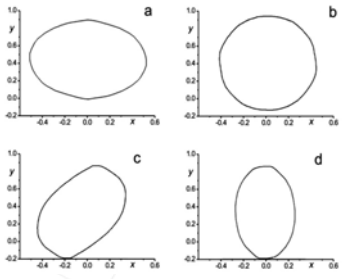


diagram rozptýlení

rozmítnutý diagram rozptýlení

- KRUHOVÝ GRAF
  - Slouží k vizuálnímu posouzení normality na základě kombinace šikmosti a špičatosti.
  - V QC Expertu zelený kruh (elipsa) je optimální tvar pro normální rozdělení, černý představuje data. Pro normální data se křivky „téměř“ kryjí.



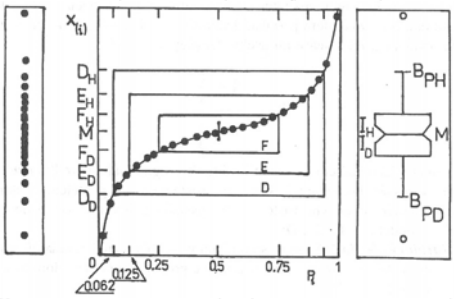
a) rovnoměrné r.
b) normální r.
c) exponenciální r.
d) Laplaceovo r.

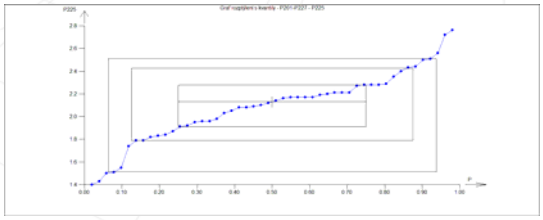
4

Univerzita Palackého v Olomouci

Diagnostické grafy EDA

- GRAF ROZPTÝLENÍ S KVANTILY (osa x: pořadová pravděpodobnost P_i, osa y: pořádková statistika)
  - 3 obdélníky: F – kvartilový, E – oktilový, D – sedecilový.
  - Ukazuje lokální koncentrace dat, odchylky v symetrii (nesigmoidální tvar, poloha obdélníků) a indikuje OB (mimo sedecilový obdélník D).
  - Body grafu jsou vizuálně i významově shodné s kvantilovým grafem. Vzájemná poloha obdélníků odpovídá symetrii, resp. asymetrii rozdělení. Vodorovná příčka uprostřed nejmenšího obdélníku označuje medián, svislá úsečka na příčce odpovídá intervalu spolehlivosti mediánu.



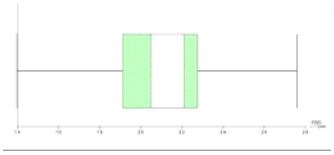


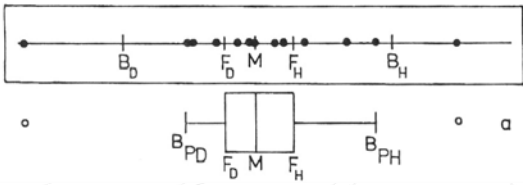
5

Univerzita Palackého v Olomouci

Diagnostické grafy EDA

- KRABICOVÝ GRAF (osa x: úměrná hodnotám x_i, osa y: interval)
  - Větší obdélník ohraničuje vnitřních 50% dat, horní okraj zeleného obdélníku odpovídá 75% kvantilu, spodní okraj zeleného obdélníku odpovídá 25% kvantilu, střed bílého pruhu v zeleném obdélníku odpovídá mediánu, šířka pruhu odpovídá IS mediánu, dva černé proužky jsou vnitřní hradby. Body mimo vnitřní hradby jsou OB.
  - **Vnitřní hradby:** $B_H = F_H + 1{,}5R_F$ $B_D = F_D - 1{,}5R_F$
    kde $R_F = F_H - F_D$





- **ODCHYLKY V ŠIKMOSTI** (+ je aritmetický průměr)

- **ODCHYLKY VE ŠPIČATOSTI** (+ je aritmetický průměr)

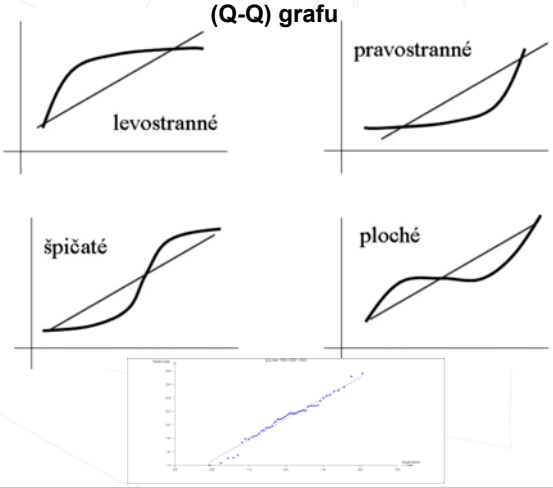
6

Univerzita Palackého v Olomouci

Diagnostické grafy EDA

- KVANTIL-KVANTILOVÝ (Q-Q) GRAF (osa x: kvantilová funkce normálního (obecně teoretického) rozdělení, osa y: pořádková statistika)
  - Graf pro diagnostiku normality a OB.
  - Pro normální data bez OB má tvar přímky; pro normální data s OB má tvar přímky s koncovými body ležícími mimo tuto přímku; pro systematicky zešikmená má nelineární konvexní nebo konkávní tvar. Pro data s vyšší špičatostí než odpovídá normálnímu rozdělení, má tvar konkávně-konvexní . Pro data s nižší špičatostí než odpovídá normálnímu rozdělení, má tvar konvexně-konkávní.
  - Výhoda Q-Q grafu: možnost vizuálně posoudit, zda je nelinearita způsobena jen několika body, nebo všemi daty.
  - Většinou platí: protíná-li spojnice mezi body přímku > 4×, přikloníme se k normalitě, protíná-li spojnice mezi body přímku < 4×, přikloníme se k porušení normality.

Odchylky od normality v kvantil-kvantilovém (Q-Q) grafu

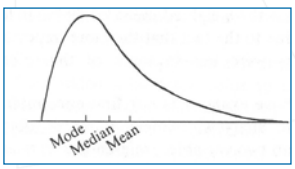


7

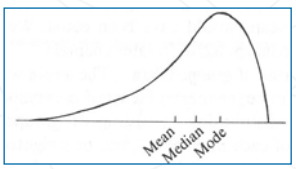
Univerzita Palackého v Olomouci

Diagnostické grafy EDA

- GRAF HUSTOTY PRAVDĚPODOB-NOSTI (osa x: hodnoty x_i, osa y: odhad hustoty pravděpodobnosti (f(x))
  - Graf identifikuje odchylky v g_1, g_2, lokální koncentrace dat a OB.
  - V případě normality jsou si zelená křivka normálního rozdělení a červená křivka experimentálních dat blízké.



$g_1 > 0$



$g_1 < 0$

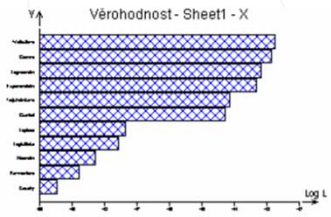
- PRAVDĚPODOBNOSTNÍ (P-P) GRAF (osa x: P_i, osa y: standardizovaná distribuční funkce)
  - Porovnává data s normálním (modrá křivka), Laplaceovým (zelená křivka) a rovnoměrným (červená křivka). Která křivka leží nejblíže černé přímce, to rozdělení odpovídá datům.
  - P-P graf je citlivý na odchylky od teoretického rozdělení ve střední části, Q-Q graf v oblasti konců.
- **LAPLACEOVO PRAVIDLO:** statistickou zvláštnost v datech považujeme za prokázanou, pokud byla odhalena alespoň ve 3 grafech EDA.

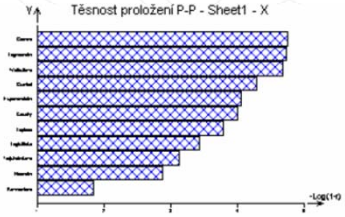
8

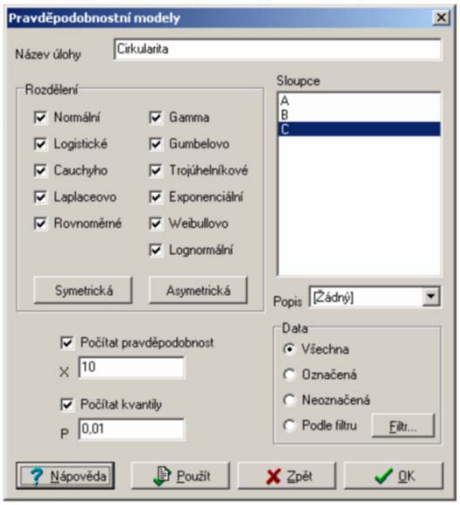
Univerzita Palackého v Olomouci

Porovnání rozdělení dat v QC Expertu

- Modul pravděpodobnostní modely umožňuje porovnat experimentální data s 11 rozděleními.
  - Graf věrohodností L: nejlepší rozdělení z hlediska věrohodnosti je nejvýše a má nejvyšší hodnotu logL.
  - Graf těsnosti proložení: porovnává těsnost proložení v P-P grafu pomocí hodnot odvozených z korelačního koeficientu –log(1–R). Rozdělení s nejvyšší hodnotou nejlépe vystihují distribuční funkci dat.







9

Univerzita Palackého v Olomouci

Základní předpoklady o datech

10

Univerzita Palackého v Olomouci

Ověření základních předpokladů

- Celá řada statistických metod předpokládá, že analyzovaná data splňují určité předpoklady. Nejsou-li tyto předpoklady splněny, vede použití „klasických" statistických metod ke zkresleným výsledkům.
- Proto je třeba předpoklady ověřit a nejsou-li splněny zvolit vhodnou statistickou metodu k analýze těchto souborů dat, např. transformaci dat.
- **ZÁKLADNÍ PŘEDPOKLADY:**
  - **nezávislost náhodného výběru dat,**
  - **rozdělení náhodného výběru je normální,**
  - **homogenita výběru (nepřítomnost OB).**
- Mezi základní předpoklady se také řadí:
  - posouzení hodnot koeficientu šikmosti (g_1) a špičatosti (g_2) → odchylky od normality,
  - posouzení velikosti náhodného výběru: rozsah výběru *n* ovlivňuje preciznost odhadů parametrů polohy a rozptýlení a projeví se při konstrukci IS. U velmi malých výběrů může být výsledek (např. šířka IS) více ovlivněn velikostí výběru než variabilitou v datech.

11

Univerzita Palackého v Olomouci

Nezávislost výběru

- Pokud se podmínky pro měření dat mění s časem, projeví se to vznikem trendu mezi prvky výběru a v datech je závislost. Závislá data indikují přítomnost systematické chyby. Závislost může být způsobena např. časovými změnami v měřícím procesu, nekonstantností podmínek, zanedbáním některých faktorů, nenáhodným výběrem vzorků.
- Závislost musí být proměřována před uspořádáním naměřených dat.
- Obvykle se ověřuje testováním významnosti autokorelačního koeficientu ρ_A např. von Neumannovým testem.

$$t_n = \frac{T_1 \cdot \sqrt{n+1}}{1 - T_1}, kde\ T_1 = \left(1 - \frac{T}{2}\right) \cdot \sqrt{\frac{n^2 - 1}{n^2 - 4}}\ a\ T = \frac{\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Pokud jsou prvky výběru vzájemně nezávisle a platí **H_0: $\rho_A = 0$** má veličina t_n Studentovo rozdělení s n+1 stupni volnosti. Pokud $|t_n| > t_{(1-\frac{\alpha}{2}, n+1)}$, H_0 o nezávislosti se zamítá. Alternativní hypotézou je **H_1: $\rho_A \neq 0$**.

12

Univerzita Palackého v Olomouci

Nezávislost výběru

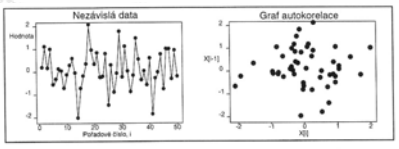
Lineární závislost prvků jednoho souboru - **AUTOKORELACE**

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

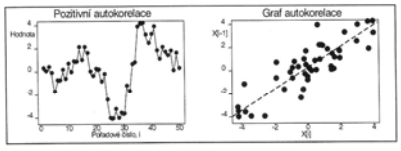
ρ_k autokorelační koeficient k-tého řádu

autokorelace I. řádu sousední hodnoty

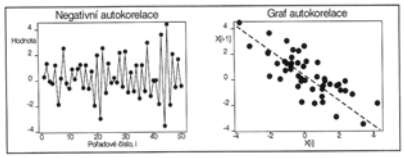
autokorelace II. řádu hodnoty „přes jednu“



Obr. 3-10 A, B Nezávislá data, koeficient autokorelace $\rho_1 = 0$



Obr. 3-11 A, B Autokorelovaná data, $\rho_1 = +0.8$



Obr. 3-12 A, B Autokorelovaná data, $\rho = -0.8$

- Znaménkový test (v QC Expertu): testuje náhodnost ve střídání hodnot vyšších a nižších než průměr. Je-li toto střídání příliš pravidelné a vyskytují-li se nepravděpodobně dlouhé sekvence po sobě jdoucích dat nad nebo pod průměrem, jsou data označena jako závislá.

13

Univerzita Palackého v Olomouci

Normalita výběru

- Normalita výběru patří k základním předpokladům, protože je na ní založena celá „klasická“ statistická analýza dat.
- Statistické testy jsou obecně k odhalení odchylek od normality méně citlivé než diagnostické grafy EDA. Závěr by měl být učiněn na základě kombinace závěrů EDA a statistických testů.
- Testy normality:
  - Momentový test (dle Jarque-Beera) založený na shodě šikmosti (g_1) a špičatosti (g_2) s normálním rozdělením:

$$\chi^2_{exp} = \frac{g_1^2}{D(g_1)} + \frac{[g_2 - E(g_2)]^2}{D(g_2)}$$

Vypočtené χ^2_{exp} srovnáváme s $\chi^2_{krit}(1-\alpha; 2)$. Je-li $\chi^2_{exp} > \chi^2_{krit}$, předpoklad normality se zamítá, je-li $\chi^2_{exp} < \chi^2_{krit}$, normalita se přijímá; $E(g_2)$ – střední hodnota g_2, $D(g_i)$ – rozptyl g_i.

  - Test normality D'Agostina – posuzuje výběrové momenty dat. Obecně je tento test podstatně citlivější na odchylky od normality, než momentový test. V QC Expertu je v modifikacích pro menší výběry (n < 100) a větší výběry (n > 100).
  - Test normality Kolmogorovův-Smyrnovův je založený na rozdílu teoretické a výběrové distribuční funkce korigovaný pro odhady μ a σ.

14

Univerzita Palackého v Olomouci

Homogenita výběru

- Homogenní výběr předpokládá, že všechny jeho prvky pocházejí ze stejného rozdělení s konstantním rozptylem σ^2.
- Nehomogenita bývá způsobena přítomností odlehlých bodů (OB), tj. hodnot, které se co do velikosti výrazně liší od ostatních.
- OB silně zkreslují odhady polohy a rozptylu, což často vede ke znehodnocení celé statistické analýzy.
- Způsoby identifikace OB:
  - Grafy EDA.
  - Metoda modifikovaných vnitřních hradeb B^* (použita v QC Expertu), body ležící mimo modifikované vnitřní hradby jsou OB:

$$B_D^* = F_D - \mathrm{K} \cdot R_F$$
$$B_H^* = F_H + \mathrm{K} \cdot R_F$$

F_D – dolní kvartl, F_H – horní kvartil, R_F – intervartilové rozpětí K ≈ 2,25 – 3,6/n

  - Statistické testy, např. Dean-Dixonův nebo Grubbsův. Tyto testy závisí na velikosti souboru dat.

15

Univerzita Palackého v Olomouci

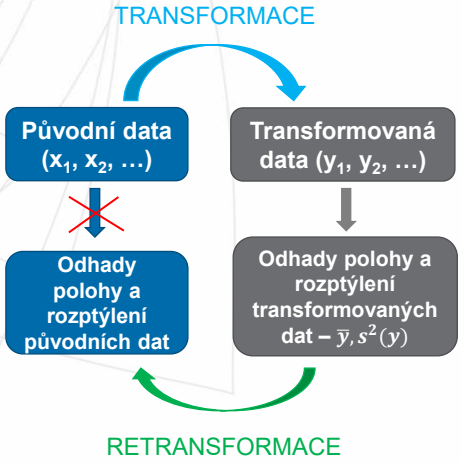
Transformace dat

16

Univerzita Palackého v Olomouci

Transformace dat – úvod

- Pokud analýzou dat zjistíme, že rozdělení se systematicky odlišuje od normálního (asymetrie, odlehlé body, nehomogenita), vzniká problém, jak data vyhodnotit. Často lze použít pro vyhodnocení takových dat jejich transformaci, která vede ke stabilizaci rozptylu, zesymetričtění rozdělení a někdy i k normalitě rozdělení.
- Obvykle nastává jedna z následujících 3 situací:
  1. Zamítnuta normalita v ZP, nejsou OB ⇒ TRANSFORMACE.
  2. Přijata/zamítnuta normalita v ZP, nalezeny OB, které „nelze“ vyloučit ⇒ TRANSFORMACE.
  3. Přijata/zamítnuta normalita v ZP, nalezeny OB, které „lze“ vyloučit ⇒ VYLOUČENÍ OB.
- PRINCIP: je vyhledána vhodná transformace, která zajistí největší přiblížení normalitě, transformace se provede, vypočte se průměr a IS transformovaných dat. Vypočtené údaje se přepočítají (retransformují) do původních dat.



17

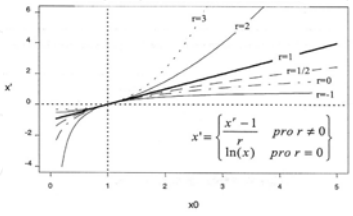
Univerzita Palackého v Olomouci

Základní druhy transformací

- MOCNINNÁ TRANSFORMACE:
  - symetrizující transformace, která je dnes nahrazována Boxovou-Coxovou transformací,
  - optimální odhad λ se hledá minimalizací asymetrie (klasického či robustního koeficientu šikmosti). Pro $\lambda = 0$ jde o logaritmickou transformaci.
- BOXOVA-COXOVA TRANSFORMACE:
  - přibližuje rozdělení výběru k normálnímu z hlediska šikmosti a špičatosti, λ se určuje stejně jako u mocninné transformace,
  - takto definovaná transformace je použitelná pouze pro kladná data,
  - tvar transformační funkce pro některé parametry uvádí obrázek vpravo:

$$y = g(x) \begin{cases} x^{\lambda} & \lambda > 0 \\ -x^{-\lambda} & \lambda < 0 \\ \ln x & \lambda = 0 \end{cases}$$

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

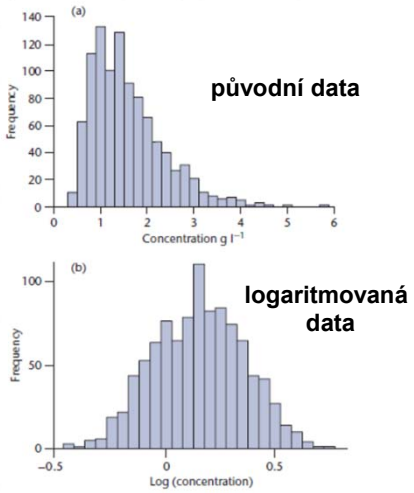


18

Univerzita Palackého v Olomouci

Další druhy transformací

- EXPONENCIÁLNÍ TRANSFORMACE:
  - založena na minimalizaci asymetrie.
- TRANSFORMACE STABILIZUJÍCÍ ROZPTYL:
  - vyžaduje nalezení transformace y = g(x), ve které je $s^2(y)$ konstantní.
- LOGARITMICKÁ TRANSFORMACE:
  - **y = log(x)** případně y = ln(x),
  - symetrizující transformace snadno proveditelná v tabulkovém procesoru,
  - př. na obr.: koncentrace protilátek v krevním séru u mužů.

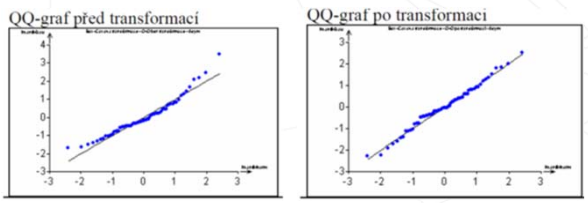


19

Univerzita Palackého v Olomouci

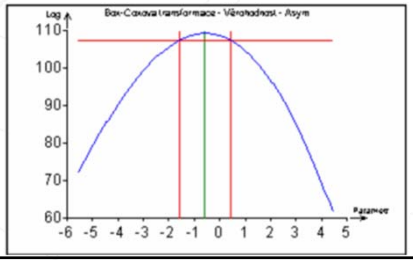
Posouzení statistické přínosnosti transformace

- Q-Q grafy:
  - Porovnání Q-Q grafů pro původní a transformovaná data.
  - Je-li tvar bodů v grafu transformovaných dat blíže přímce než pro původní data, je transformace úspěšná/přínosná.



- Graf logaritmu věrohodnostní funkce na λ:
  - Maximum = optimální parametr λ. Vodorovná přímka odpovídá 95% IS maxima věrohodnosti a svislé přímky odpovídají IS λ, tj. $<\lambda_D, \lambda_H>$. Obsahuje-li tento interval +1, není transformace přínosná.

$$\ln L(\lambda) = -\frac{n}{2}\ln s^2(y) + (\lambda - 1)\sum_{i=1}^{n}\ln x_i$$



20

Univerzita Palackého v Olomouci

Zpětná transformace

- Po vhodné transformaci vyčíslíme $\bar{y}, s^2(y)$ a následně pomocí zpětné transformace odhadneme retransformované parametry původních proměnných, tedy retransformovaný (opravený) průměr $\bar{x}_R$ a retransformovaný rozptyl $s^2_{\bar{x}_R}$. Následně pomocí nich spočteme IS retransformovaného průměru.
- Zpětnou transformaci (retransformaci) na původní proměnnou (data) provádíme, protože požadujeme odhady parametrů původních dat a ne těch transformovaných.
- Lze použít dva přístupy:
  1. nekorektní: provedení prosté zpětné transformace $\bar{x}_R = g^{-1}(\bar{y})$,
  2. korektní: vychází se z Taylorova rozvoje funkce y = g(x) v okolí $\bar{y}$.

- Výsledkem transformace je retransformovaný (opravený) průměr $\bar{x}_R$ a jeho IS. Tento postup vede k lepším odhadům polohy a rozptýlení zvláště pro data z asymetrických rozdělení.

21