Teoretická část

1 Základní grafy v R

Pro vytváření základních grafů v R používáme vestavěný balíček `graphics` [8], který obsahuje mnoho užitečných funkcí pro tvorbu grafických prvků. První kapitola se soustředí na tyto funkce tohoto balíčku a v dalších kapitolách jsou popsány funkce balíčků dalších (například `lattice`, `ggplot2`,...), které zastávají podobné funkce, avšak s různým rozsahem nastavení [9].

V následujících příkladech nejsou grafy doplněny o barvy, popisky os, legendy ani názvy a to především proto, že záměrem této kapitoly je popsat základní grafy a funkce pro jejich tvorbu v prostředí R. Všechny tyto prvky mohou být přidány do grafu, ale tím by příkazy obsahovali irelevantní parametry vzhledem k zaměření této kapitoly. Základní funkce `plot(x)` jejímž voláním se obdrží pole s grafickou reprezentaci proměnné "x", by při doplnění kódu o veškeré parametry vypadala následovaně [9]:

```
plot(x, main = "Název grafu", xlab = "popis osy x",
+    ylab = "popis osy y", col = c("red", "black", "green"))
```

Záměrem je tedy používání příkazů s pouze relevantními parametry.

1.1 Bodový graf

Bodový graf je rychlým způsobem, jak znázornit vztahy a souvislosti mezi proměnnými datasetu, případně k zjištění jejich neexistence. Data jsou zobrazeny v kartézském souřadném systému a mají pro každou hodnotu proměnné dané místo na vodorovné a svislé ose. V případě existence závislostí mezi proměnnými lze tuto závislost interpolovat přímkou, křivkou či dalším vhodným vyobrazením této závislosti.

Pro vytvoření bodového grafu v základním prostředí R (pomocí `graphics`) použijeme funkci `plot()`, která má tento typ grafu předdefinovaný pro numerické hodnoty. Viz obrázek 1 (a). Nečíselná data vytvoří jiný typ grafu.

```
plot(cars)
```

1.2 Liniový graf

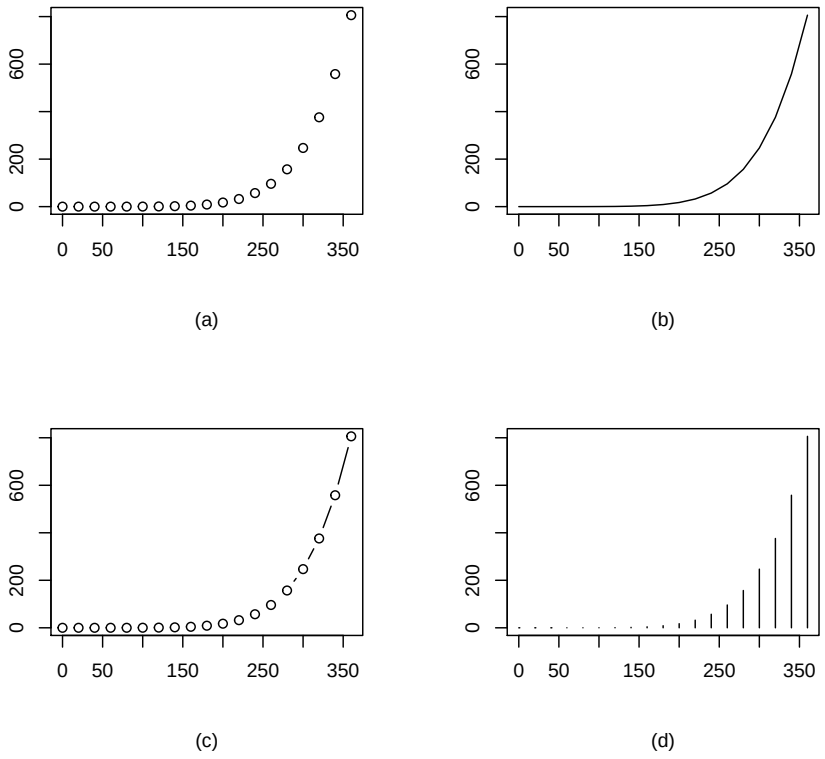
Jediný rozdíl mezi bodovým a liniovým grafem je, že jeden zobrazuje body a druhý je spojuje.[9] (viz. obrázek 1 (a), (b)). Pro vykreslení liniového grafu se používá již několikrát zmíněná funkce `plot()`, kterou doplníme o požadovaný typ vykreslení:

```r
plot(x, type="l")
```

V tabulce 1 jsou uvedené některé základní atributy parametru `type`, které mohou být použity [7]:

	Anglický popis	Český popis
p	points	bodový
l	lines	liniový
b	both	složený
h	histogram	histogram
n	no plotting	bez vykreslení

Tabulka 1: Základní atributy parametru ‘type‘



Obrázek 1: Porovnání základních typů grafů: (a) - bodový, (b) - liniový, (c) - složený, (d) - histogram

Popis a všechny atributy dalších parametrů funkce `plot()` lze nalézt v nápovědě zadáním příkazu `?plot()`.

1.3 Vykreslení rozdělení v R

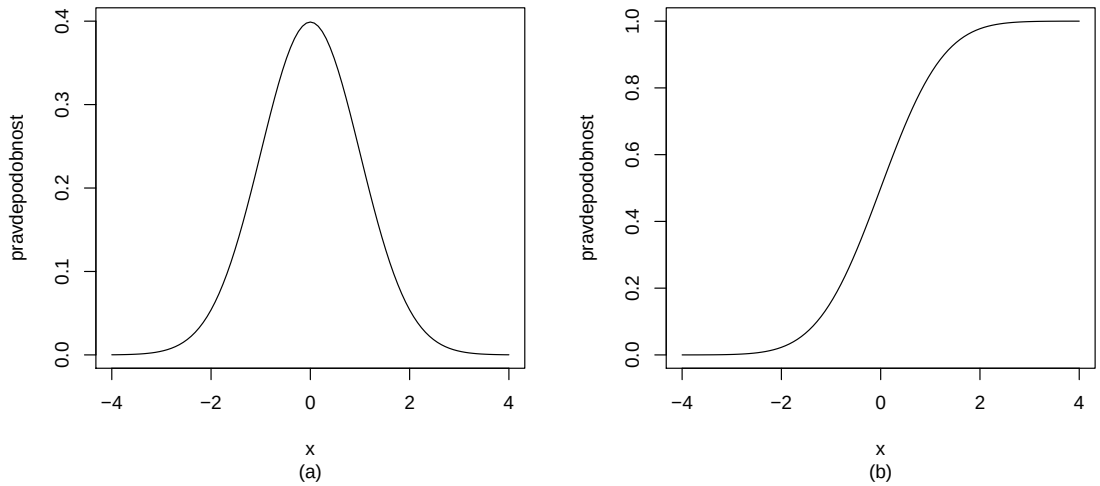
Teorie pravděpodobnosti je základem statistiky a R má hodně nástrojů pro práci s pravděpodobnosti, rozdělením pravděpodobnosti a náhodnými proměnnými. R má zkrácený název pro každé rozdělení pravděpodobnosti. [9] Tyto názvy slouží k identifikaci funkcí spojených s rozděleními. Například zkrácený název "norm" pro normální rozdělení, "exp" pro exponenciální rozdělení a další. Funkce pak mají formu:

Funkce	Účel
dxxxx	Hustota pravděpodobnosti
pxxxx	Distribuční funkce
qxxxx	Kvantilová funkce
rxxxx	Generátor náhodných čísel z daného rozdělení

Tabulka 2: Funkce pro práci s rozděleními

Funkce v R lze vykreslovat pomocí funkce `curve()` z balíčku `graphics`. Lze vykreslit jak standardní funkce, tak i funkce definované uživatelem. Například hustotu pravděpodobnosti normálního rozdělení a její distribuční funkci můžeme vykreslit tímto způsobem (Obrázek 2):

```
curve(dnorm(x))
curve(pnorm(x))
```



Obrázek 2: Hustota pravděpodobnosti normálního rozdělení (a) a její distribuční funkce (b)

1.3.1 Q-Q graf a P-P graf

Q-Q (*quantile-quantile*) graf a P-P (*probability–probability* nebo *percent–percent*) graf (Obrázek 3) se používají hlavně k testování normality při průzkumové analýze dat. Další způsob, jak zjistit zda-li data mají normální rozdělení je sestrojení histogramu (viz. sekce 1.4.1), avšak použití Q-Q grafu je přesnější.

Princip Q-Q grafu spočívá v porovnání dvou rozdělení pravděpodobnosti pomocí vykreslení jejich kvantilů proti sobě. Na jedné ose se nacházejí teoretické kvantily normálního rozdělení a na druhé ose kvantily naměřené (pozorované). Pokud data mají přesně normální rozdělení, všechny body grafu leží na přímce 45°. [9] [2]

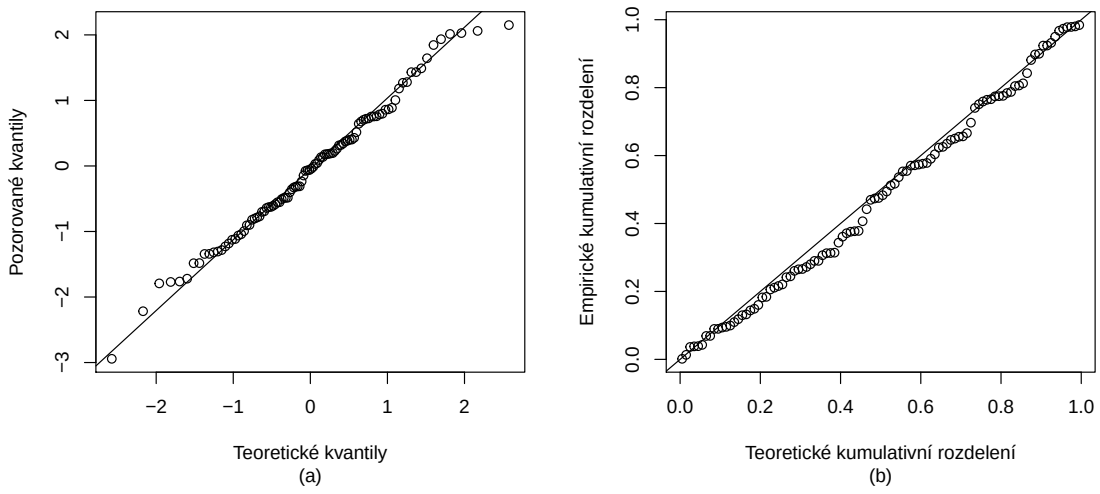
Princip P-P grafu je obdobný jako u Q-Q grafu: vykreslují se dvě distribuční funkcí proti sobě (jedná teoretická a jedná pozorovaná) a pokud všechny body grafu leží přibližně na přímce, jedná se o normální rozdělení. Z velké části se P-P graf používá k vyhodnocení koeficientu šikmosti rozdělení.[6]

V R se Q-Q graf vykreslí takto:

```
qqnorm(x)
qqline(x)
```

P-P graf v R lze vykreslit například následovně:

```
plot(ppoints(length(x)), sort(pnorm(x)))
abline(0,1)
```



Obrázek 3: Q-Q Graf (a) a P-P Graf (b)

1.4 Sloupcový graf

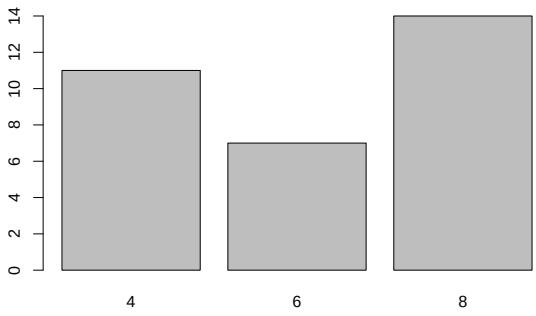
Sloupcový graf je jedním z nejvíce používaných způsobů vizualizace dat. Obvykle se používá pro zobrazení kvantitativních hodnot na ose y a kvalitativních na ose x. Výška sloupců může reprezentovat jak četnosti výskytu hodnot, tak i samotné hodnoty.[1]

V R lze tento typ grafu vykreslit pomocí funkce `barplot()`. V příkladu (Obrázek 4) je použit data set `mtcars`, konkretně atribut `cyl` - počet válců v motoru.

```
table(mtcars$cyl)
```

```
##
##  4  6  8
## 11  7 14
```

```
barplot(table(mtcars$cyl))
```



Obrázek 4: Ukázka jednoduchého sloupcového grafu

1.4.1 Histogram

Sloupcový graf s četnostmi na souvislé ose je taky známý jako histogram.[1] Četnosti mohou být absolutní či relativní. Absolutní četnost zobrazuje počet statistických jednotek s hodnotou znaku, který patří do určitého intervalu. Podíl příslušné četnosti a rozsahu datového souboru se nazývá relativní četnost.[5] Šířka sloupce reprezentuje jednotlivé intervaly, které mají stejnou délku. Pro výpočet optimální délky intervalu existují různé metody. Základní histogram se vytváří pomocí funkci `hist()` a její atribut `breaks` udává buď hranice intervalů, jejich preferovaný počet nebo metodu výpočtu intervalu. V R jsou vestavěny 3 metody výpočtu:

1. Sturges [4]

```
hist(x, breaks = "Sturges")
```

$$k = [log_2(n)] + 1$$

Kde k je počet intervalů a n je počet prvků neboli počet pozorování výběru x. Tato metoda je výchozí pro funkci `hist()`.

2. Scott [4]

```
hist(x, breaks = "Scott")
```

Scotovo pravidlo je následující:

$$h = \frac{3.5\sigma}{n^{\frac{1}{3}}}$$

kde σ je směrodatná odchylka a h je předpokládaná šířka intervalu.

Počet intervalů může být vypočítán pomocí vztahu:

$$k = \left[\frac{max(x) - min(x)}{h}\right]$$

Případně oba vztahy lze shrnout do jednoho:

$$k = \left[n^{\frac{1}{3}} \frac{max(x) - min(x)}{3.5\sigma}\right]$$

3. Freedman–Diaconis [3]

```
hist(x, breaks = "FD")
```

Freedman–Diaconisovo pravidlo pro stanovení předpokládané šířky intervalu je:

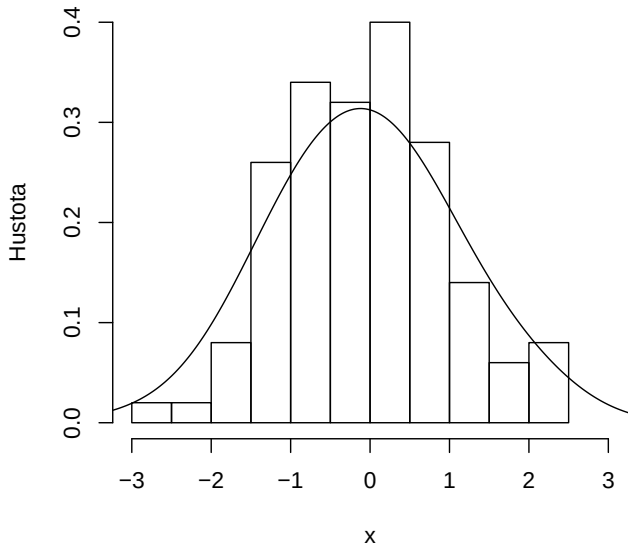
$$h = 2\frac{IQR(x)}{n^{\frac{1}{3}}}$$

Po dosazení:

$$k = \left[n^{\frac{1}{3}} \frac{max(x) - min(x)}{2IQR(x)}\right]$$

kde IQR je mezikvartilové rozpětí, které definujeme jako rozdíl třetího a prvního kvartilů.

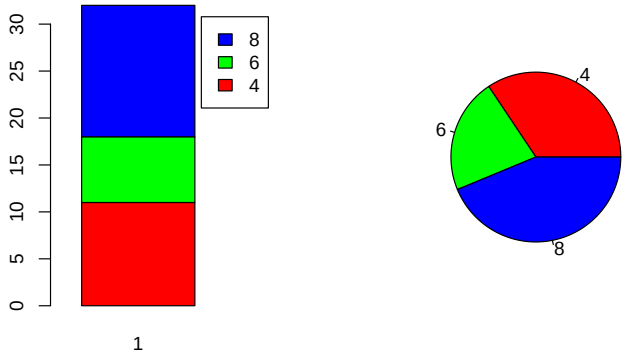
Histogram je jedním ze standardních způsobů, používaných k odhadu tvaru rozdělení, přesto se ale tento způsob považuje za nepřesný, vzhledem k ovlivnění tvaru počtem použitých intervalů. Při normálním rozdělení by měl histogram mít zvoncovitý tvar schodný s Gaussovou křivkou (Obrázek 5).



Obrázek 5: Histogram s odhadem hustoty pravděpodobnosti

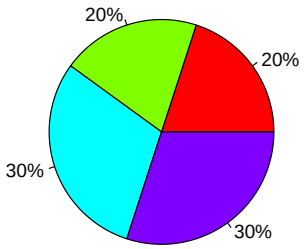
1.4.2 Koláčový graf

Koláčový graf představuje plný kruh (360°), který je rozdělen na jednotlivé výseče pro znázornění číselných proporci mezi proměnnými. Koláčový graf je tvořen transformaci skládaného sloupcového grafu do polárního souřadnicového systému (Obrázek 6). [10]



Obrázek 6: Skládaný sloupcový graf transformovaný do polárního souřadnicového systému

Jednoduché koláčové grafy se vykreslují pomoci funkci `pie()` (Obrázek 7).

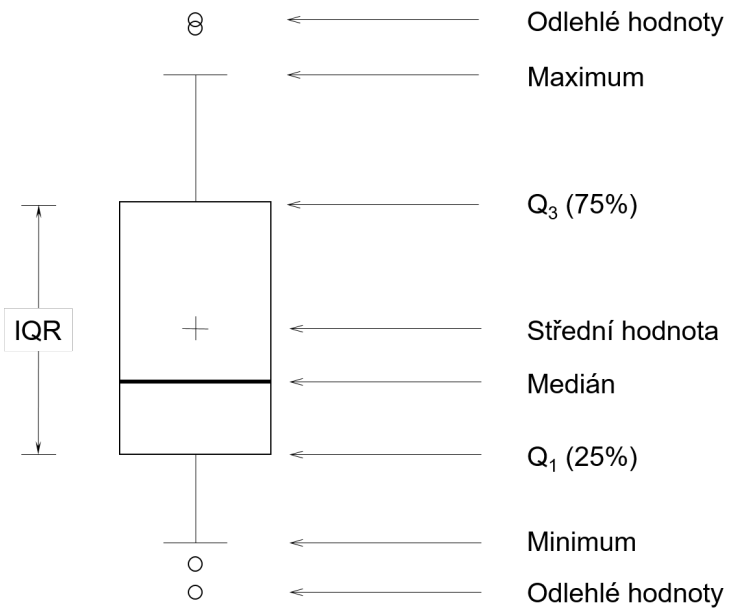


Obrázek 7: Ukázka jednoduchého koláčového grafu

1.5 Krabicový graf

Krabicový graf poskytuje rychlé a jednoduché vizuální shrnutí datasetu a v prostředí R se vykreslí pomocí funkce `boxplot()` z balíčku `graphics`. Obrázek 8 znázorňuje typický krabicový graf, kde silná čára je medián, krabice kolem ní určuje polohu prvního a třetího kvartilů (dolní Q_1 kvantil 25% a horní Q_3 kvantil 75%). "Vousy" (*whiskers*) nad a pod krabici znázorňují rozpětí dat bez odlehlých hodnot. Odlehlé hodnoty jsou definovaný jako hodnoty ležící ve větší vzdálenosti od krabice než 1,5 × IQR, kde IQR je mezikvartilové rozpětí (*interquartile range*) neboli $Q_3 - Q_1$.

```
boxplot(x)
```



Obrázek 8: Boxplot

[1] Chang, W. 2012. *R graphics cookbook: Practical recipes for visualizing data*. O'Reilly Media.

[2] Cleveland, W.S. 1994. *The elements of graphing data*. Hobart Press.

[3] Histogram - wikipedia: [vid. 6.8.2017]. *https://en.wikipedia.org/wiki/Histogram*.

[4] Maciejewski, R. 2011. *Data representations, transformations, and statistics for visual reasoning*. Morgan & Claypool Publishers.

[5] Novovičová, J. 2006. *Pravděpodobnost a matematická statistika*. Praha: Vydavatelství ČVUT,

[6] P–P plot - wikipedia: [vid. 11.8.2017]. *https://en.wikipedia.org/wiki/P%E2%80%93P_plot*.

[7] R: Generic x-y plotting: [vid. 11.5.2017]. *https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/plot.html*.

[8] R: The r graphics package: [vid. 22.4.2017]. *https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/graphics-package.html*.

[9] Teetor, P. 2011. *R cookbook: Proven recipes for data analysis, statistics, and graphics*. O'Reilly Media.

[10] Wilkinson, L., Wills, D., Rope, D., Norton, A. and Dubbs, R. 2006. *The grammar of graphics*. Springer New York.