M6120 – 4. cvičení : **M6120cv04** (*EDA Exploratory Data Analysis - průzkumová analýza dat*)

A. Úvod

Účelem průzkumové analýzy dat je

- odhalit zvláštnosti dat,
- ověřit předpoklady pro následné statistické zpracování.

V průzkumové analýze dat se užívají především **grafické metody** a ty pak slouží pro

- zjednodušení popisu dat,
- identifiikaci typu rozdělení výběru,
- pro zlepšení rozdělení dat.

Při průzkumové analýze dat se využívá především **robustních kvantilových charakteristik**, které umožňují sledování lokálního chování dat. Mezi základní statistické charakteristiky patří

- stupeň symetrie a špičatosti rozdělení výběru
- lokální koncentrace dat
- přítomnost vybočujících dat

B. Typy dat:

(1) **Nominální** proměnná je taková, o jejíž dvou hodnotách můžeme pouze říci, zda jsou stejné či různé (škola, fakulta, obor, krevní skupiny: A, B, O, A/B). Hodnotami mohou být texty (písmena), případně i číselné kódy.

(2) **Ordinální (pořadová)**, u jejíž dvou hodnot můžeme navíc určit pořadí (úroveň spokojenosti, vzdělání). Jako hodnoty lze použít text, datum, číslo.

(3) **Intervalová (rozdílová)** proměnná je taková, pro jejíž dvě hodnoty můžeme navíc (k možnostem ordinální proměnné) vypočítat, o kolik je jedna hodnota větší (resp. menší) než druhá (měsíční příjem domácnosti, počet dětí v rodině). Hodnotami jsou tedy čísla.

(4) **Poměrová (podílová)** proměnná je ta, pro jejíž dvě hodnoty můžeme navíc (k možnostem intervalové proměnné) vypočítat, kolikrát je jedna hodnota větší (resp. menší) než druhá, tzn. jedná se pouze o kladné hodnoty (počet členů domácnosti).

Nominální a *ordinální* proměnné jsou souhrnně označovány jako **kvalitativní**; *intervalové* a *poměrové* proměnné jsou souhrnně označovány jako **kvantitativní**.

Kvantitativní proměnné můžeme podle jiného hlediska dělit na

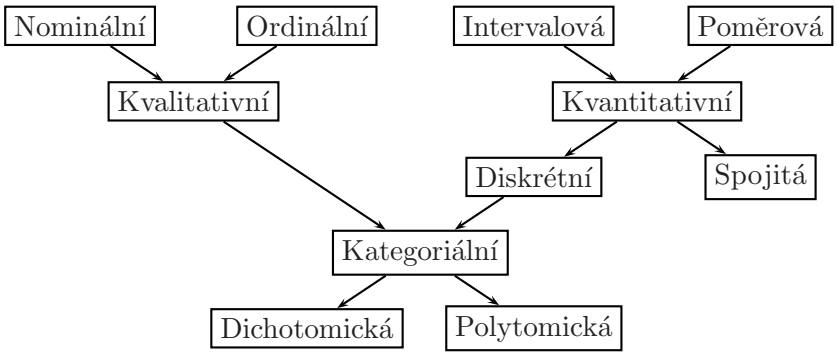
(a) **Diskrétní**, které nabývají pouze celočíselných obměn (počet válců automobilu), a

(b) **Spojité**, jež mohou nabývat libovolných hodnot z určitého intervalu (věk respondenta, cena výrobku, roční příjem domácnosti).

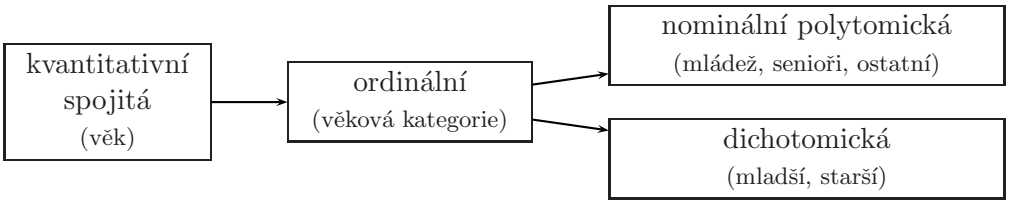
Nominální, *ordinální* a *kvantitativní diskrétní* proměnné můžeme souhrnně označit jako **kategoriální** (obměny těchto proměnných nazýváme kategoriemi). Podle jiného hlediska je můžeme dělit na

(i) **dichotomické (alternativní, binární)**, které nabývají pouze dvou kategorií (ekonomicky aktivní a neaktivní, kuřák a nekuřák), a

(ii) **polytomické (množné)**, jež nabývají více než dvou kategorií (rodinný stav, obor).



Proměnné určitého typu můžeme převádět, např.



C. Základní popisné statistiky

Cílem teorie odhadu je **na základě náhodného výběru** odhadnout

- rozdělení pravděpodobnosti,
- popřípadě některé parametry tohoto rozdělení,
- anebo nalézt odhad nějaké funkce parametrů $\boldsymbol{\theta}$, tj. $\gamma(\boldsymbol{\theta})$.

Rozdělení náhodné veličin X při neznámých parametrech $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ je vyčerpávajícím způsobem určeno distribuční funkcí

$$F(x;\boldsymbol{\theta}) = P(X \leq x) \quad \text{pro } x \in \mathbb{R},\ \boldsymbol{\theta} \in \boldsymbol{\Theta}.$$

Informaci obsaženou v náhodném výběru lze zužitkovat k popisu **distribuční funkce** pomocí **výběrové (empirické) distribuční funkce**. Mějme $\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq F(x;\boldsymbol{\theta})$.

Zaveďme tzv. **indikátor množiny** předpisem: $I_B(x) = \begin{cases} 1 & x \in B, \\ 0 & x \notin B \end{cases}$

a pro $x \in \mathbb{R}$ **indikátor jevu**: $I_i(x) = I_{(-\infty,x>}(X_i) = \begin{cases} 1 & X_i \leq x, \\ 0 & X_i > x. \end{cases}$ pro $i = 1, \ldots, n.$

Potom $I_1(x), \ldots, I_n(x)$ jsou nezávislé náhodné veličiny se stejným alternativním rozdělením pravděpodobností s parametrem $\pi \in (0,1)$, tj. $\perp\!\!\!\perp\{I_1, \ldots, I_n\} \simeq A(\pi)$. Parametr π je roven pravděpodobnosti úspěchu, tj.

$$P(I_i(x) = 1) = P(X_i \leq x) = F(x;\boldsymbol{\theta}) \quad \Rightarrow \quad \boxed{\perp\!\!\!\perp\{I_1, \ldots, I_n\} \simeq A(\pi = F(x;\boldsymbol{\theta}))}.$$

Položme

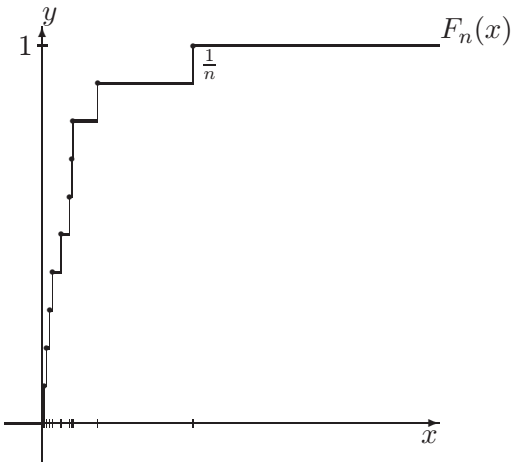
$$\begin{aligned} Y(x) &= \textstyle\sum_{i=1}^{n} I_i(x) \\ F_n(x) &= \frac{Y(x)}{n} \end{aligned}$$

a postupně počítejme

$$\boxed{EF_n(x)} = E\tfrac{Y(x)}{n} = \tfrac{1}{n}Y_n = \tfrac{1}{n}\sum_{i=1}^{n} I_i(x) = \tfrac{1}{n} \cdot n\ F(x;\boldsymbol{\theta}) = \boxed{F(x;\boldsymbol{\theta})}.$$

Protože posloupnost $\{F_n(x)\}_{n=1}^{\infty}$ splňuje jak slabý, tak silný zákon velkých čísel, tak platí

$$\boxed{\begin{aligned} \lim_{n\to\infty} P(|F_n(x) - F(x;\boldsymbol{\theta})| \geq \varepsilon) &= 0 \\ P(\lim_{n\to\infty} F_n(x) = F(x;\boldsymbol{\theta})) &= 1 \end{aligned}}.$$



Z uvedených vztahů je vidět, že pokud rozsah výběru bude dostatečně velký, lze **distribuční funkci rozdělení**, z něhož výběr pochází, **dostatečně přesně aproximovat** pomocí **výběrové (empirické) distribuční funkce**.

Předpokládejme, že rozdělení, z něhož výběr pochází, má konečné druhé momenty se střední hodnotou μ a rozptylem σ^2, což budeme dále značit

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\} \simeq \mathcal{L}(\mu, \sigma^2).$$

Tedy pro každé $i = 1, \ldots, n$ platí

$$\begin{aligned} EX_i &= \mu \\ DX_i &= \sigma^2 \end{aligned}.$$

Potom tyto charakteristiky zřejmě závisí na parametru $\boldsymbol{\theta}$, neboť

$$\begin{aligned} \mu &= \textstyle\int_{-\infty}^{\infty} x dF(x;\boldsymbol{\theta}) \\ \sigma^2 &= \textstyle\int_{-\infty}^{\infty} (x-\mu)^2 dF(x;\boldsymbol{\theta}) \end{aligned},$$

proto bude lépe značit je $\boxed{\mu(\boldsymbol{\theta})}$ a $\boxed{\sigma^2(\boldsymbol{\theta})}$ místo μ a σ^2.

Všimněme si dále, že pro každé $x \in \mathbb{R}$ je $\boxed{F_n(x) = F_n(X_1, \ldots, X_n)}$ statistikou, tím také náhodnou veličinou (která nabývá hodnot mezi nulou a jedničkou) a tím i funkcí elementárního jevu $\omega \in \Omega$.

Zvolíme-li ω libovolně, ale pevně a uvažujeme-li $\boxed{F_n(x)}$ jako funkci proměnné x, pak lze snadno odvodit, že je tato funkce **distribuční funkcí** nějaké náhodné veličiny a lze zavést její střední hodnotu a rozptyl

$$\begin{aligned} \mu_n &= \int_{-\infty}^{\infty} x dF_n(x;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} X_i \\ \sigma_n^2 &= \int_{-\infty}^{\infty} (x-\mu)^2 dF(x;\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n}(X_i-\mu_n)^2 \end{aligned} .$$

Zřejmě μ_n a σ_n^2 jsou borelovské funkce náhodného výběru a tedy statistiky a lze je považovat za odhady parametrických funkcí $\mu(\boldsymbol{\theta})$ a $\sigma^2(\boldsymbol{\theta})$. Lze očekávat, že čím bude rozsah náhodného výběru větší, tím bude odhad uvedených parametrických funkcí kvalitnější.

Připomeňme si další momentové charakteristiky a z nich odvozené číselné charakteristiky (závislost na neznámém parametru $\boldsymbol{\theta}$ pro přehlednost dále vypustíme)

obecné momenty	$\mu'_k = EX^k$	pro $k \in \mathbb{N}$
centrální momenty	$\mu_k = E(X-EX)^k$	pro $k \in \mathbb{N}$
střední hodnota	$\mu = \mu'_1$	
rozptyl	$\sigma^2 = \mu_2$	
směrodatná odchylka	$\sigma = \sqrt{\sigma^2}$	
šikmost (*skewness*)	$\gamma_1 = \frac{\mu_3}{\mu_2^{3/2}} = \frac{\mu_3}{\sigma^3}$	
špičatost (*kurtosis*)	$\gamma_2 = \frac{\mu_4}{\mu_2^2} = \frac{\mu_4}{\sigma^4}$	
popř. tzv. *excess kurtosis*	$\gamma_2^* = \frac{\mu_4}{\mu_2^2} - 3 = \frac{\mu_4}{\sigma^4} - 3$	

kterým odpovídají dva nejdůležitější výběrové protějšky (vypočítané z náhodného výběru $X_1, \ldots, X_n$, resp. z jejich realizací $x_1, \ldots, x_n$)

výběrový průměr	$\bar{x} = \widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$
a výběrová směrodatná odchylka	$s = \widehat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$.

Velmi důležité jsou také **kvantily**, které definujeme pomocí kvantilové funkce

$$Q(p) = \inf\{x \in \mathbb{R} : F(x) = P(X \le x) \ge p\} \qquad \text{kde } p \in (0,1),$$

a to především

dolní kvartil	$x_{0.25} = Q_1 = Q(0.25)$
medián	$x_{0.5} = Q_2 = Q(0.5)$
horní kvartil	$x_{0.75} = Q_3 = Q(0.75)$
interkvartilé rozpětí	$IQR = x_{0.75} - x_{0.25}$

Empirická (výběrová) kvantilová funkce

Je definována pomocí náhodného výběru

$$\perp\!\!\!\perp\{X_1, \ldots, X_n\}$$

takto

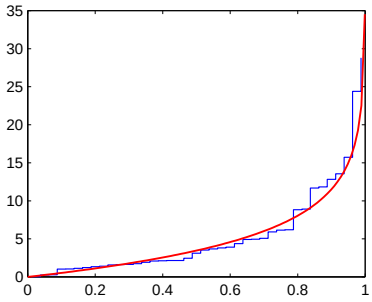
$$Q_{emp}(p_i) = X_{(i)} \quad \text{pro} \quad p_i = \frac{i-\frac{1}{2}}{n},$$

kde

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

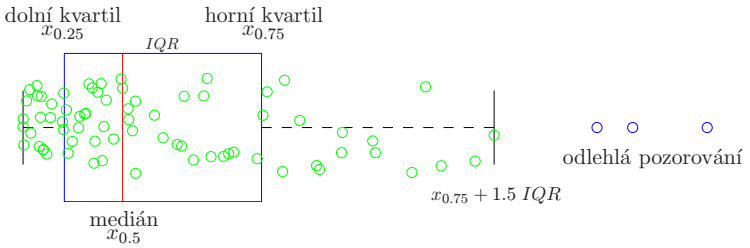
jsou tzv. **pořádkové statistiky**, tj. uspořádaný náhodný výběr.

Teoritická a empirická kvantilová funkce exponenciálního rozdělení



D. Grafické prostředky průzkumové analýzy dat

- ***NOMINÁLNÍ ZNAKY***: *sloupcový* nebo *koláčový* graf četností znaků
- ***ORDINÁLNÍ ZNAKY***:
  - *diagram rozptýlení*, kdy se jednorozměrná data zobrazují rozptýleně (jittering), aby byla patrna četnost jednotlivých hodnot znaků.
  - *krabicový graf* (*box plot*, *box and whisker plot*)
    - obyčejný



    - vrubový (s intervalem spolehlivosti pro medián).
- ***KVANTITATIVNÍ ZNAKY***:
  - *HISTOGRAM* je nejstarším a v praxi patrně nějpoužívanějším neparametrickým odhadem hustoty.

    Při jeho konstrukci rozložíme reálnou osu na stejně dlouhé intervaly (s výjimkou krajních) a na každém z nich neznámou hustotu odhadneme konstantou rovnou podílu počtu pozorování, jež do daného intervalu padnou, k hodnotě nh_n, kde h_n je délka intervalu a n je celkový počet dat.

    Největším problémem histogramu je **optimální volba počtu subintervalů** k_{opt} a jejich délky h_{opt}:
    - Sturgersovo pravidlo $\boxed{k_{opt} = \lceil 1 + \log_2 n \rceil \doteq \lceil 1 + 3.3 \log_{10} n \rceil}$ kde $\lceil x \rceil$ značí celočíselnou část čísla x plus jedna.

* Jestliže však data nejsou normální, ale zešikmená nebo s jinou špičatostí než normální, Doane (1976) doporučuje zvýšit počet subintervalů o
$$\boxed{\lceil \log_2(1 + \widehat{\gamma}_1 \sqrt{\tfrac{n}{6}}) \rceil}$$
kde $\widehat{\gamma}_1$ je odhad standardizovaného koeficientu šikmosti.
* Pro přibližně symetrická rozdělení výběru lze k_{opt} počítat podle vztahu
$$\boxed{k_{opt} = \lceil 2\sqrt{n} \rceil}$$
(viz. Meloun, Militký (1994)).
* V širokém rozmezí velikostí výběrů je možné užít výraz
$$\boxed{k_{opt} = \lceil 2.46(n-1)^{0.4} \rceil}$$
(viz. Meloun, Militký (1994)).
* Pokud se neočekává příliš sešikmené rozdělení výběru, je výhodné volit konstatní délku třídních intervalů h_{opt} a Scott(1979) navrhuje jako optimální volit $\boxed{h_{opt} = 3.5\hat{\sigma} n^{-\frac{1}{3}}}$, kde $\hat{\sigma}$ je nějaký odhad směrodatné odchylky.
* Freedman a Diaconis (1981) navrhují robustnější odhad $\boxed{h_{opt} = 2\ IQR\ n^{-\frac{1}{3}}}$ kde IQR je **interkvartilové rozpětí**.
* Pro komplikovanější tvary výběrových rozdělení je třeba zvětšit počet třídních intervalů, nebo případně užít speciálních postupů hledání třídních intervalů nekonstatních délek.
* MATLAB používá v proceduře histfit $\boxed{k_{opt} = \lceil \sqrt{n} \rceil}$
Viz. `nbins = ceil(sqrt(n));`
* Jazyk R umožňuje volby: *Sturgersovo pravidlo* (`breaks="Sturges"` – implicitní volba), další možností je *Scottovo pravidlo* (`breaks="Scott"`), popř. *Freedmanovo–Diaconisovo pravidlo* (`breaks="Freedman-Diaconis"` nebo zkráceně `breaks="FD"` – doporučená volba).

– ***KVANTILOVÉ GRAFY*** (*Q–Q grafy*, *Q–Q plots*, *Quantile–quantile plots*) pomocí kterých můžeme např. porovnávat
* teoretické a výběrové kvantily,
* kvantily dvou výběrů.

Kontrola normality: v tomto případě se Q–Q graf nazývá také ***rankitový graf***. Je-li $X_{(1)}, \ldots, X_{(n)}$ setřídný náhodný výběr ze standardizovaného normálního rozdělení $N(0,1)$ s distribuční funkcí $\Phi(x)$. Pak

$$EX_{(i)} \approx \eta_i = \Phi^{-1}\left(\frac{i - 3/8}{n + 1/4}\right) \qquad \text{viz. Blom (1958).}$$

Je-li $X_{(1)}, \ldots, X_{(n)}$ setřídný náhodný výběr z normálního rozdělení $N(\mu, \sigma^2)$, pak

$$EX_{(i)} \approx \mu + \sigma\eta_i$$

a graf realizací $x_{(i)}$ oproti η_i by přibližně měla být přímka.

Rankitový graf tedy umožňuje orientační zařazení výběrového rozdělení do skupin podle šikmosti, špičatosti a délky konců. Konvexní, resp. konkávní, průběh Q–Q grafu zde indikuje sešikmené rozdělení výběru, zatímco esovitý průběh ukazuje na rozdílnost v délce konců ve srovnání s normálním rozdělením. Je možné indikovat i směs normálních rozdělení nebo přítomnost vybočujících bodů.

– ***PRAVDĚPODOBNOSTNÍ GRAF (P–P plot)***: pravděpodobnostní grafy jsou alternativou ke Q–Q grafům. Slouží k porovnání distribuční funkce výběru (vyjádřené přes pořadovou pravděpodobnost) se standardizovanou distribuční funkcí zvoleného teoretického rozdělení. V případě normálního rozdělení se P–P graf nazývá také *normal probability plot*.

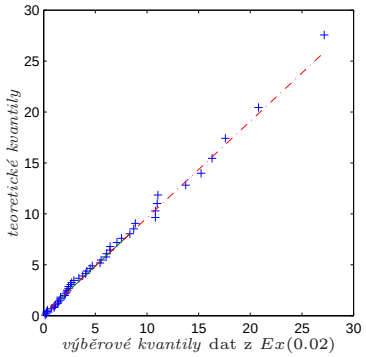
Na následujících třech obrázcích budeme demonstrovat použití Q–Q grafů pro simulovaná data z exponenciálního, Poissonova a normálního rozdělení.

Pokud jsou generovaná data ze stejné rodiny rozdělení, body leží zhruba na přímce a platí

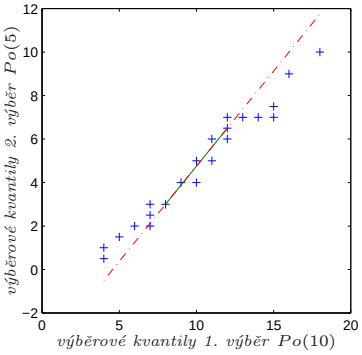
$$X_{(i)} \approx Q(p_i) = F^{-1}(p_i) \quad \text{pro} \quad X \sim F(x) \quad \text{a} \quad Y_{(i)} \approx a + bQ(p_i) \quad \text{pro} \quad Y \sim F\left(\tfrac{x-a}{b}\right).$$

Pocházejí-li z různých rozdělení, část bodů leží výrazně mimo přímku.

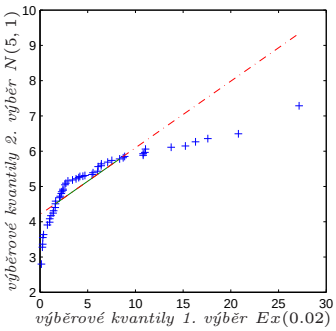
Exponenciální rozdělení



Poissonovo rozdělení



Exponenciální a normální rozdělení



Příklad 1: Posouzeni výsledků Amthaureova IQ testu

Ve 20. století vznikly s rozvojem zkoumání osobnosti stovky testů inteligence. Mnohé z nich však zapadly v zapomnění a dnes se již nepoužívají. Současná psychologie většinou využívá testů inteligence jen několik, které se ukázaly být nejpřesnější a nejkvalitnější. Mnohé z nich jsou sice desetiletí staré, ale důkladně prověřené a neustále vylepšované.

Test struktury inteligence publikoval v roce 1953 Rudolf Amthauer. Test inteligence si klade za cíl nejen určit hodnotu IQ, ale také zjistit strukturu inteligence (zda převažuje inteligence slovní, vizuální, prostorová, apod.).

Test je určen osobám starším 13 let. Lépe se hodí pro testování osob s průměrným a nadprůměrným intelektem, protože je obtížný a náročný na rychlost myšlení i pozornost.

Test je složen z 9 typů otázek. Je tvořen úkoly, kde se pracuje se slovy – doplňuje se vhodné slovo do věty, vyřazuje se slovo, které nepatří mezi ostatní, hledají se analogická

slova a tvoří se pojmy nadřazené. Nalezneme zde také úlohy týkající se matematických schopností – početní úlohy a číselné řady. A nakonec jsou zde obsaženy úkoly na prostorovou představivost a 2D inteligenci.

Popis dat je v souboru `s204.inf`, samotná data jsou uložena v souboru `s204.txt`.

Nejprve načteme popisný soubor pomocí příkazu `readLines`. Díky tomu, že je příkaz v závorkách, ihned se zobrazí obsah proměnné `popis`.

```
> fileTxt <- paste(data.library, "s204.inf", sep = "")
> con <- file(fileTxt)
> (popis <- readLines(con))
```

```
[1] "S204: Posouzeni vysledku Amthaureova testu u 98 studentu"
[2] "[1] skore Amthaureova testu"
[3] "V ramci prijimaciho rizeni absolvuji uchazeci o studium na vysoke skole "
[4] "Amthaueruv test struktury inteligence. Vysledky tohoto testu se vyjadruji"
[5] "prostrednictvim  tzv. hrubeho skore. Ze studentu prijatych ke studiu"
[6] "behem 4 let byl proveden nahodny vyber 98 studentu."
[7] ""
```

```
> close(con)
```

Nyní načteme datový soubor `s204.txt` pomocí příkazu `scan` a vytvoříme proměnnou typu `vector`.

```
> fileDat <- paste(data.library, "s204.txt", sep = "")
> skore <- scan(fileDat)
> str(skore)
```

```
 num [1:98] 77 105 110 88 128 104 94 104 129 96 ...
```

Protože máme k dispozici kvantitativní data, budeme u nich zkoumat polohu, variabilitu, šikmost, špičatost, typ rozdělení, apod.

Velmi stručnou charakteristiku dat (hlavně co do charakteristiky polohy) dostaneme v R pomocí příkazu `summary`

```
> summary(skore)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   72.0   101.2   109.0   109.9   120.0   148.0
```

Zjistíme tak

- minimální a maximální hodnoty (viz příkazy `min` a `max`)

  ```
  > min(skore)
  ```

  ```
  [1] 72
  ```

  ```
  > max(skore)
  ```

```
[1] 148
```

- průměr (příkaz `mean`),

```
> mean(skore)

[1] 109.8776
```

- dále výběrový medián (příkaz `median`)

```
> median(skore)

[1] 109
```

- výběrový dolní a horní kvartil (příkazy `quantile(skore,0.25)` a `quantile(skore,0.75)`)

```
> quantile(skore, 0.25)

   25%
101.25

> quantile(skore, 0.75)

75%
120
```

Variabilitu dat zjistíme

- pomocí výběrového rozptylu či výběrové směrodatné odchylky (má stejnou měrnou jednotku jako vstupní data)

```
> var(skore)

[1] 276.0467

> sd(skore)

[1] 16.61465
```

- pomocí rozsahu dat, což je hodnota minima a maxima (velmi hrubá míra variability)

```
> range(skore)

[1]  72 148

> diff(range(skore))

[1] 76
```

- pomocí interkvartilového rozpětí (robustní míra variability)

```
> diff(quantile(skore, probs = c(0.25, 0.75)))

  75%
18.75
```

Šikmost a špičatost vypočítáme pomocí příkazů, které získáme z balíčku `moments`

```
> library(moments)
> skewness(skore)


[1] -0.02934328
```

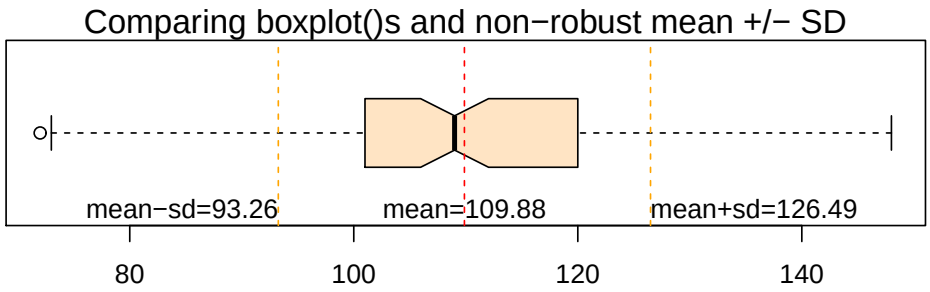
```
> kurtosis(skore)
```

```
[1] 2.931691
```

Pomocí krabicového grafu můžeme snadno posoudit polohu, variabilitu dat, odlehlá pozorování.

Do našeho grafu navíc ještě přidáme polohu výběrového průměru a výběrový průměr plus/minus výběrová směrodatná odchylka.

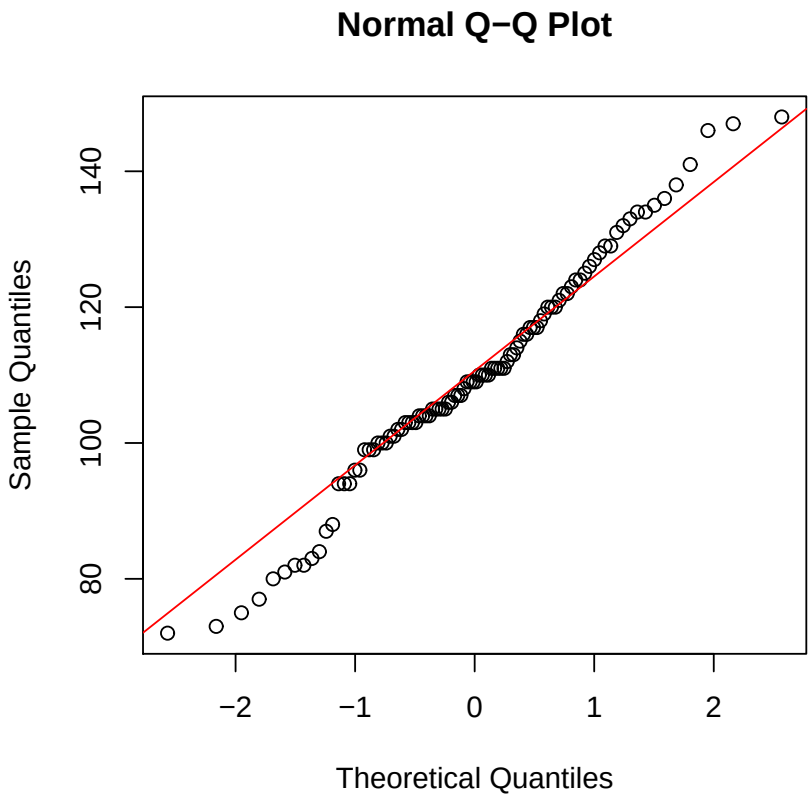
```
> rb <- boxplot(skore, notch = TRUE, horizontal = TRUE, col = "bisque")
> mtext("Comparing boxplot()s and non-robust mean +/- SD", side = 3, line = 0.04,
    cex = 1.25)
> Mean <- mean(skore)
> SD <- sd(skore)
> abline(v = Mean, col = "red", lty = 2)
> abline(v = Mean - SD, col = "orange", lty = 2)
> abline(v = Mean + SD, col = "orange", lty = 2)
> mtext(paste("mean=", round(Mean, 2), sep = ""), side = 1, line = -1.04,
    at = Mean)
> mtext(paste("mean-sd=", round(Mean - SD, 2), sep = ""), side = 1, line = -1.04,
    at = Mean - SD, adj = 1)
> mtext(paste("mean+sd=", round(Mean + SD, 2), sep = ""), side = 1, line = -1.04,
    at = Mean + SD, adj = 0)
```



Obrázek 1: Krabicový graf (vrubovaný) pro hodnoty Amthauerova IQ testu.

Normalitu graficky ověříme následujícím způsobem

```
> qqnorm(skore)
> qqline(skore, col = "red")
```

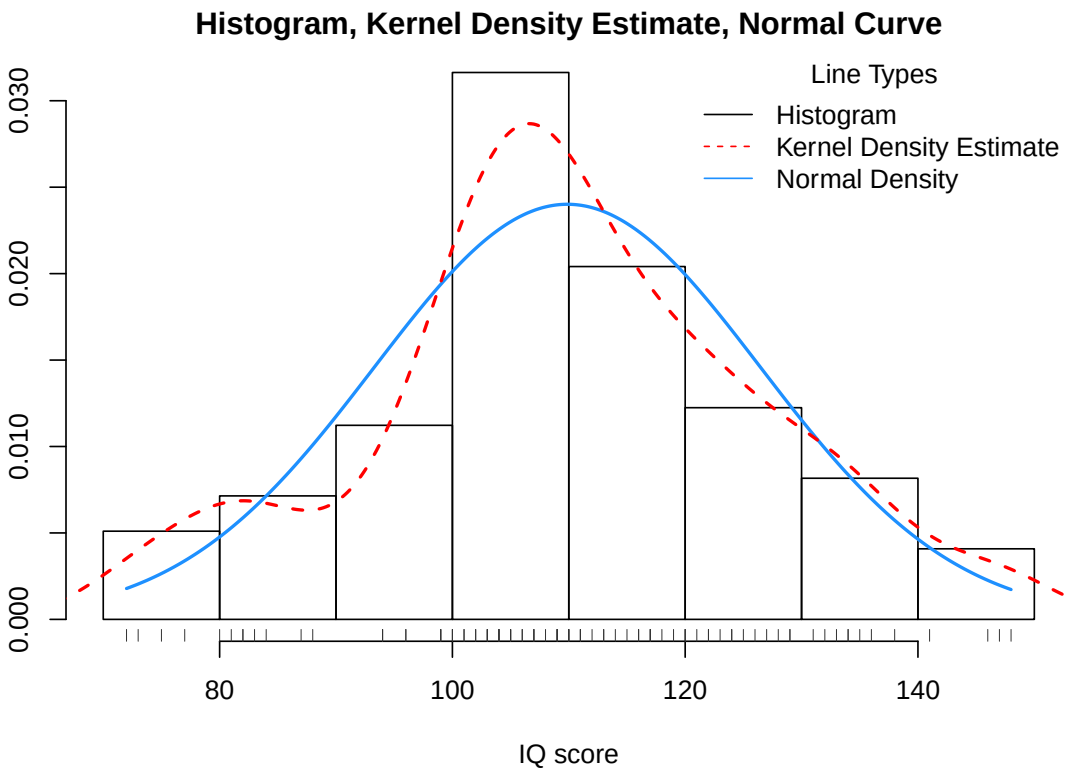


Obrázek 2: Ověření normality pro hodnoty Amthauerova IQ testu.

Vidíme, že nejnižší a nejvyšší hodnoty leží mimo přímku.

Nyní vytvoříme histogram, jádrový odhad hustoty vstupních dat a navíc ještě vykreslíme normální hustotu, kde střední hodnota a rozptyl jsou určeny výběrovými odhady. Tento graf již více odhalí příčiny, proč normalita nedopadla nejlépe.

```
> x <- skore
> par(mar = c(4, 2, 1, 0) + 0.75)
> h <- hist(x, probability = TRUE, breaks = "FD", xlab = "IQ score",
    main = "Histogram, Kernel Density Estimate, Normal Curve")
> xfit <- seq(min(x), max(x), length = 512)
> yfit <- dnorm(xfit, mean = mean(x), sd = sd(x))
> lines(xfit, yfit, col = "dodgerblue", lwd = 2)
> lines(density(x, n = 512), lwd = 2, col = "red", lty = 2)
> rug(x, side = 1, ticksize = 0.02, col = "grey20")
> legend("topright", legend = c("Histogram", "Kernel Density Estimate",
    "Normal Density"), col = c("black", "red", "dodgerblue"),
    lty = c(1, 2, 1), bty = "n", title = "Line Types")
```



Obrázek 3: Histogram, jádrový odhad hustoty, normální hustota pro hodnoty Amthauerova IQ testu.

Příklad 2: Data on Fertility and Contraception in Developing Countries

Popis dat je v souboru `robey.cbk`, samotná data jsou uložena v souboru `robey.dat`.

Nejprve načteme popisný soubor pomocí příkazu `readLines`. Díky tomu, že je příkaz v závorkách, ihned se zobrazí obsah proměnné `popis`.

```
> fileTxt <- paste(data.library, "robey.cbk", sep = "")
> con <- file(fileTxt)
> (popis <- readLines(con))

 [1] "Data on Fertility and Contraception in Developing Countries"
 [2] "[1] Nation"
 [3] "[2] Region"
 [4] "        Africa"
 [5] "        Asia       = Asia and Pacific"
 [6] "        Latin-Amer = Latin America and Caribbean"
 [7] "        Near-East  = Near East and North Africa"
 [8] "[3] Total Fertility Rate"
 [9] "[4] Percent of Contraceptors among Women of Childbearing Age"
[10] "Source of data: Robey, Shea, Rutstein, and Morris (1992) The reproductive"
[11] "revolution: New survey findings. Technical Report M-11, Population Reports."


> close(con)
```

Nyní načteme datový soubor `robey.dat` pomocí příkazu `read.table`. Na rozdíl od příkazu `scan` příkaz `read.table` vytváří ihned datový rámec.

```
> fileDat <- paste(data.library, "robey.dat", sep = "")
> data <- read.table(fileDat, header = FALSE)
```

Proměnné přejmenujeme a vypíšeme prvních a posledních šest pozorování.

```
> names(data) <- c("Nation", "Region", "Tot.Fert.Rate",
    "PercentContraceptors")
> str(data)
```

```
'data.frame':        50 obs. of  4 variables:
 $ Nation             : Factor w/ 50 levels "Bangladesh","Belize",..: 4 6 7 14 22 24 25 26 29 30 ...
 $ Region             : Factor w/ 4 levels "Africa","Asia",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ Tot.Fert.Rate      : num  4.8 6.5 5.9 6.1 6.5 6.4 6.8 2.2 7.3 5.7 ...
 $ PercentContraceptors: int  35 9 16 13 27 6 5 75 4 6 ...
```

```
> head(data)
```

```
    Nation Region Tot.Fert.Rate PercentContraceptors
1 Botswana Africa           4.8                   35
2  Burundi Africa           6.5                    9
3 Cameroon Africa           5.9                   16
4    Ghana Africa           6.1                   13
5    Kenya Africa           6.5                   27
6  Liberia Africa           6.4                    6
```

```
> tail(data)
```

```
    Nation    Region Tot.Fert.Rate PercentContraceptors
45   Egypt Near_East           4.6                   40
46  Jordan Near_East           5.5                   35
47 Morocco Near_East           4.0                   42
48 Tunisia Near_East           4.3                   51
49  Turkey Near_East           3.4                   60
50   Yemen Near_East           7.0                    7
```

Pomocí příkazu `summary` dostaneme jednoduché popisné statistiky.

```
> summary(data)
```

```
       Nation          Region   Tot.Fert.Rate   PercentContraceptors
 Bangladesh: 1   Africa    :18   Min.   :1.700   Min.   : 4.00
 Belize    : 1   Asia      :10   1st Qu.:3.600   1st Qu.:12.25
 Bolivia   : 1   Latin_Amer:16   Median :4.600   Median :41.00
 Botswana  : 1   Near_East : 6   Mean   :4.688   Mean   :37.44
 Brazil    : 1                   3rd Qu.:5.975   3rd Qu.:55.00
 Burundi   : 1                   Max.   :7.300   Max.   :77.00
 (Other)   :44
```

Absolutní a relativní četnosti kvalitativní proměnné `Region` získáme například takto

```
> table(data$Region)

    Africa       Asia Latin_Amer  Near_East
        18         10         16          6

> prop.table(table(data$Region))

    Africa       Asia Latin_Amer  Near_East
      0.36       0.20       0.32       0.12
```

Dále nás zajímají základní popisné statistiky proměnné `Tot.Fert.Rate` za jednotlivé regiony, které lze získat více způsoby, a to buď pomocí příkazu `by` nebo `tapply`.

- Průměry

```
> (MeansFerts <- with(data, by(Tot.Fert.Rate, Region, mean)))

Region: Africa
[1] 5.855556
------------------------------------------------------------
Region: Asia
[1] 3.54
------------------------------------------------------------
Region: Latin_Amer
[1] 4.05
------------------------------------------------------------
Region: Near_East
[1] 4.8

> str(MeansFerts)

 by [1:4(1d)] 5.86 3.54 4.05 4.8
 - attr(*, "dimnames")=List of 1
  ..$ Region: chr [1:4] "Africa" "Asia" "Latin_Amer" "Near_East"
 - attr(*, "call")= language by.default(data = Tot.Fert.Rate, INDICES = Region, FUN = mean)

> mode(MeansFerts)

[1] "numeric"

> class(MeansFerts)

[1] "by"
```

- Mediány

```
> with(data, by(Tot.Fert.Rate, Region, median))

Region: Africa
[1] 6.1
------------------------------------------------------------
Region: Asia
[1] 3.45
------------------------------------------------------------
Region: Latin_Amer
[1] 3.9
------------------------------------------------------------
Region: Near_East
[1] 4.45
```

- Variabilita

```
> with(data, tapply(Tot.Fert.Rate, Region, sd))
```

```
    Africa       Asia Latin_Amer  Near_East
  1.169325   1.285993   0.925923   1.282186
```

- Počet pozorování

```
> (Pocty <- with(data, tapply(Tot.Fert.Rate, Region, length)))
```

```
    Africa       Asia Latin_Amer  Near_East
        18         10         16          6
```

```
> str(Pocty)
```

```
 int [1:4(1d)] 18 10 16 6
 - attr(*, "dimnames")=List of 1
  ..$ : chr [1:4] "Africa" "Asia" "Latin_Amer" "Near_East"
```

```
> mode(Pocty)
```

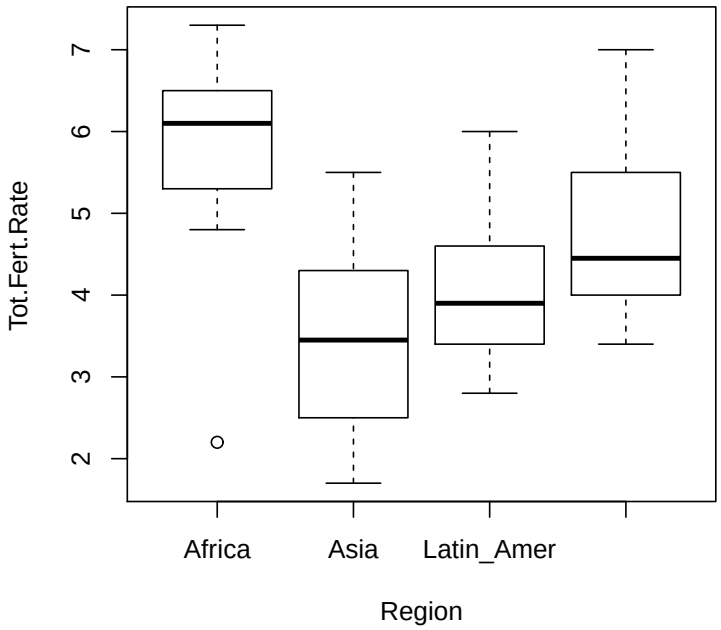
```
[1] "numeric"
```

```
> class(Pocty)
```

```
[1] "array"
```

Všimněme si, jaký graf získáme příkazem

```
> plot(Tot.Fert.Rate ~ Region, data = data)
```



Obrázek 4: Příkaz `plot` v případě kvantitativní versus kvalitativní proměnná pro "robey" data.

A nyní vykreslíme několik různých grafů, a to scatter plot kvantitativních proměnných a po boku jejich histogramy.

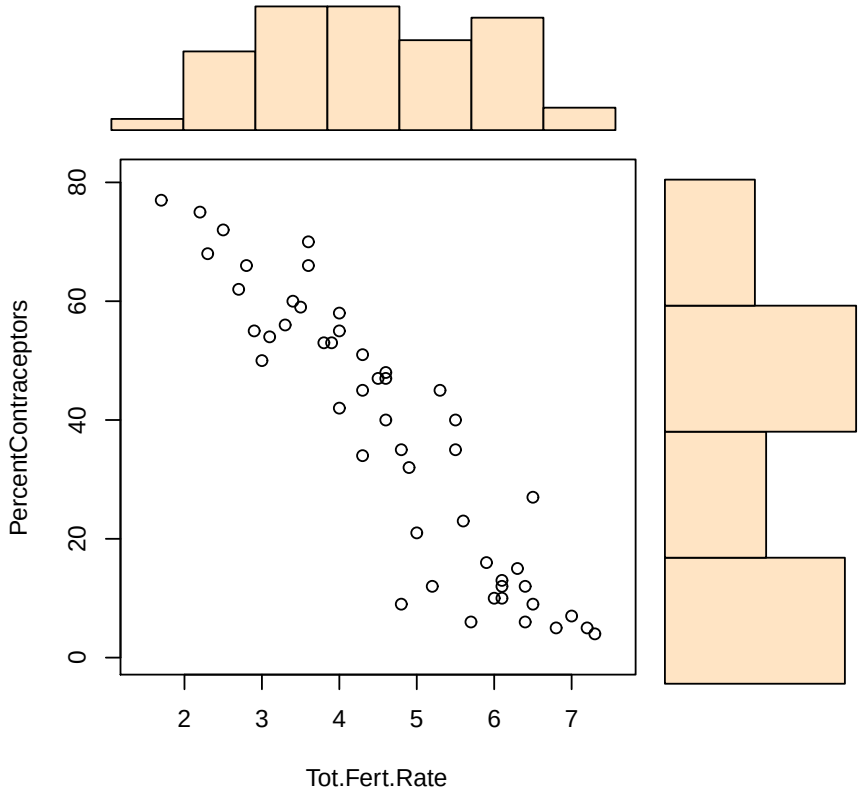
```
> ix <- 3
> iy <- 4
> x <- data[, ix]
> y <- data[, iy]
> def.par <- par(no.readonly = TRUE)
> xhist <- hist(x, breaks = "FD", plot = FALSE)
> str(xhist)
```

```
List of 7
 $ breaks     : num [1:8] 1 2 3 4 5 6 7 8
 $ counts     : int [1:7] 1 7 11 11 8 10 2
 $ intensities: num [1:7] 0.02 0.14 0.22 0.22 0.16 0.2 0.04
 $ density    : num [1:7] 0.02 0.14 0.22 0.22 0.16 0.2 0.04
 $ mids       : num [1:7] 1.5 2.5 3.5 4.5 5.5 6.5 7.5
 $ xname      : chr "x"
 $ equidist   : logi TRUE
 - attr(*, "class")= chr "histogram"
```

```
> yhist <- hist(y, breaks = "FD", plot = FALSE)
> str(yhist)
```

```
List of 7
 $ breaks     : num [1:5] 0 20 40 60 80
 $ counts     : int [1:4] 16 9 17 8
 $ intensities: num [1:4] 0.016 0.009 0.017 0.008
 $ density    : num [1:4] 0.016 0.009 0.017 0.008
 $ mids       : num [1:4] 10 30 50 70
 $ xname      : chr "y"
 $ equidist   : logi TRUE
 - attr(*, "class")= chr "histogram"
```

```
> top <- max(c(xhist$counts, yhist$counts))
> xrange <- c(range(x)[1] - 0.05 * diff(range(x)), range(x)[2] +
      0.05 * diff(range(x)))
> yrange <- c(range(y)[1] - 0.05 * diff(range(y)), range(y)[2] +
      0.05 * diff(range(y)))
> layout(matrix(c(2, 0, 1, 3), 2, 2, byrow = TRUE), width = c(3,
      1), height = c(1, 3), respect = TRUE)
> par(mar = c(4, 4, 1, 1))
> plot(x, y, xlim = xrange, ylim = yrange, xlab = names(data)[ix],
      ylab = names(data)[iy])
> par(mar = c(0, 3, 1, 1))
> barplot(xhist$counts, axes = FALSE, ylim = c(0, top), space = 0,
      col = "bisque")
> par(mar = c(3, 0, 1, 1))
> barplot(yhist$counts, axes = FALSE, xlim = c(0, top), space = 0,
      horiz = TRUE, col = "bisque")
> par(def.par)
```



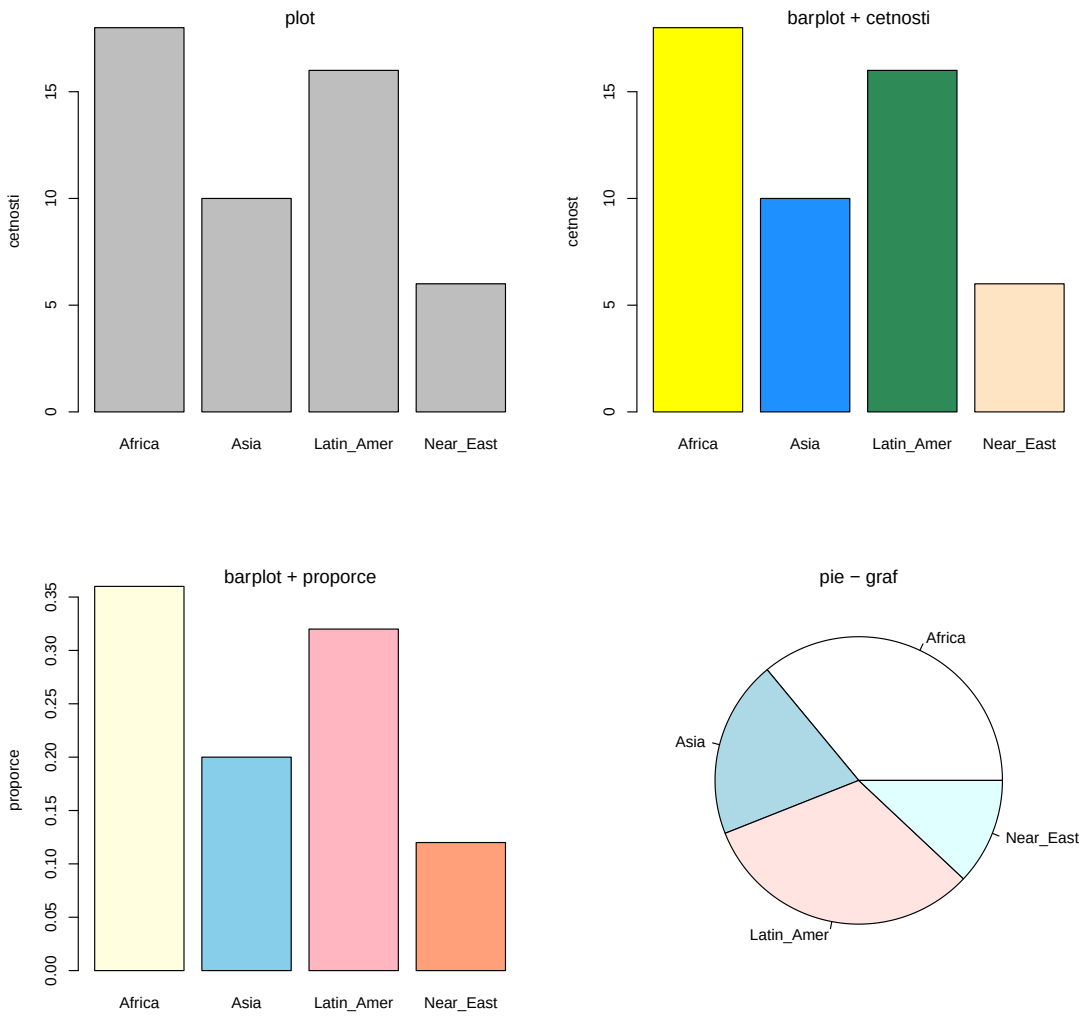
Obrázek 5: Scatter plot a marginální histogramy kvantitativních proměnných z "robey" dat.

Pro kvalitativní proměnnou `Region` vykreslíme nejprve tři různé sloupcové grafy a nakonec také koláčový graf.

Všimněme si, že pro kvantitativní proměnnou lze pomocí příkazu `plot` vytvořit sloupcový graf.

V příkazu `barplot` pracujeme s tabulkami absolutních, popř. relativních četností.

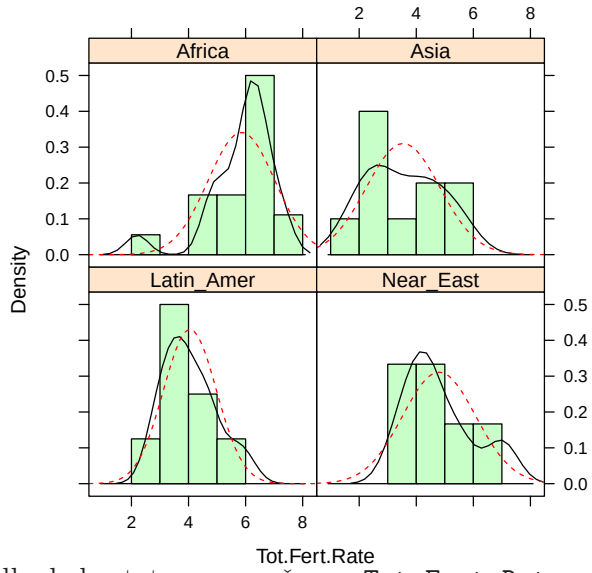
```
> par(mfrow = c(2, 2), bty = "n")
> plot(data$Region, ylab = "cetnosti")
> mtext("plot")
> barplot(table(data$Region), ylab = "cetnost", col = c("yellow",
    "dodgerblue", "seagreen", "bisque"))
> mtext("barplot + cetnosti")
> barplot(prop.table(table(data$Region)), ylab = "proporce", col = c("lightyellow",
    "skyblue", "lightpink", "lightsalmon"))
> mtext("barplot + proporce")
> pie((table(data$Region)))
> mtext("pie - graf")
```



Obrázek 6: Různé grafy pro proměnnou `Region` v "robey" data.

Nyní použijeme poněkud složitější příkaz `histogram` z knihovny `lattice`, který se hodí, chceme–li znázornit histogram kvantitativní proměnné `Tot.Fert.Rate` pro různé varianty kvalitativní proměnné `Region`. Do jednotlivých panelů navíc přidáme odhady normální hustoty a jádrový odhad hustoty.

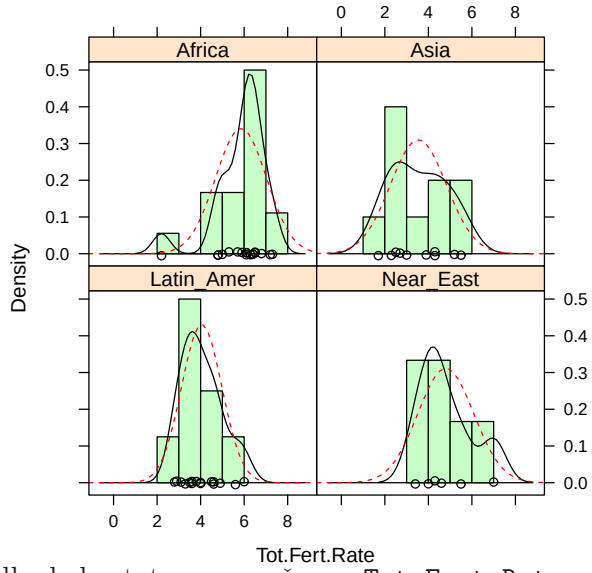
```
> library(lattice)
> print(histogram(~Tot.Fert.Rate | Region, data, as.table = TRUE,
     breaks = "FD", type = "density", panel = function(x, ...) {
         panel.histogram(x, ...)
         panel.densityplot(x, col = "black", ...)
         panel.mathdensity(dmath = dnorm, col = "red", lty = 2,
             args = list(mean = mean(x), sd = sd(x)))
     }))
```



Obrázek 7: Různé odhady hustot pro proměnnou `Tot.Fert.Rate` podle proměnné `Region` v "robey" data – 1. způsob.

Téměř stejný graf dostaneme následujícím způsobem

```
> library(lattice)
> print(densityplot(~Tot.Fert.Rate | Region, data, as.table = TRUE,
     panel = function(x, ...) {
         panel.histogram(x, breaks = "FD", type = "density", ...)
         panel.densityplot(x, col = "black", ...)
         panel.mathdensity(dmath = dnorm, col = "red", lty = 2,
             args = list(mean = mean(x), sd = sd(x)))
     }))
```



Obrázek 8: Různé odhady hustot pro proměnnou `Tot.Fert.Rate` podle proměnné `Region` v "robey" data – 2. způsob.

E. Úkol:

(a) Načtěte soubor informací `s211.inf` a dat `s211.txt`. Prohledněte si oba soubory.

(b) Zjistěte četnosti jednotlivých hodnot IQ.

(c) Vykreslete četnosti jednotlivých hodnot vhodným grafem.

(d) Na základě tabulky četností vykreslete empirickou distribuční funkci.

(e) Vykreslete krabicový graf.

(f) Vykreslete histogram spolu s odhadem normální hustoty a jádrovým odhadem hustoty.

Reference

[1] Blom, G. (1958): *Statistical Estimates and Transformed Beta Variates*, New York: Wiley.

[2] Doane, D.P. (1976): *Aesthetic Frequency Classifications*, Amer. Statist. 30, 181-183.

[3] Freedman, D., Diaconis, P. (1981): *On the Histogram as a Density Estimator: L_2 Theory*, Zeitschrift fur Wahrscheinlichtkeitstheorie und verwandte Gebiete 57, 453-476.

[4] Meloun, M., Militký, J. (1994): *Statistické zpracování experimentálních dat*, edice PLUS.

[5] Scott, D.W. (1979): *On Optimal and Data-Based Histograms*, Biometrika 66, 605-610.

[6] Scott, D.W. (1991): *Multivariate Denstity Estimation*, J. Wiley & Sons.

[7] Sturgers, H.A. (1926): *The Choice of a Class Interval*, J. Amer. Statist. Assoc 21, 65-66.