ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

FAKULTA ŽIVOTNÍHO PROSTŘEDÍ

BAKALÁŘSKÁ PRÁCE

2018 Irina Georgievová

ČESKÁ ZEMĚDĚLSKÁ UNIVERZITA V PRAZE

FAKULTA ŽIVOTNÍHO PROSTŘEDÍ

KATEDRA VODNÍHO HOSPODÁŘSTVÍ
A ENVIRONMENTÁLNÍHO MODELOVÁNÍ

Vizualizace enviromentálních dat

BAKALÁŘSKÁ PRÁCE

Vedoucí práce: **doc. Ing. Martin Hanel, Ph.D.**

Bakalant: **Irina Georgievová**

2018

Česká zemědělská univerzita v Praze

Fakulta životního prostředí

ZADÁNÍ BAKALÁŘSKÉ PRÁCE

Autorka práce:	Irina Georgievová
Studijní program:	Krajinářství
Obor:	Vodní hospodářství
Vedoucí práce:	doc. Ing. Martin Hanel, Ph.D.
Garantující pracoviště:	Katedra vodního hospodářství a environmentálního modelování
Jazyk práce:	Čeština
Název práce:	**Vizualizace environmentálních dat**
Název anglicky:	**Visualization of environmental data**
Cíle práce:	Představení klíčových poznatků týkajících se vizualizace a průzkumové analýzy dat z teoretického hlediska i z hlediska praktické implementace v R. Zhodnoceny budou jak nástroje obsažené v základní distribuci R, tak nástroje dostupné v balících lattice, grid, ggplot2, raster, rasterVis, případně i nástroje pro tvorbu dynamických vizualizací (htmlwidgets, shiny apod.).
Metodika:	- rešerše základních poznatků - popis vizualizačních prostředků se zaměřením na využití v hydrologii, porovnání výhod/nevýhod - popis nejpoužívanějších R balíků, jejich základních funkcí a demonstrace jejich využití
Doporučený rozsah práce:	40-60 stran
Klíčová slova:	vizualizace dat, grammar of graphics, průzkumová analýza dat

Doporučené zdroje informací:

1. WICKHAM, H. *Ggplot2 : elegant graphics for data analysis*. Dordrecht: Springer, 2009. ISBN 978-0-387-98140-6.

Předběžný termín obhajoby: 2017/18 LS - FŽP

Elektronicky zamítnuto: 25. 4. 2017
doc. Ing. Martin Hanel, Ph.D.
Vedoucí katedry

Prohlášení:
Prohlašuji, že jsem bakalářskou práci *Vizualizace enviromentálních dat* zpracovala samostatně. Veškerou literaturu a další podkladové materiály uvádím v seznamu na straně

V Praze dne .....

.....
Irina Georgievová

Poděkování:

Obsah

Úvod

Teoretická část

1 Grammar of graphics

1.1 Historie vizualizace dat (stručná)

1.1.1 Minard & Playfair

1.2 Zásady vizualizace dat

1.2.1 Tuft

(*The Visual Display of Quantitative Information*; *Tufte's principles*)

1.2.2 Cleveland

(*Elements of Graphing Data*; *Visualizing Data*)

2 Základní grafy v R

Pro vytváření základních grafů v R používáme vestavěný balíček `graphics` [18], 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 [20].

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 [20]:

```
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.

2.1 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)
```

2.2 Liniový graf

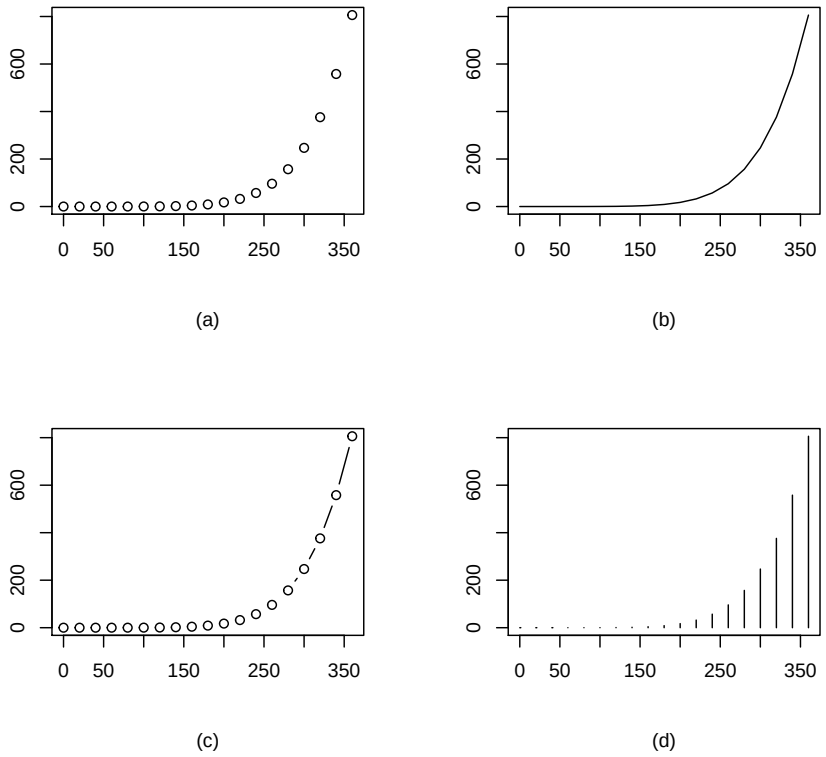
Jediný rozdíl mezi bodovým a liniovým grafem je, že jeden zobrazuje body a druhý je spojuje.[20] (viz. obrázek 1 (a), (b)). Pro vykreslení liniového grafu se používá již několikrát zmíněná funkce `plot()`, kterou doplníme o požadovaný typ vykreslení:

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

V tabulce 1 jsou uvedené některé základní atributy parametru `type`, které mohou být použity [17]:

	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()`.

2.3 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. [20] Tyto názvy slouží

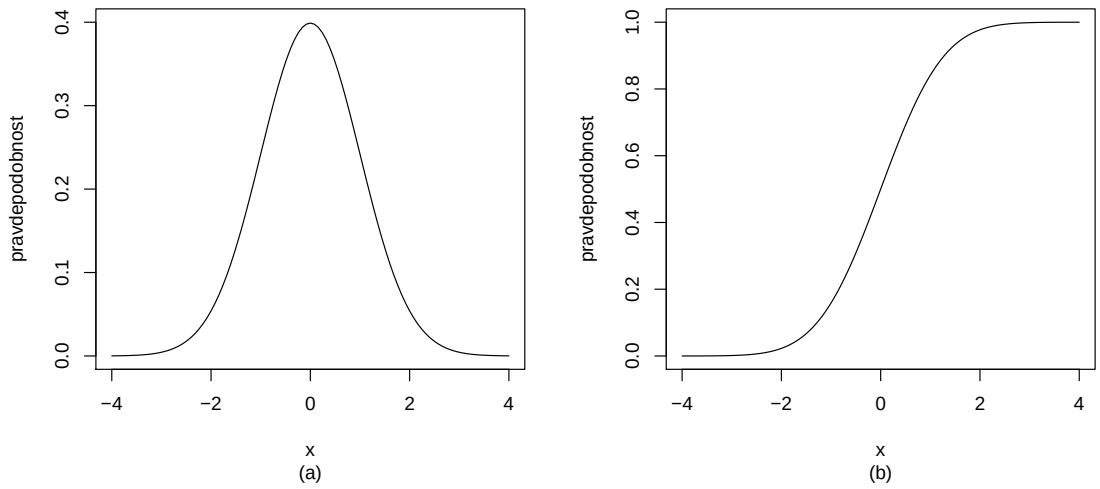
k identifikaci funkcí spojených s rozděleními. Například zkrácený název "norm" pro normální rozdělení, "exp" pro exponenciální rozdělení a další. Funkce pak mají formu:

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)

2.3.1 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 3.4. 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°. Vztah hustoty rozdělení a Q-Q grafu je znázorněn na obrázku 12. [20] [4]

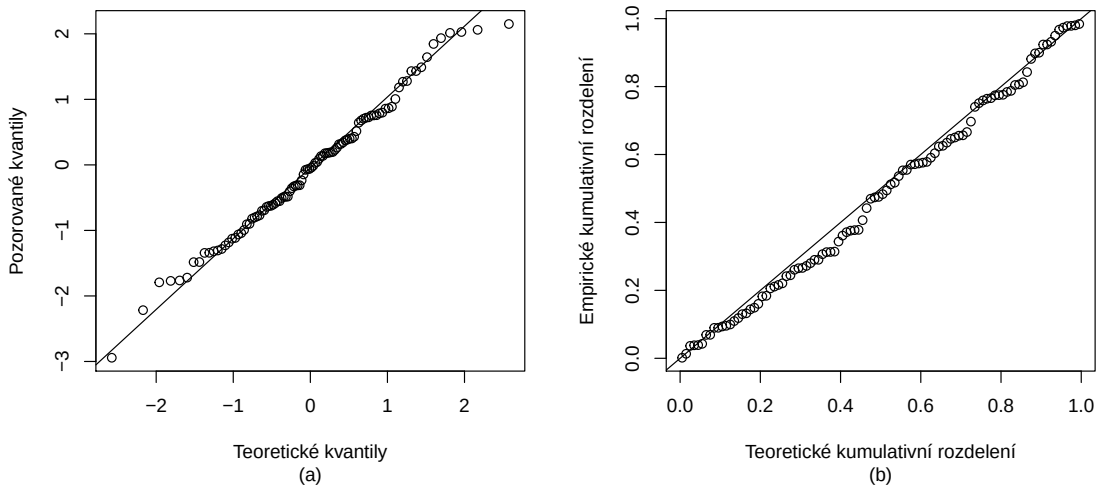
Princip P-P grafu je obdobný jako u Q-Q grafu: vykreslují se dvě distribuční funkcí proti sobě (jedná teoretická a jedná pozorovaná) a pokud všechny body grafu leží přibližně na přímce, jedná se o normální rozdělení. Z velké části se P-P graf používá k vyhodnocení koeficientu šikmosti rozdělení.[16]

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)
```

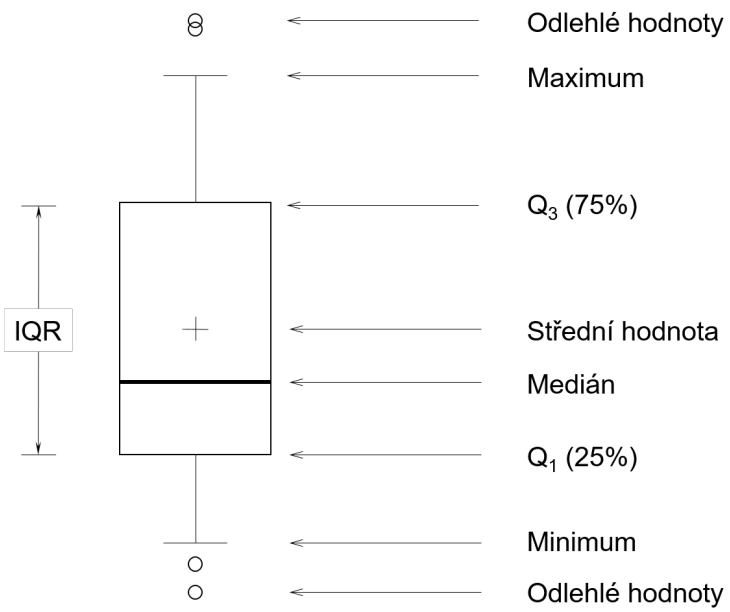


Obrázek 3: Q-Q Graf (a) a P-P Graf (b)

2.3.2 Krabicový graf

Krabicový graf poskytuje rychlé a jednoduché vizuální shrnutí datasetu. V základním prostředí R se vykreslí pomocí funkce `boxplot()` z balíčku `graphics`. Obrázek 4 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 4: Boxplot

2.4 Sloupcový graf

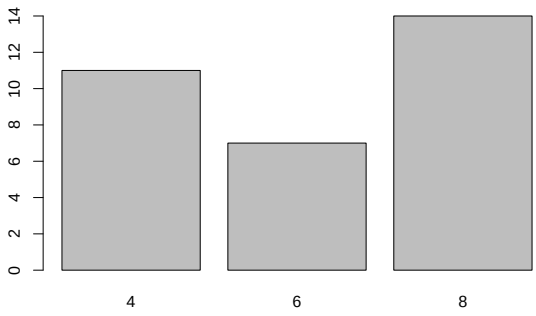
Sloupcový graf je jedním z nejvíce používaných způsobů vizualizace dat. Obvykle se používá pro zobrazení kvantitativních hodnot na ose y a kvalitativních na ose x. Výška sloupců může reprezentovat jak četnosti výskytu hodnot, tak i samotné hodnoty.[3]

V R lze tento typ grafu vykreslit pomocí funkce `barplot()`. V příkladu (Obrázek 5) 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 5: Ukázka jednoduchého sloupcového grafu

2.4.1 Histogram

Sloupcový graf s četnostmi na souvislé ose je taky známý jako histogram.[3] 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.[13] 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 [10]

```
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 [10]

```
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 [8]

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

Freedman–Diaconisovo pravidlo pro stanovení předpokládané šířky intervalu je:

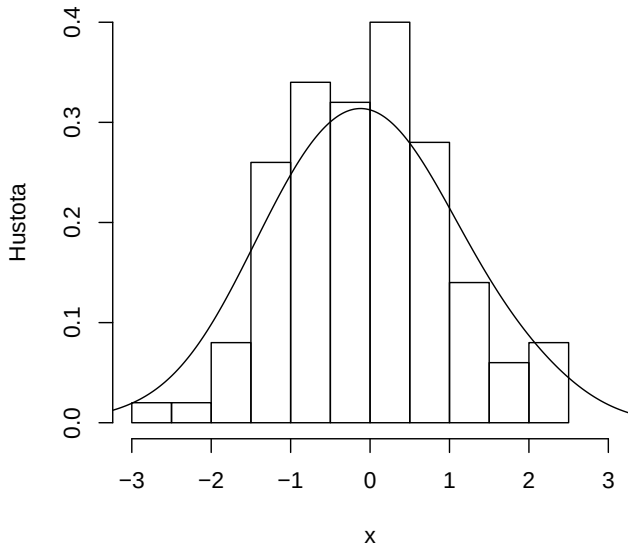
$$h = 2\frac{IQR(x)}{n^{\frac{1}{3}}}$$

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.

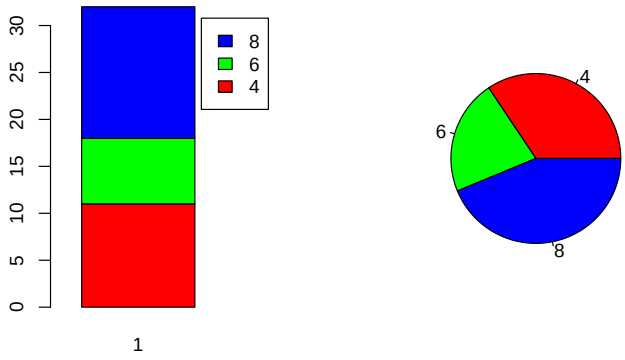
Histogram je jedním ze standardních způsobů, používaných k odhadu tvaru rozdělení, přesto se ale tento způsob považuje za nepřesný, vzhledem k ovlivnění tvaru počtem použitých intervalů. Při normálním rozdělení by měl histogram mít zvoncovitý tvar schodný s Gaussovou křivkou (Obrázek 6).



Obrázek 6: Histogram s odhadem hustoty pravděpodobnosti

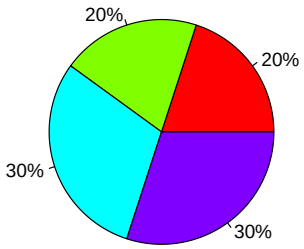
2.4.2 Koláčový graf

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 7). [25]



Obrázek 7: Skládaný sloupcový graf transformovaný do polárního souřadnicového systému

Jednoduché koláčové grafy se vykreslují pomoci funkci `pie()` (Obrázek 8).



Obrázek 8: Ukázka jednoduchého koláčového grafu

2.4.3 Číslicový histogram (*stem-and-leaf*)

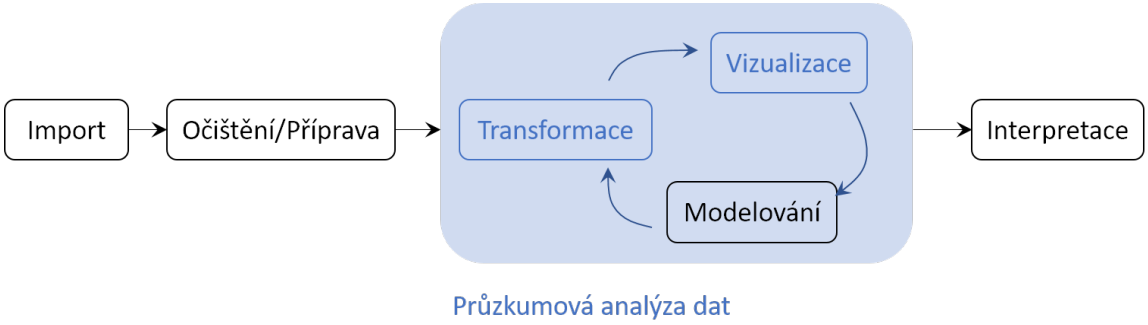
Číslicový histogram je jinak známy jako *stem-and-leaf plot* a podobně histogramu pomáhá vizualizovat tvar rozdělení. Jedná se spíš o historický typ grafu, populární v osmdesátých letech, kdy vykreslování velkých datasetu bylo obtížnější. Vstupní údaje se rozdělí na dva sloupce rozdělené vertikální linii. Pravý sloupec obsahuje listy (*leaf*) - poslední číslici čísla či číslo po desetinné čárce a levý sloupec obsahuje stonek (*stem*) - zbylé číslice (čísla před desetinnou čárkou). Každý stonek se uvádí pouze jednou, aniž by se nějaký vynechal i kdyby to znamenalo, že nebude mít žádné listy. Listy se uvádějí v rostoucím pořadí. Tak v příkladu uvedeným níže, v prvním řádku stonkem je číslice -2, listy jsou číslice 9 a 2 a tak víme že v datasetu se vyskytli čísla -2.9 a -2.2. Tento typ grafu v prostředí R se vykresluje pomoci funkci `stem()`:

```
stem(x)
```

```
##
##   The decimal point is at the |
##
##   -2 | 92
##   -1 | 887755333332211100
##   -0 | 998888776666666555554433332111100
##    0 | 00111222223333444445677778888899
##    1 | 0233445689
##    2 | 0012
```

Fisher & Tukey [23]

3 Průzkumová analýza dat



Obrázek 9: Posloupnost datové analýzy

Úkolem průzkumové analýzy dat (*Explanatory Data Analysis*, zkráceně EDA) je vizualizace a transformace dat systematickým způsobem za účelem maximálního pochopení dat, určení vztahu mezi nimi a posouzení jejích kvality. EDA je důležitou části datové analýzy a měla by být jedním z jejích prvních kroků.

Zařazení průzkumové analýzy dat do procesu datové analýzy je zobrazeno v diagramu 9. Prvním krokem datové analýzy je **import** dat. Obecně v tomto případě to znamená nahrání obdržených dat ze souboru či databáze do prostředí R. Bez tohoto kroku datová analýza nemůže být vykonána. V momentě když data jsou importována do R je dobře je **očistit** neboli **přípravit**. Tím je myšleno ukládání dat v konzistentní a systematické formě, odpovídající sémantice původního datasetu. Zkratka očištěné data jsou taková data, ve kterých sloupce odpovídají proměnným a řádky odpovídají pozorováním. Taková příprava dat usnadňuje další práci s nimi.

Jakmile data jsou očištěná je obvyklým krokem jejich transformace. **Transformaci** se rozumí omezení pozorování (například dle zájmového území, povodí), vytváření nových proměnných, na základě již existujících, agregace (např. z denního do měsíčního kroku) a výpočet souhrnných statistik (střední hodnoty, kvantily atd.), odstranění odlehlých pozorování, normalizace. Po tom co jsou data očištěná a obsahují veškeré potřebné proměnné je možné na ně aplikovat dva nejdůležitější nástroje k zjištění informaci: vizualizaci a modelování. Tyto nástroje mají svoje výhody a nevýhody a jakákoliv skutečná analýza se na ně opakovaně obrací.

Vizualizace je schopná odhalit neočekávané chování dat a poukázat na další směr analýzy. Vizualizaci lze odhalit nevhodně zvolená či špatně připravená data a nekorektní dotazování. I přesto že vizualizace je dobrým nástrojem datové analýzy, její aplikace na větší datasety je značně náročná a interpretace výsledku je na analytikovi.

Modelování je doplňek vizualizace. Jedná se o zásadně matematický a výpočetní nástroj, který se obecně hodí i na větší datasety. Téměř každý model musí splňovat své předpoklady, které by měli být ověřené před jejich aplikaci, na rozdíl od vizualizace, která žádnými předpoklady nedisponuje.[24]

Důležitou součásti analýzy je **interpretace** výsledků a formulace závěrů. Vyhodnocuje se jak dobře zvolený model či vizualizace slouží k pochopení dat a jejich popisu. Je také důležitý si uvědomit komu se výsledky interpretují, kdo je cílová skupina. Dobře provedené grafické výstupy podložené jejich správnou interpretaci jsou jedním z nejlepších způsobů prezentaci dát.

Průzkumová analýza dat není specifikována jako konkretní soubor pravidel a postupu, ale jako přístup k analýze dat. Obvykle zahrnuje následující kroky:

- vyhledávání vybočujících (odlehlých) pozorování,
- náhrada chybějících hodnot,
- transformace dat,
- změny typu proměnných,
- ověřování normality

3.1 Odlehlá pozorování

Odlehlá pozorování (*outliers*) jsou významně odlišná vůči ostatním hodnotám datasetu. Definice toho, jak moc odlišná taková pozorování mají být je dáno analytikem na základě konkretního datasetu a kontextu problematiky. Tato pozorování mohou být indikátorem chybných dat nebo vzácných událostí. Důvody proč se tato pozorování vyskytují by měli by být pečlivě zkoumány. Dále je důležité posoudit jak je jimi ovlivněn výsledek analýzy. Případně zdali je předpoklady metody připouštějí.

Hledání odlehlých, vybočujících pozorování a jiných anomálií pro jednotlivé veličiny se dá udělat graficky například pomoci boxplotu (viz. sekce 2.3.2), bodových grafů (2.1) nebo číslicových histogramů (2.4.3). Dají se také vypočítat pomocí různých statistik, například metoda *jackknife*, která bude popsána v následující kapitole (3.1.1). V momentech, kdy je vizualizace obtížná (velké datasety, větší množství navzájem se ovlivňujících proměnných, atd.), využívají se nástroje vícerozměrné, například Mahalanobisovy vzdálenosti (3.1.2), *leverages* (3.1.3) a další.

3.1.1 *Jackknife*

Metoda byla původně představená Johnem W. Tukey v roce 1958 v "*The Annals of Mathematical Statistic*" [22] a jedná se o speciální případ metody *bootstrap* (více o metodě B. Efron a R. Tibshirani v "*An Introduction to the Bootstrap*" [6]).

Postup metody *jackknife* je založen na celkem jednoduché myšlence. Zjišťují se souhrnné statistiky podsouborů (*Jackknife Samples*), které se vytvářejí postupným vypouštěním jednotlivých pozorování z původního datasetu. Jinými slovy existuje n unikátních Jackknife podsouborů a i-tý Jackknife podsoubor je definován jako vektor.

Pomocí porovnání souhrnných statistik původního datasetu a vytvořených Jackknife podsouborů se odhadne vliv jednotlivých pozorování na původní dataset. Jedná ze souhrnných statistik, kterou lze použit je střední hodnota $\bar{x}$. Tak, pro původní

dataset obsahující n pozorování střední hodnota se stanoví dle vzorce $\bar{x} = \frac{1}{n}\sum_{i=1}^{n}\bar{x}_i$. Střední hodnota Jackknife podsouborů se vyhodnotí následovně:

$$\bar{x}_i = \frac{1}{n-1}\sum_{j=1,j\neq i}^{n} x_j, \quad \text{kde } i = 1,\ldots,n.$$

Porovnání se pak provede dle vzorce $Var(\bar{x}) = \frac{n-1}{n}\sum_{i=1}^{n}(\bar{x}_i - \bar{x})^2$, kde $Var(\bar{x})$ je odhad rozptylu, který indikuje, jak moc jednotlivá pozorování ovlivňují dataset, tzn. přítomnost odlehlých pozorování. Metoda může být také použitá k odhadu skutečné, neovlivněné střední hodnoty datasetu. [12]

3.1.2 Mahalanobisovy vzdálenosti

K měření vzdálenosti mezi objekty se často používá euklidovská vzdálenost. Euklidovská vzdálenost je jednoduchá na výpočet a interpretaci, ale není schopná brát v úvahu vztahy mezi daty. Za tímto účelem lze použit mahalanobisovou vzdálenost. Je definovaná matice $\boldsymbol{X}(n \times p)$, obsahující n objektů $\boldsymbol{x}_i$ a p proměnných. Euklidovská vzdálenost mezi vektorem i-tého řádku $\boldsymbol{x}_i(1 \times p)$ této matici a vektoru středních hodnot řádku $\bar{\boldsymbol{x}}(1 \times p)$ se počítá jako

$$ED_i = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T}, \quad \text{pro } i = 1,\ldots,n$$

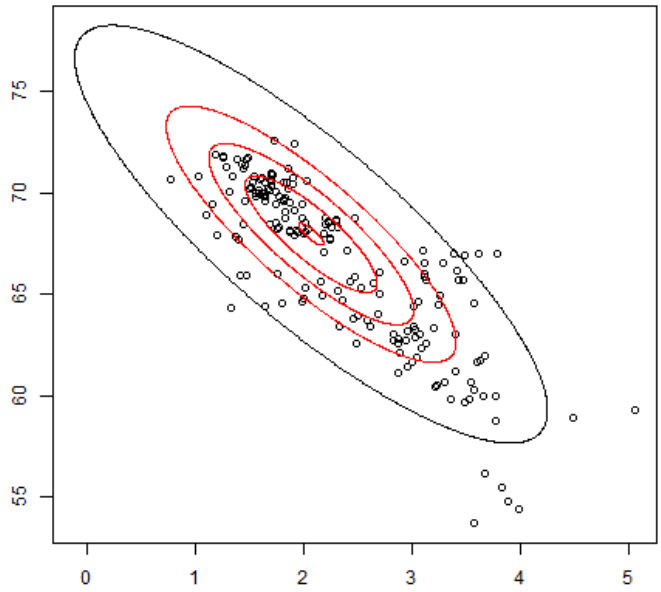
zatímco mahalanobisová vzdálenost se počítá jako

$$MD_i = \sqrt{(\boldsymbol{x}_i - \bar{\boldsymbol{x}})\boldsymbol{C}_x^{-1}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T}, \quad \text{pro } i = 1,\ldots,n$$

kde $\boldsymbol{C}_x$ je kovarianční matice. [5]

Na obrázku 10 jsou znázorněny elipsy mahalanobisových vzdálenosti, kde každá elipsa představuje vzdálenost od průměru. Z tohoto je zřejmý že vzdálenost roste pomaleji ve směru korelace. Pozorování které je výrazně vzdáleno od středu ale leží ve směru závislosti má nižší mahalanobisovou vzdálenost než pozorování které je stejně vzdáleno od středu ale neleží ve směru závislosti. Tato vlastnost mahalanobisových vzdálenosti umožňuje identifikaci odlehlých pozorování.

Metoda byla představená P.C. Mahalanobisem v roce 1936 ve článku *"On the Generalized Distance in Statistics"* [11]. Mahalanobisové vzdálenosti se používají nejenom k nalezení odlehlých pozorování, ale i ke zkoumání reprezentativity mezi dvěma data sety, aplikuje se v algoritmu k-nejbližších sousedů, v diskriminační analýze a má mnoho dalších uplatnění.



Obrázek 10: Mahalanobisovy vzdálenosti

3.1.3 Leverages

Leverage (případně též efekt, vliv nebo projekční h prvek) se používá v regresní analýze k měření velikosti vlivu pozorování na regresní odhad. Princip metody spočívá v kontrole diagonálních prvků projekční matici $\boldsymbol{H}$, která je produktem metody nejmenších čtverců a je definována $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Model lineární regrese může být zapsán: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon}$, kde vektor vysvětlované proměnné je $\boldsymbol{y}$, matice vysvětlujících proměnných je $\boldsymbol{X}$, vektor regresních koeficientů, který se odhaduje, je $\boldsymbol{\beta}$ a vektor náhodné složky je $\boldsymbol{\varepsilon}$. Metoda nejmenších čtverců poskytuje řešení regresní rovnici: $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Lze dosadit: $\hat{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Výsledný vektor pak má tvar $\hat{\boldsymbol{y}} = \boldsymbol{H}\boldsymbol{y}$, kde $\boldsymbol{H}$ je projekční matice. [2]

3.2 Náhrada chybějících pozorování

Problem chybějících pozorování spočívá v neschopnosti některých metod k jejich zpracovávání. Takové hodnoty lze vynechat nebo doplnit (nahradit) jednou z řady metod. Vynechání hodnot vede k nežádoucímu zmenšení datasetu, proto je výhodnější chybějící údaje doplnit. Nejednodušším nástrojem pro náhradu chybějících hodnot je aritmetický průměr příslušné proměnné. Tento způsob může vést ke zkresleným odhadům (neplatí-li předpoklad, že chybějící údaje jsou zcela náhodné) a podhodnocuje variabilitu a kovariancí datasetu a proto se nedoporučuje v případě vyššího podílu chybějících údajů. Další možnou metodou je náhrada náhodným číslem generovaným

z příslušného rozdělení (parametry se odhadují z výběru). V tomto případě se respektuje variabilita datasetu, ale nerespektuje se jeho kovariance. Chybějící údaje lze také odvozovat pomocí známých hodnot na základě jednoduché lineární regresní funkce. Tato metoda respektuje nejenom variabilitu vzorku, ale i jeho korelační strukturu. [15]

3.3 Transformace dat

Jedním z cíle transformací dat je dosažení srovnatelnosti proměnných: sjednocení měřítka, variaci a typu proměnných. Hlavním využitím je splnění podmínek vyžadovaných metodami, například podmínky normality, kde se snažíme přivést data na normální rozdělení pro snížení vlivu rušivých proměnných, odlehlých hodnot, snížení vztahu mezi střední hodnotou a rozptylem atd. [21] Rozdělujeme lineární transformaci (centrování, normování) a nelineární transformaci (plynoucí z typu a charakteru dat).

Lineární transformace zachovává lineární vztahy mezi proměnnými. Jedním z příkladu takové úpravy dat je metoda centrování, která se používá u vícerozměrných analýz. Podstata metody spočívá v zachování měřítka vzorku při změně hodnot: od původních hodnot se odečítá průměr proměnné (od prvků sloupce se odečte jejich sloupcový průměr), průměry získaných nových proměnných se rovnají nule. Matematický zápis by mohl být zapsán následovně:

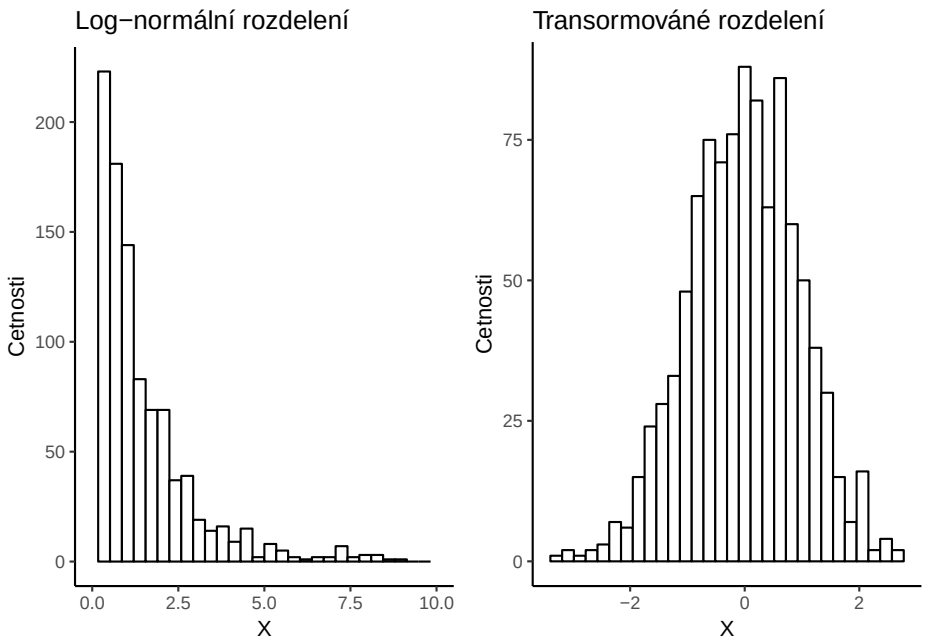
$$v_{ij} = x_{ij} - \bar{x}_j$$

Vektor průměrů $\bar{\boldsymbol{v}}$ je nulový, kovariance a korelace proměnných zůstává nezměněná. [7] Další často využívanou metodou je metoda normalizace dat. Tato metoda transformuje měřítka vzorků pro možnost jejich porovnání, "eliminuje" jednotky měření, po úpravě střední hodnota vzorku odpovídá nule a odchylka jedničce (normální rozdělení).

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma(x_j)}$$

$\sigma(x_j)$ je směrodatná odchylka sloupce proměnné, vektor průměrů $\bar{\boldsymbol{z}}$ je nulový a kovariance vektoru nových proměnných se shoduje s korelaci původního vektoru. [1]

Nelineární transformace plyne z typu dat a mění (snižuje či zvyšuje) lineární vztahy mezi proměnnými a to znamená, že nezachovává korelaci mezi nimi. Pokud data mají charakter absolutní četností, používá se odmocninová transformace $X' = \sqrt{X}$, pokud odpovídají lognormálnímu rozdělení, používá se logaritmická transformace $X' = \log_{10} X$ atd. Logaritmus náhodné veličiny s log-normálním rozdělením po úpravě má normální rozdělení (viz obrázek 11). Logaritmická transformace může být použita pouze u nezáporných rozdělení. [26] [9]

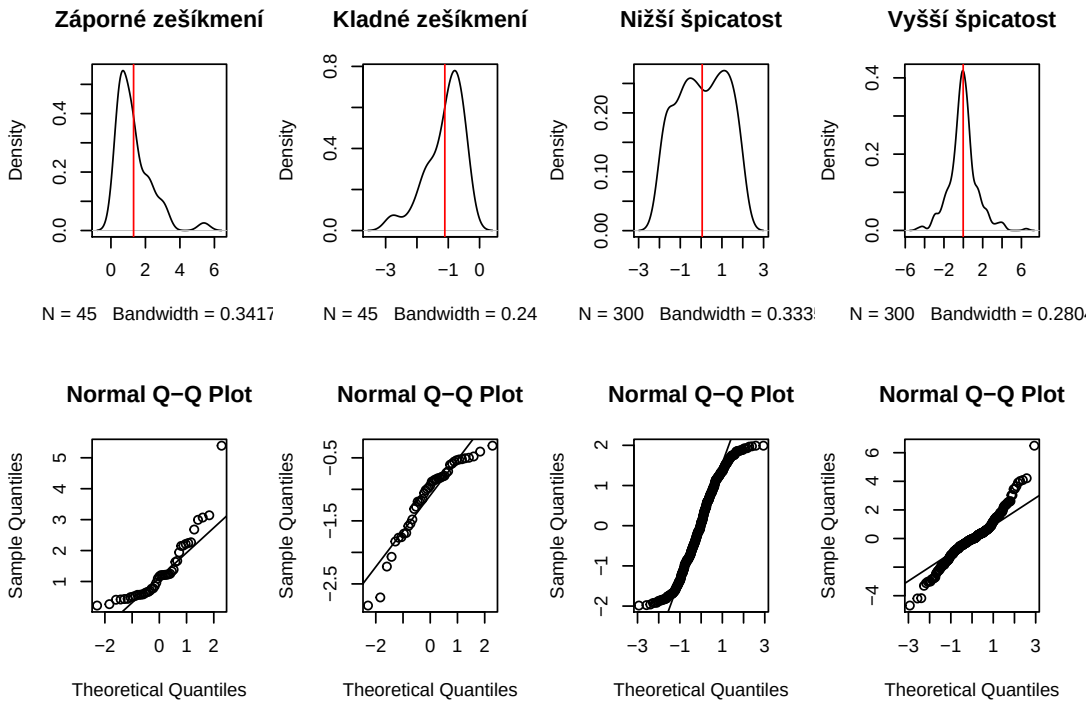


Obrázek 11: Log-normální rozdělení transformováné na normální

3.4 Ověřování normality

Důležitým aspektem popisu proměnné je tvar jejího rozdělení, který udává četnosti hodnot z různých rozsahů proměnné. Většina statistických testů a metod se zakládá na předpokladu, že proměnná má normální rozdělení. Z tohoto důvodu je vhodné ověřovat normalitu rozdělení analyzovaného vzorku. Ověřené statistické testy poskytují přesné výsledky, pokud nejsou porušený předpoklady normálnosti.

Zjistit zda-li vzorek pochází z normálního rozdělení lze grafickým posouzením nebo pomocí testů normality. Mezi nástroje grafického posouzení normality se řádí histogram rozdělení četnosti (kapitola 2.4.1), graf výběrové distribuční funkce (2.3), Q-Q graf a P-P graf (2.3.1). Vztah hustoty rozdělení a Q-Q grafu je znázorněn na obrázku 12. Dále existuje řada testů normality, zde budou popsané Shapiro-Wilk (SW) test a Jarqua-Bera (JB) test.



Obrázek 12: Vztah hustoty rozdělení a Q-Q grafu pro různá narušení normality

Shapiro-Wilk test byl poprvé představen v roce 1965 S.S. Shapiro a M. Wilkem [19]. Metoda dokáže pracovat se vzorky velikosti 12 až 5000 elementů. Nulová hypotéza tohoto testu předpokládá, že vzorek má normální rozdělení. Pokud p-hodnota je menší, než zvolená hladina významnosti, zamítá se nulová hypotéza, jinými slovy vzorek nemá normální rozdělení. Statistika testu vypadá následovně:

$$W = \frac{(\sum_{i=1}^{n} a_i x_{(i)})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

kde $x_{(i)}$ je i-tý nejmenší prvek (statistika i-tého řádu), $\bar{x}$ je průměr vzorku, n je počet pozorování.

Jarqua-Bera test závisí na koeficientech šikmosti a špičatosti. Statistika JB testu může být zapsána:

$$T = n\left(\frac{(\sqrt{b_1})^2}{6} + \frac{(b_2 - 3)^2}{24}\right),$$

kde n je velikost vzorku, $\sqrt{b_1}$ je koeficient šikmosti vzorku a b_2 je koeficient špičatosti. Nulová a alternativní hypotézy se schodují s SW testem. Používá se pro větší datasety (nad 2000 elementů). [14]

Praktická část

4 Praktická vizualizace dat

4.1 Prostředí R

4.1.1 Balíčky

4.1.2 ...

4.2 Balíčky pro vizualizaci dat

4.2.1 ggplot2

4.2.2 lattice

4.2.3 rgl

4.3 Balíčky pro interaktivní vizualizaci dat (htmlwidgets)

4.3.1 plotly

4.3.2 dygraphs

4.3.3 leaflet

4.3.4 ggvis

4.4 Balíčky pro prostorová data

4.4.1 ggmap

4.5 ...

4.5.1 raster

4.5.2 rasterVis

4.6 Balíčky pro webové aplikace

4.6.1 shiny

4.6.2 flexdashboard

4.6.3 dashboard

Literatura

[1] Abdi, H. and Williams, L. 2010. Normalizing data. encyclopedia of research design. *Thousand Oaks, CA: Sage*. (2010).

[2] Cardinali, C. 2014. Observation influence diagnostic of a data assimilation system. *Advanced Data Assimilation for Geosciences: Lecture Notes of the Les Houches School of Physics: Special Issue*. (2014).

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

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

[5] De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D.L. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems*. 50, 1 (2000).

[6] Efron, B. and Tibshirani, R. 1994. *An introduction to the bootstrap*. Taylor & Francis Ltd.

[7] Hebák, P., Hustopecký, J., Jarošová, E. and Pecáková, I. 2007. *Vícerozměrné statistické metody (1)*. Informatorium, Praha.

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

[9] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. 2004. *Applied linear statistical models*. McGraw-Hill/Irwin.

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

[11] Mahalanobis, P.C. 1936. On the generalised distance in statistics. *Proceedings National Institute of Science*. 2, 1 (1936).

[12] McIntosh, A. 2016. The jackknife estimation method. *arXiv*. (2016).

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

[14] Öztuna, D., Elhan, A.H. and Tüccar, E. 2006. Investigation of four different normality tests in terms of type 1 error rate and power under different distributions.

Turkish Journal of Medical Sciences. 36, 3 (2006).

[15] Pecáková, I. 2014. Problém chybějících dat v dotazníkových šetřeních. *Politická ekonomie*. 2014, (Jan. 2014).

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

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

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

[19] SHAPIRO, S.S. and WILK, M.B. 1965. An analysis of variance test for normality (complete samples)†. *Biometrika*. 52, 3-4 (1965). DOI:https://doi.org/10.1093/biomet/52.3-4.591.

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

[21] Transformation of data: [vid. 3.2.2018]. *http://statisticalconcepts.blogspot.cz/2010/02/transformation-of-data-validity-of.html*.

[22] Tukey, J.W. 1958. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*. 29, (1958).

[23] Tukey, J.W. 1977. *Exploratory data analysis*. ADDISON WESLEY PUB CO INC.

[24] Wickham, H. and Grolemund, G. 2017. *R for data science*. O'Reilly Media.

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

[26] Zumel, N. and Mount, J. 2014. *Practical data science with r*. Manning.