PRŮZKUMOVÁ ANALÝZA JEDNOROZMĚRNÝCH DAT

Exploratory Data Analysis (EDA)

- Reprezentativní náhodný výběr:
  1. Prvky výběru x_i jsou vzájemně nezávislé.
  2. Výběr je homogenní, tj. všechna x_i jsou ze stejného rozdělení.
  3. Předpokládá se, že jde o normální rozdělení.
  4. Všechny x_i mají stejnou P, že budou zařazeny do výběru.

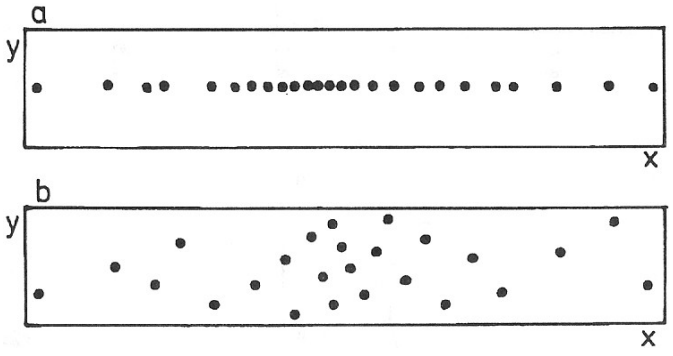
CÍL EDA: nalezení zvláštností statistického chování dat (odchylek od předpokládaného rozdělení).

- **OBECNÝ POSTUP** (vždy dodržujeme!):
  1. diagnostické grafy EDA.
  2. ověření základních předpokladů (ZP) o výběru statistickými testy.
     Jsou-li ZP splněny (4.), nejsou-li splněny (3.)
  3. transformace dat.
  4. vyčíslení parametrů polohy, rozptýlení a tvaru.

- Pořádková statistika – vzestupně setříděné prvky výběru.
- Pořadová pravděpodobnost je definována jako $P_i = i/(n+1)$,
  v EDA se používá: $P_i = \frac{i-1/3}{n+1/3}$.

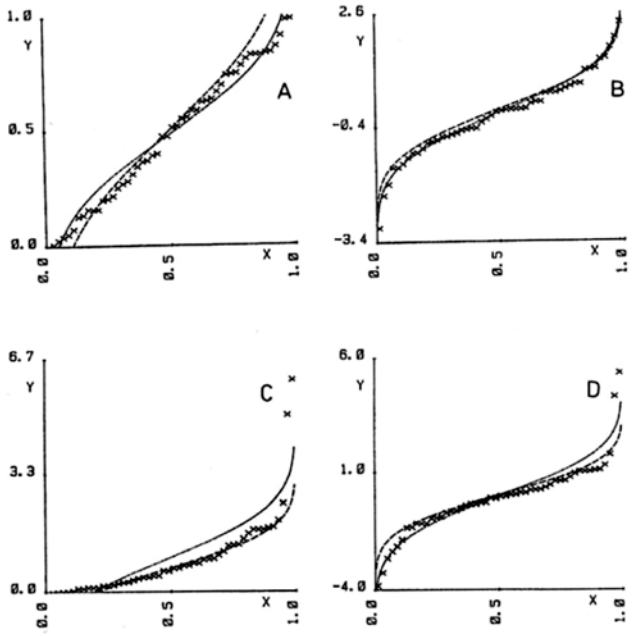
Grafy identifikace statistických zvláštností dat

- **DIAGRAM ROZPTÝLENÍ** (osa x: hodnoty x_i, osa y: nemá význam)
  - Zobrazuje všechna data ve skutečném měřítku na ose x
  - jednorozměrná projekce kvantilového grafu do osy x,
  - ukazuje lokální koncentrace dat a indikuje OB,
  - modifikace – rozmítnutý diagram rozptýlení.



✘ **KVANTILOVÝ GRAF** (osa x: pořadová pravděpodobnost P_i, osa y: pořádková statistika x_i)

- zobrazuje kvantily dat proložené kvantilovou funkcí normálního rozdělení,
- přehledné znázornění dat, tvar rozdělení – symetrii,
- ukazuje lokální koncentrace dat a OB,
- *zelená křivka* (průměr a s^2), *červená křivka* (M a mediánová odchylka) – podle toho, která z křivek lépe prokládá data, je vhodné použít buď průměr nebo medián!



Obr. 2.5 Kvantilové grafy (robustní --- a klasické ...) pro výběry z rozdělení (A) rovnoměrného, (B) normálního, (C) exponenciálního a (D) Laplaceova

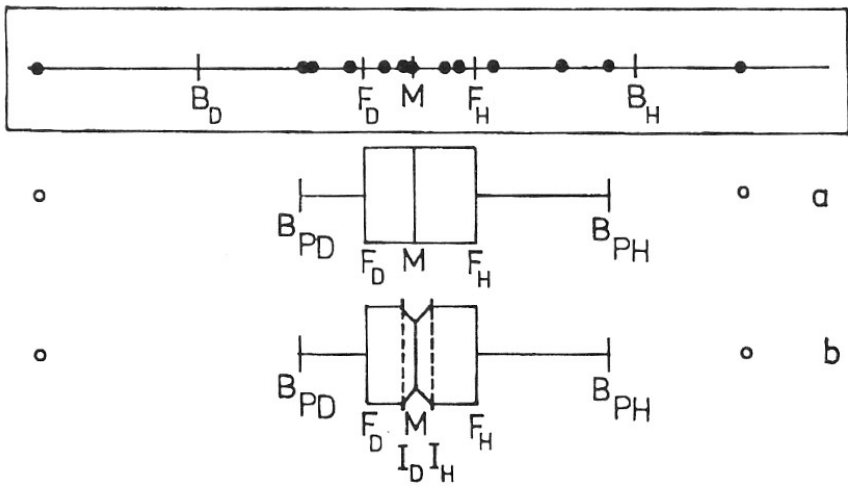
✘ **KRABICOVÝ GRAF** (osa x: úměrná hodnotám x_i, osa y: interval)

- znázornění mediánu (robustního odhadu polohy),
- posouzení symetrie v okolí kvartilů a u konců rozdělení,
- identifikaci OB,
- modifikace – vrubový krabicový graf.

Vnitřní hradby: $B_H = F_H + 1{,}5R_F$

$B_D = F_D - 1{,}5R_F$,

kde $R_F = F_H - F_D$, pro data z normálního rozdělení $B_H - B_D \approx 4{,}2$



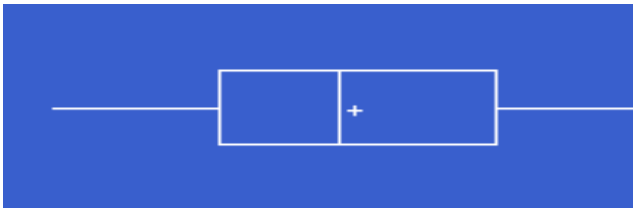
(Robustní) interval spolehlivosti M: $L_{1,2} = M \pm \frac{1{,}5 \cdot R_F}{\sqrt{n}}$

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

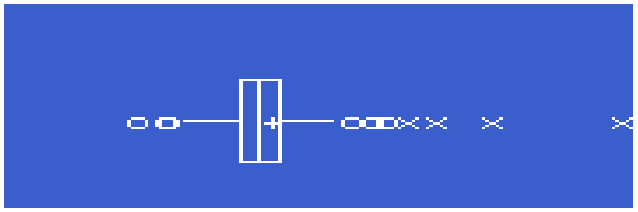
$g_1 > 0$

$g_1 < 0$

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



$g_2 < 3$

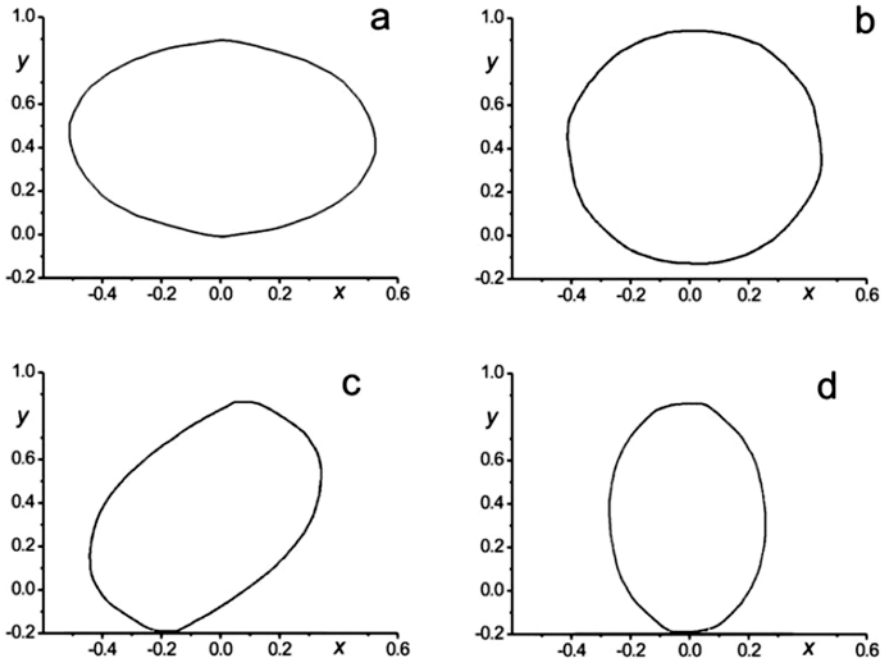


$g_2 > 3$

✘ **GRAF SYMETRIE** – citlivý indikátor symetrie rozdělení. V ideálním případě leží body na horizontální přímce, která představuje M. Přerušované meze = IS mediánu M. Pro asymetrické rozdělení vykazují body výrazný trend rostoucí pro g1 < 0 a klesající pro g1 > 0, překračující přerušované meze.

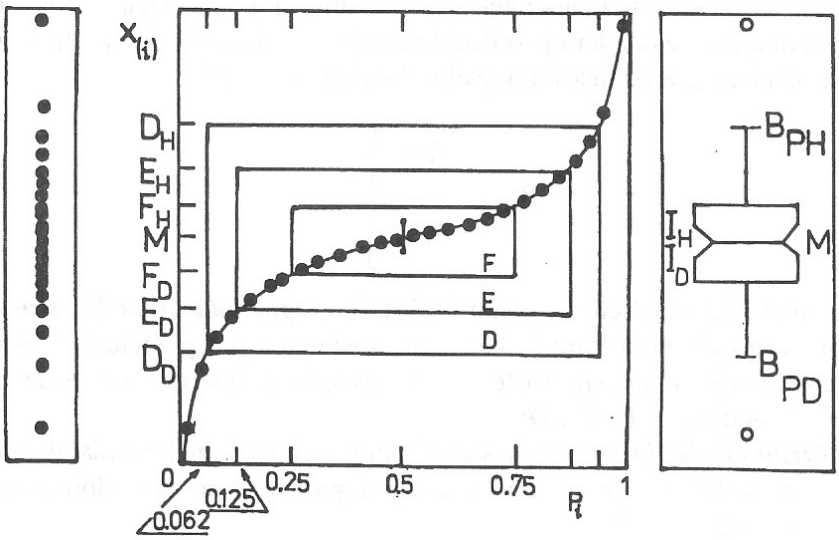
✘ **GRAF ŠPIČATOSTI** – má podobný význam jako předchozí graf. Pro výrazně nenormální špičatost dat vykazují body v grafu výrazný trend.

✘ **KRUHOVÝ GRAF** – slouží k vizuálnímu posouzení normality na základě kombinace šikmosti a špičatosti. (*Zelený* kruh (elipsa) je optimální tvar pro normální rozdělení, *černý* představuje data. Pro normální data se obě křivky „téměř“ kryjí.)



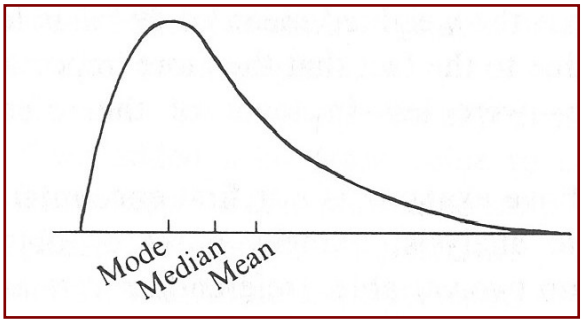
Kruhový graf pro výběry z rozdělení a) rovnoměrného, b) normálního, c) exponenciálního a d) Laplaceova.

✘ **GRAF ROZPTÝLENÍ S KVANTILY** (osa x: P_i, osa y: pořádková statistika x_i)
- body grafu jsou významově i vizuálně shodné s kvantilovým grafem,
- pro symetrické rozdělení má křivka (kvantilová funkce) sigmoidální tvar,
- vzájemná poloha obdelníků (F, E, D) odpovídá symetrii, resp. asymetrii,
- body mimo obdelník D jsou OB.

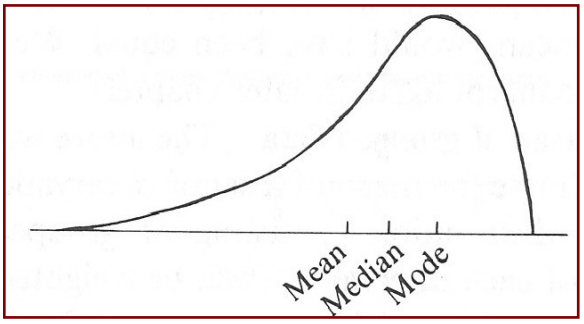


✘ **GRAF HUSTOTY PRAVDĚPODOBNOSTI** (osa x: proměnná x, osa y: (jádrový) odhad hustoty pravděpodobnosti f(x))

- umožňuje identifikovat odchylky v g_1, g_2, nehomogenitu dat a OB,
- v případě normality jsou si *zelená křivka* normálního rozdělení a *červená křivka* experimentálních dat blízké.



$g_1 > 0$



$g_1 < 0$

pro $g_1 = 0$ platí $\bar{X} = \tilde{X}_{0,5} = X_M$

✘ **HISTOGRAM** (osa x: proměnná x, osa y: hustota pravděpodobnosti y)

- posouzení symetrie a špičatosti rozdělení,
- identifikace odlehlých bodů a lokálních koncentrací dat.

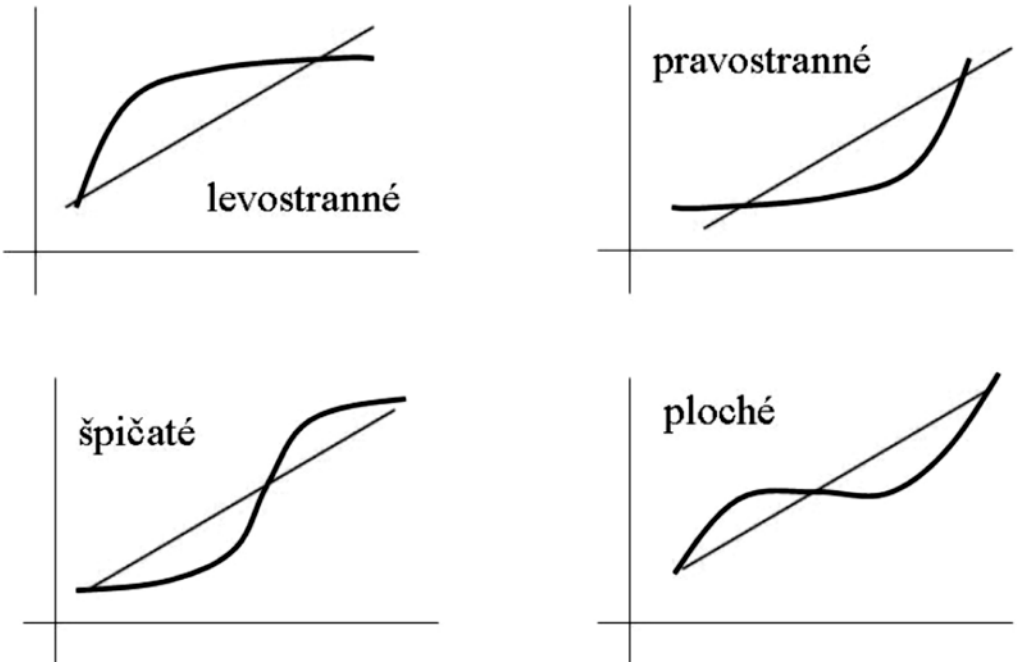
✘ **KVANTIL-KVANTILOVÝ GRAF** (osa x: kvantilová funkce teoretického rozdělení, osa y: pořádková statistika)

- nejpoužívanější diagnostický graf pro porovnání experimentálních dat s teoretickým rozdělením,
- pokud je teoretické rozdělení shodné s experimentálními daty, budou body v grafu ležet na přímce,
- porovnání s normálním rozdělením – RANKITOVÝ GRAF.

✘ **RANKITOVÝ GRAF** (osa x: kvantilová funkce normálního rozdělení, osa y: pořádková statistika)

- umožňuje identifikovat odchylky v g_1 a g_2,
- nalezení OB – body na koncích vzdálené od přímky,
- VÝHODA GRAFU – možnost vizuálního posouzení zda jsou odchylky od normálního rozdělení způsobeny jen několika body, nebo všemi daty.

Diagnostikování rozdělení kvantil-kvantilovým grafem



✘ **PRAVDĚPODOBNOSTNÍ 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.