

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:	Irína 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 balíčcích lattice, grid, ggplot2, raster, rasterVis, případně i nástroje pro tvorbu dynamických vizualizací (htmlwidgets, shiny apod.).
Metodika:	<ul style="list-style-type: none">- 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:

Abstrakt

Vložte abstrakt o rozsahu cca 100–200 slov. K problému vícenásobné marginalizace (VM) dochází, pokud články dodavatelského řetězce nastavují cenu svého výstupu způsobem, který by optimalizoval zisk v podmínkách prodávajícího na trhu s koncovým zbožím. V takové situaci dochází k cenové spirále a výsledná cena koncové produkce svou přílišnou výši poškozuje jak spotřebitelský užitek, tak i zisk řetězce jako celku. Kvantifikace dopadu VM byla již předmětem našeho dřívějšího výzkumu. Tento příspěvek se zabývá otázkou, za jakých podmínek k cenové spirále dochází a jakým způsobem probíhá konvergence k výsledným (rovnovážným) cenám. Pro tyto účely byl navržen a implementován model řetězce v podobě multiagentního systému, se kterým je možné analyzovat celý proces pomocí počítačové simulace.

Klíčová slova: vizualizace dat, grammar of graphics, průzkumová analýza dat

Abstract

Double (or multiple) marginalization is often identified as the main source of a decentralized supply chain's (SC's) inefficiency. In its core lies the fact that if the agents constituting the SC choose their output prices according to the golden rule of profit maximization (which normally applies to a single firm that produces independently and sells directly to the end consumer), the prices in the SC tend to spiral up to an inefficient level (equilibrium prices) where both the consumer surplus and the SC's total profit are diminished. The level of equilibrium prices and their impact on the SC's profit and efficiency had been studied in our earlier works. Our focus in this paper was the properties of the process of convergence of the prices inside the SC to equilibrium levels. The analysis was carried out using computer experiments with an agent-based simulation model of a SC with limited information. Only serial chain structure was considered.

Keywords: Data visualization, grammar of graphics, exploratory data analysis

Obsah

Úvod	9
Teoretická část	10
1 Vizualizace dat	10
1.1 Historie vizualizace dat	10
1.2 Zásady vizualizace dat	13
1.2.2 Edward Tufte	13
1.2.1 Wiliam S. Cleveland	14
1.3 Grammar of graphics	14
2 Základní grafy v R	15
2.1 Bodový graf	15
2.2 Liniový graf	15
2.3 Vykreslení rozdělení v R	16
2.3.1 Q-Q graf a P-P graf	18
2.3.2 Krabicový graf	19
2.4 Sloupcový graf	20
2.4.1 Histogram	20
2.4.2 Koláčový graf	22
2.4.3 Číslícový histogram (<i>stem-and-leaf</i>)	23
3 Průzkumová analýza dat	24
3.1 Odlehlá pozorování	25
3.1.1 <i>Jackknife</i>	25
3.1.2 Mahalanobisovy vzdálenosti	26
3.1.3 Leverages	27
3.2 Náhrada chybějících pozorování	28
3.3 Transformace dat	28
3.4 Ověřování normality	29
4 Praktická vizualizace dat	31
4.1 Prostředí R	31
4.1.1 Balíčky	31
4.1.2	31
4.2 Balíčky pro vizualizaci dat	31
4.2.1 ggplot2	31
4.2.2 lattice	31
4.2.3 rgl	31
4.3 Balíčky pro interaktivní vizualizaci dat (htmlwidgets)	31
4.3.1 plotly	31
4.3.2 dygraphs	31
4.3.3 leaflet	31
4.3.4 ggviz	31
4.4 Balíčky pro prostorová data	31
4.4.1 ggmap	31
4.5	31

4.5.1 raster	31
4.5.2 rasterVis	31
4.6 Balíčky pro webové aplikace	32
4.6.1 shiny	32
4.6.2 flexdashboard	32
4.6.3 dashboard	32
Praktická část	32
Seznam obrázků	33
Seznam tabulek	33
Literatura	34

Úvod

Vizualizace dat vždy hrála a neustále hraje významnou roli ve vědě. Je to jednoduchý a jeden z nejlepších způsobů pochopení dat. Poskytuje jasnou představu o konfiguraci dat, odhaluje skryte struktury v datech a shrnuje informace. Proces vizualizaci je nedílnou součástí mnoha lékařských analýz a téměř všechny přírodní vědy využívají grafického zobrazení dat k vizualizaci a komunikaci svých výsledků. Dlouhou tradici prezentace dat se vyznačuje i ekonomika. Sbírané a analyzované po dobu mnoha let data se v současné době převádějí do grafické formy. Masivní příliv dat a jejich dostupnost vedli k novým metodám a novým přístupům. Kombinace programovacích dovedností, matematických a statistických znalostí a odborných znalostí týkajících se obsahu přijala název "*Data Science*". Objevily se pozice takzvaných "*information designers*", které vyvíjí vlastní softwary pro vizualizaci dat, zakládají poradenské firmy, pořádají globální workshopy nebo vytvářejí blogy s tisíci registrovanými uživateli. [22] Přes všechny výhody vizualizace, jedná se pouze o nástroj datové analýzy, obecně dostupný každému. Nesprávné či nevhodné použití tohoto nástrojů vede k tomu, že existují grafy, které se považují za moc barevné a rušivý, postrádající smysl až zavádějící. Z tohoto důvodů se obracíme na takzvané zásady vizualizace. (?)

... popsat zásady vizualizace, její zařazení do datové analýzy, moderní způsoby vizualizace (používané balíčky v R, interaktivní grafy). Aplikace.

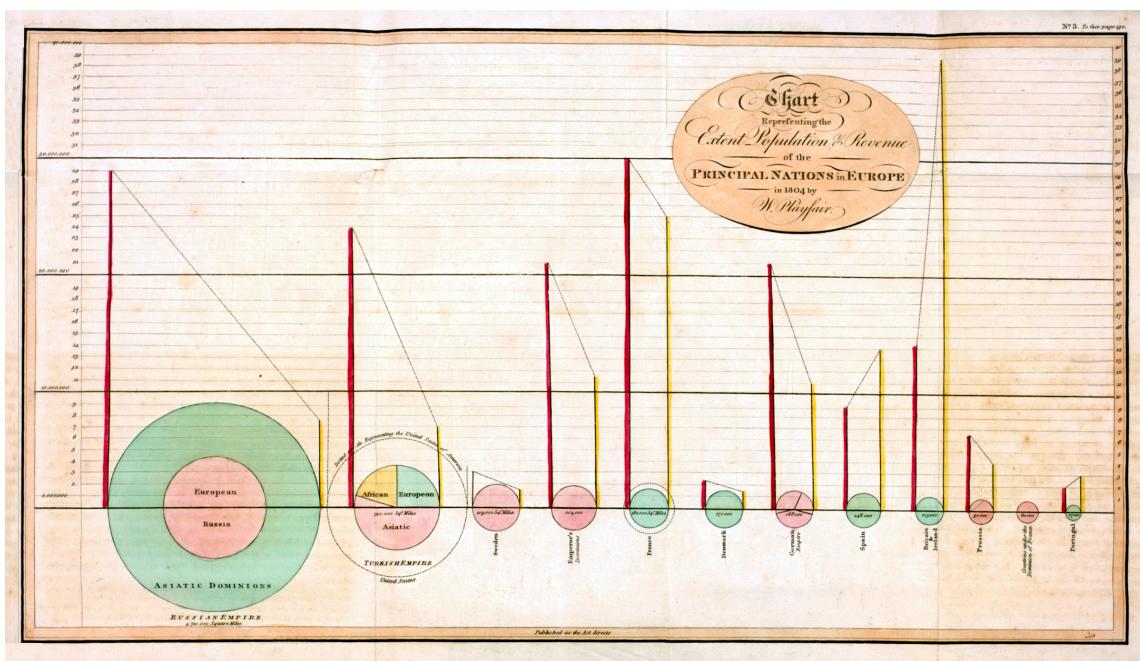
Teoretická část

1 Vizualizace dat

1.1 Historie vizualizace dat

Před 17. stoletím jediné co by se dalo klasifikovat jako vizualizaci dat byly mapy pro navigaci a průzkum, ale také diagramy, geometrická schémata a tabulky pozic hvězd a jiných nebeských těles. Postupný vývoj statistické teorie a růst zájmu o data na konci 18. století vedly k inovacím a expanzi nových grafických forem. Kartografové se pokoušeli zaznamenat více, než pouhou geografickou polohu na mapě a objevili se první pokusy o tematické mapování geologických, ekonomických a medicínských dat.

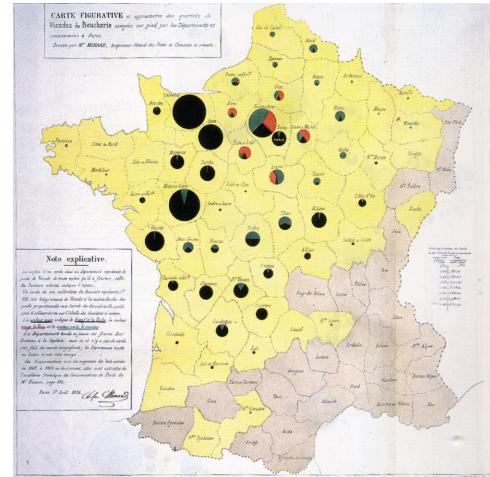
William Playfair (1759-1823) je obecně znám jako průkopník v oblasti vizualizace dat a je považován za vynálezce několika typů grafů. Například liniový a sloupcový grafy a grafy časových řád byly popsány v jeho práci z roku 1786 *"Commercial and Political Atlas"*¹. Později popsal i koláčový graf ve své práci *"Statistical Breviary"* v roce 1801. Obrázek 1 ukazuje příklad jeho kreativní kombinace různých vizualizačních technik (kruhy, koláče, linie), pomocí které se snažil porovnat daňovou zátěž mezi Británií a dalšími zeměmi. Na tomto grafu také ukázal možnost použítí více měřítek pro různé ukazatele (v grafu populace a daně).



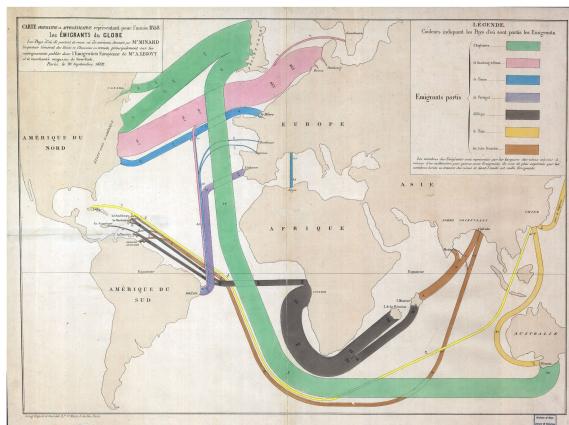
Obrázek 1: Kombinace různých vizuálních technik, Playfair 1801

¹ "Commercial and Political Atlas: Representing, by Copper-Plate Charts, the Progress of the Commerce, Revenues, Expenditure, and Debts of England, during the Whole of the Eighteenth Century"

V polovině 19. století byly vytvořeny všechny podmínky pro rychlý růst vizualizace. V důsledku rostoucí významnosti číselných informací pro sociální plánovaní, industrializaci, obchod a dopravu, byli zřízeny oficiální statistické úřady po celé Evropě. Vývoj statistické teorie, iniciovaný Gaussem a Laplaceem, měl odezvu ve společnosti a poskytl prostředky ke zpracování velkého množství dat. Pro vizualizaci se stalo dat období 1850-1900 "Zlatý věkem", s jedinečnou krásou a velkým množstvím inovací. S těmito inovacemi je hlavně spojené jméno Charlese Josepha Minarda (1781-1870). Například, Minardem bylo zavedeno použití koláčových grafů s výsečemi na mapách (obrázek 2), kde velikost koláčového grafu ukazuje sumu za oblast nebo každý grafický region na mapě a výšeče reprezentují dílčí součty za jednotlivé kategorie. Dále se také zabýval znázorněním geografických pohybu a dopravy lidí, zboží, importu a exportu úměrně jejich velikostí. Tento typ vizualizace se nazývá

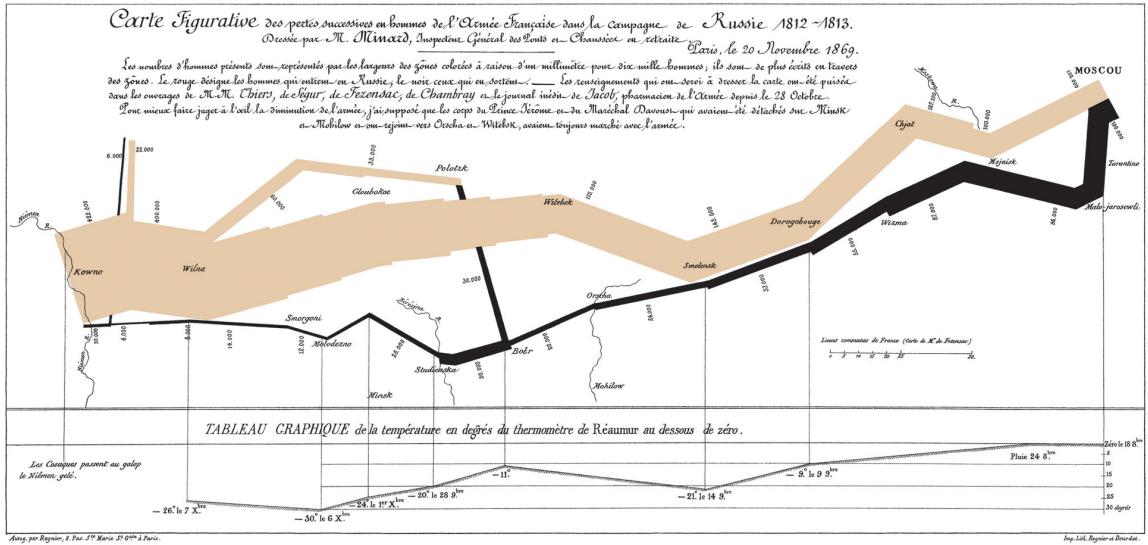


Obrázek 2: Dobytka odeslaný z celé Francie ke spotřebě v Paříži, Minard 1858



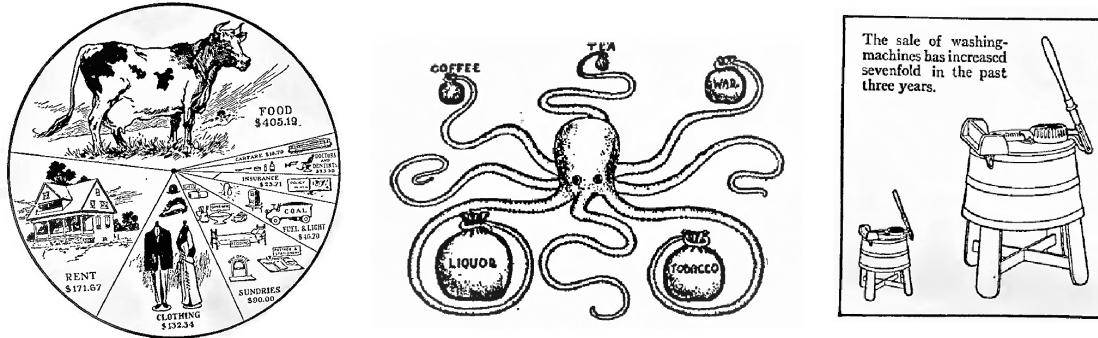
Obrázek 3: Mapa světové migrace, Minard 1858

"flow maps", viz obrázek 3. Jednou z nejslavnějších jeho práci je zobrazení postupných ztrát mužů francouzské armády během Napoleonského tažení na Moskvu v letech 1812-1813 (obrázek 4). Je považovaná za nejlepší informativní vizualizaci. I přestože v tomto grafu je celkem 6 proměnných (množství, lokace ve dvou rozměrech, postup armády, teplota, datum a skupiny), podařilo vše zobrazit tak, aniž by graf byl přeplněný a matoucí.



Obrázek 4: Postup Napoleonských vojsk v letech 1812-13, Minard 1869

Začátek 20. století je občas nazýván "moderním temným věkem" vizualizace. V letech 1900-1950 bylo jen málo grafických inovací. Nadšení pro vizualizaci, které charakterizovalo 19. století bylo nahrazeno formálními (z velké části statistickými) grafy a modely z oblasti sociologie. Hlavní zájem byl o přesná čísla, odhady parametrů, směrodatné odchylinky. Vizualizace byly považované za pouhé hezké obrázky bez schopnosti podat přesná data. [8] Ve své práci *"Graphic Methods for Presenting Facts"* z roku 1919 Willord C. Brinton [1880-1957] kritizoval a vysvětloval chyby takovýchto grafů. Například koláčový graf rozdělení rodinných příjmů (od 900\$ do 1000\$) na obrázku 5. Tento graf je příkladem nepovedené vizualizace: oko preferenčně soudí dle velikostí obrázků a ne dle uhlů výšečí. Obrázek uprostřed znázorňuje druhy utracení: je to zábavný způsob vizualizace, avšak nelze přesně určit velikost brašen, ani je porovnat mezi sebou. Další obrázek by měl čtenáři sdělit informaci, že prodej praček za poslední tří roky vzrostl sedmkrát. Z obrázku není patrný poměr sedmi ku jedné ani přesné roky kdy bylo provedeno porovnání údajů. Dále Brinton ve své práci upozorňoval, že neúspěšná prezentace dat může vést k chybným závěrům a také zmiňoval potřebu jakéhosi standardu, souhrnu "gramatických pravidel pro grafický jazyk". [2]



Obrázek 5: Ukázky vizualizaci ze začatku 20. století, Brinton 1919

Ke "znovuzrození" vizualizace došlo v polovině šedesátých let 20. století, po napsání Johnem W. Tukey [1915-2000] článku "*The Future of Data Analysis*", ve kterém vyzývá společnost k uznání analýzy dat jako samostatného oboru statistiky odlišného od matematické statistiky. [28] Brzy poté začal Tukey s vývojem široké řady nových a efektivních grafů pod společným tématem "průzkumové analýzy dat" (popsaný v jeho práci "*Explanatory Data Analysis*" z roku 1977, viz o tématu kapitola 3). [27] Mezi těmito novými grafy jsou například číslicový histogram (popsaný v kapitole 2.4.3), boxplot nebo krabicový graf (popsaný v kapitole 2.3.2) a další. Mnoho z nich je aktivně používáno ve statistické praxi a implementováno do většiny softwarů. [8]

Od roku 1975 se vyvíjí statistické vypočetní systémy a s nimi i analýza dat a jejich vizualizace. V tomto období se vizualizace začala být vnímána jako vláštní odvětví, hlavně díky Williamu S. Clevelandu a Edwardu Tufte, které založili vědecké základ problematiky. Tufte vyvinul a popularizoval zakladní principy grafické integrity a terminologii. Cleveland se zabýval studiemi grafického vnímání a kognitivních procesů, které lidi používají k pochopení grafů, rozvíjel teorii o spravném provedení vizualizaci. [11] V důsledku jejich práci současná doba se vyznačuje kvalitní, interaktivní a dynamickou vizualizaci. [8]

1.2 Zásady vizualizace dat

1.2.2 Edward Tufte

Za revoluční průlom se považuje kniha Edwarda Tufte *The Visual Display of Quantitative Information* z roku 1983. V kombinaci s dvěma následně publikovanými pracemi *Envisioning Information* z roku 1990 a *Visual Explanations* z roku 1997 jsou to nejznámější práce na téma vizualizace dat a říká se, že Tufte našel autentický způsob k definování jejího "standardu". [22] Ideální způsob vizualizace dle Tufte je ztrouchný, elegantní a informativní graf, jakým je například graf postupu Napoleonských vojsk v letech 1812-13, vytvořený Minardem (viz obrázek 4). Tufte říká, že grafická elegance se často nachází v jednoduchosti návrhu a komplexnosti dat.

1.2.1 Wiliam S. Cleveland

Kromě práci Edwarda Tufte.. Wiliam S. Cleveland ve své práci *The Elements of Graphing Data* z roku 1994

Visualizing Data 1993

1.3 Grammar of graphics

2 Základní grafy v R

Pro vytváření základních grafů v R používáme vestavěný balíček `graphics` [21], 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 [24].

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 irrelevantní parametry vzhledem k zaměření této kapitoly. Základní funkce `plot(x)` jejímž voláním se obdrží pole s grafickou reprezentací proměnné "x", by při doplnění kódu o veškeré parametry vypadala následovně [24]:

```
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 6 (a). Nečíselná data vytvoří jiný typ grafu.

```
plot(cars)
```

2.2 Liniový graf

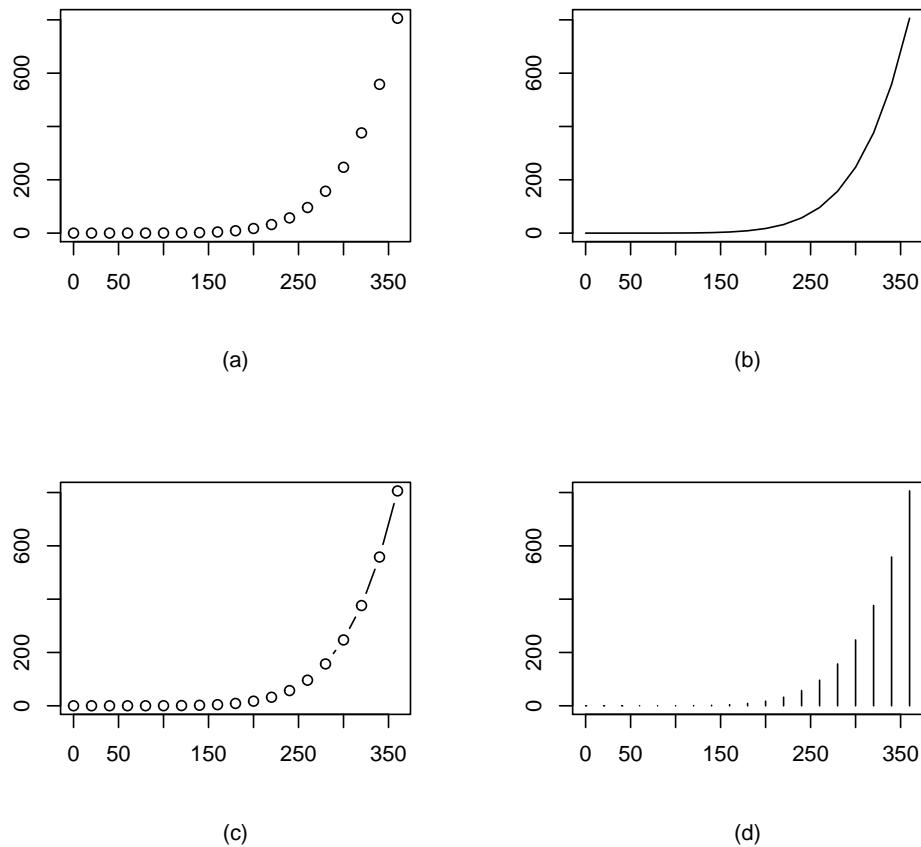
Jediný rozdíl mezi bodovým a liniovým grafem je, že jeden zobrazuje body a druhý je spojuje.[24] (viz. obrázek 6 (a), (b)). Pro vykreslení liniového grafu se používá již několikrát zmíněná funkce `plot()`, kterou doplníme o požadovaný typ vykreslení:

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

V tabulce 1 jsou uvedené některé základní atributy parametru `type`, které mohou být použity [20]:

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

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

2.3 Vykreslení rozdělení v R

Teorie pravděpodobnosti je základem statistiky a R má hodně nástrojů pro práci s pravděpodobností, rozdělením pravděpodobnosti a náhodnými proměnnými. R má zkrácený název pro každé rozdělení pravděpodobnosti. [24] Tyto názvy slouží

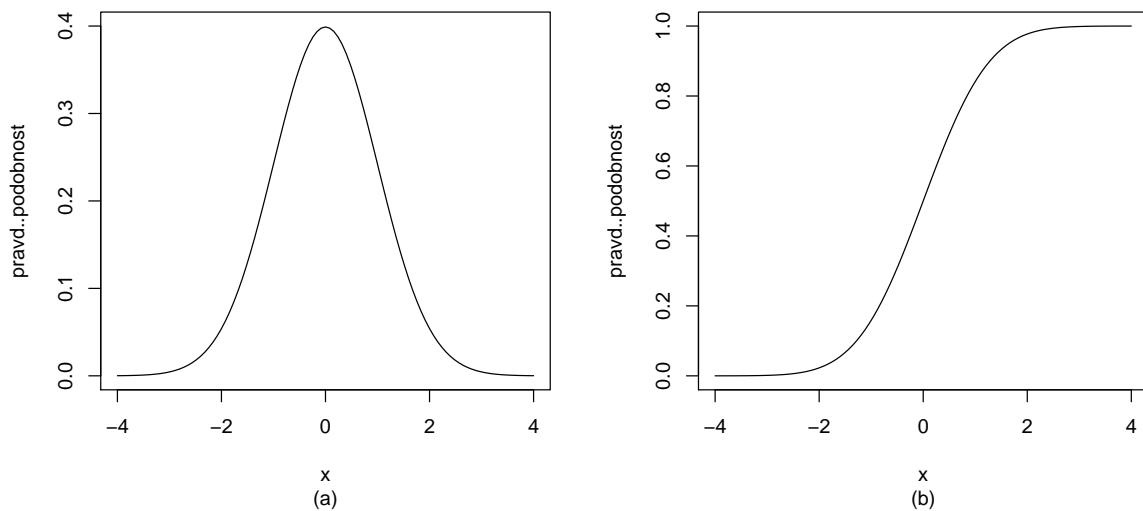
k identifikaci funkcí spojených s rozděleními. Například zkrácený název “norm” pro normální rozdělení, “exp” pro exponenciální rozdělení a další. Funkce pak mají formu:

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 7):

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



Obrázek 7: 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 8) 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 17. [24] [5]

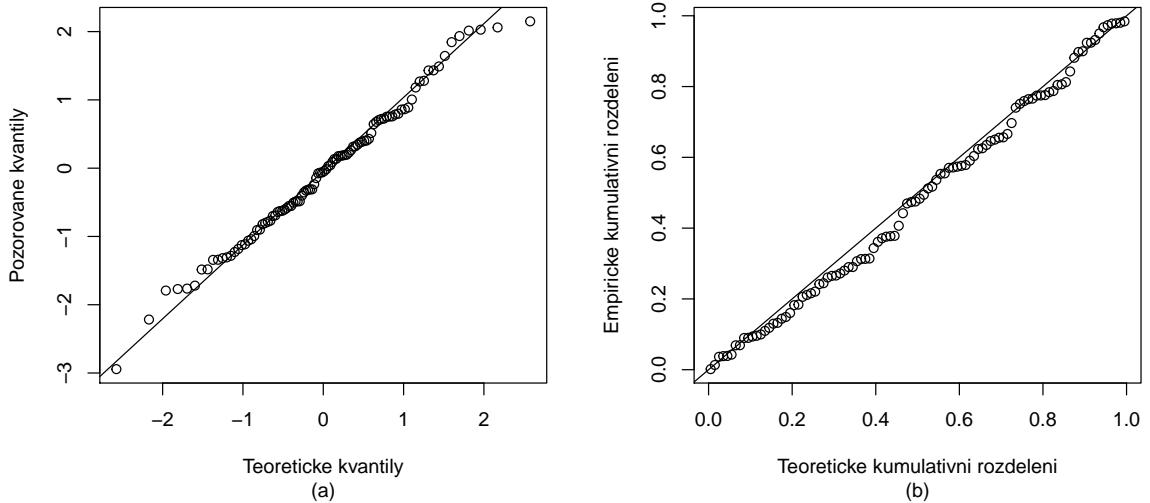
Princip P-P grafu je obdobný jako u Q-Q grafu: vykreslují se dvě distribuční funkcí proti sobě (jedná teoretická a jedná pozorovaná) a pokud všechny body grafu leží přibližně na přímce, jedná se o normální rozdělení. Z velké části se P-P graf používá k vyhodnocení koeficientu šikmosti rozdělení.[19]

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)
```

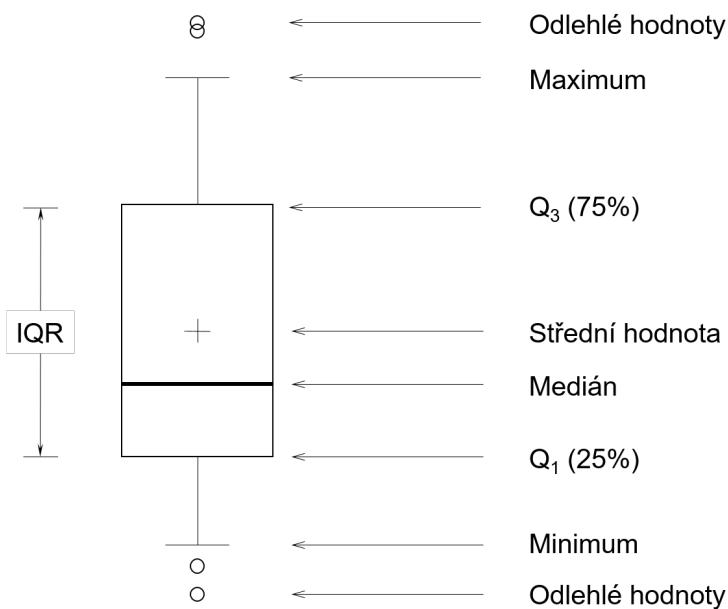


Obrázek 8: Q-Q Graf (a) a P-P Graf (b)

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 9 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 definované jako hodnoty ležící ve větší vzdálenosti od krabice než $1,5 \times \text{IQR}$, kde IQR je mezikvartilové rozpětí (*interquartile range*) neboli $Q_3 - Q_1$.

```
boxplot(x)
```



Obrázek 9: Boxplot

2.4 Sloupcový graf

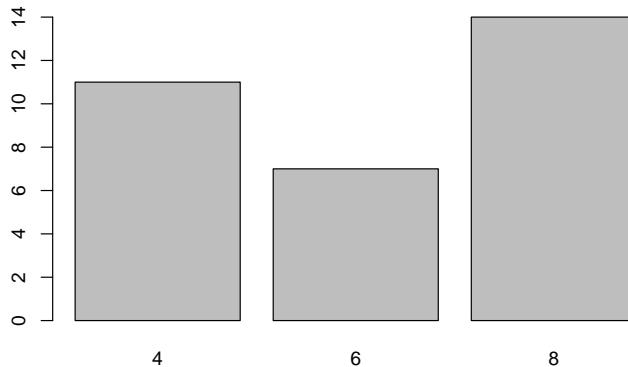
Sloupcový graf je jedním z nejvíce používaných způsobů vizualizace dat. Obvykle se používá pro zobrazení kvantitativních hodnot na ose y a kvalitativních na ose x. Výška sloupců může reprezentovat jak četnosti výskytu hodnot, tak i samotné hodnoty.[4]

V R lze tento typ grafu vykreslit pomocí funkce `barplot()`. V příkladu (Obrázek 10) je použit data set `mtcars`, konkrétně atribut `cyl` - počet válců v motoru.

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

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

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



Obrázek 10: Ukázka jednoduchého sloupcového grafu

2.4.1 Histogram

Sloupcový graf s četnostmi na souvislé ose je taky známý jako histogram.[4] 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.[16] 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 [13]

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

$$k = \lceil \log_2(n) \rceil + 1$$

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

2. Scott [13]

```
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:

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

Případně oba vztahy lze shrnout do jednoho:

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

3. Freedman–Diaconis [10]

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

Freedman–Diaconisovo pravidlo pro stanovení předpokládané šířky intervalu je:

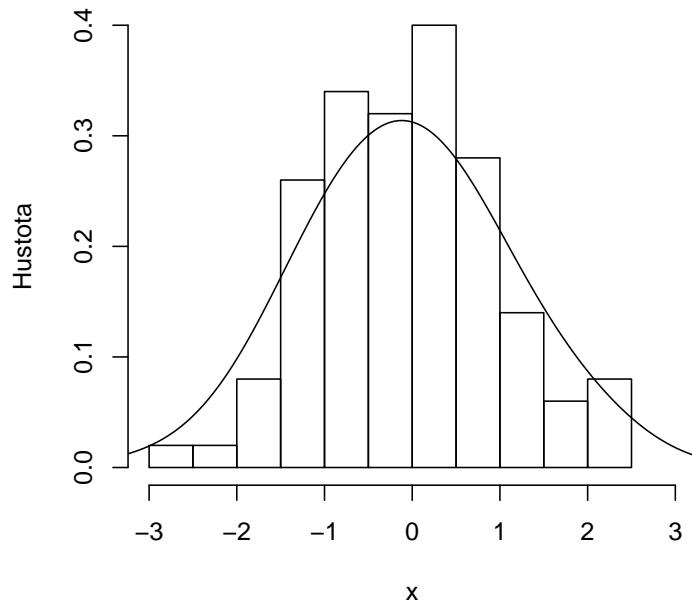
$$h = 2 \frac{IQR(x)}{n^{\frac{1}{3}}}$$

Po dosazení:

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

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

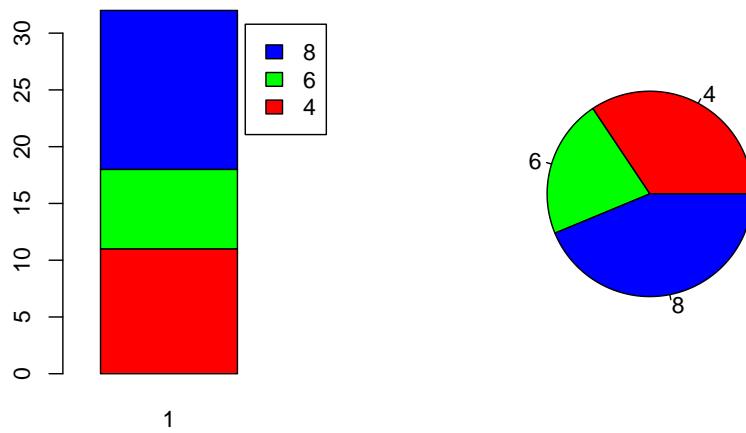
Histogram je jedním ze standardních způsobů, používaných k odhadu tvaru rozdělení, přesto se ale tento způsob považuje za nepřesný, vzhledem k ovlivnění tvaru počtem použitých intervalů. Při normálním rozdělení by měl histogram mít zvoncovitý tvar schodný s Gaussovou křivkou (Obrázek 11).



Obrázek 11: Histogram s odhadem hustoty pravděpodobnosti

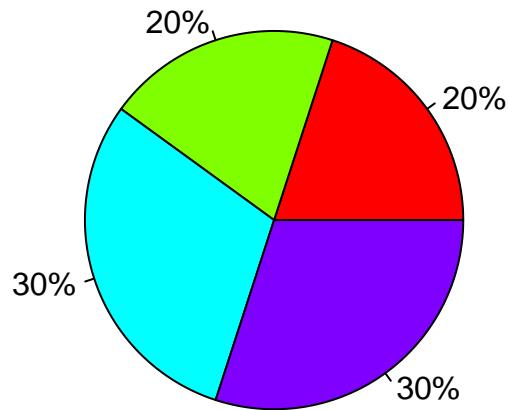
2.4.2 Koláčový graf

Koláčový graf představuje plný kruh (360°), který je rozdělen na jednotlivé výseče pro znázornění číselných proporcí mezi proměnnými. Koláčový graf je tvořen transformací skládaného sloupcového grafu do polárního souřadnicového systému (Obrázek 12). [30]



Obrázek 12: Skládaný sloupcový graf transformovaný do polárního souřadnicového systému

Jednoduché koláčové grafy se vykreslují pomocí funkci `pie()` (Obrázek 13).



Obrázek 13: Ukázka jednoduchého koláčového grafu

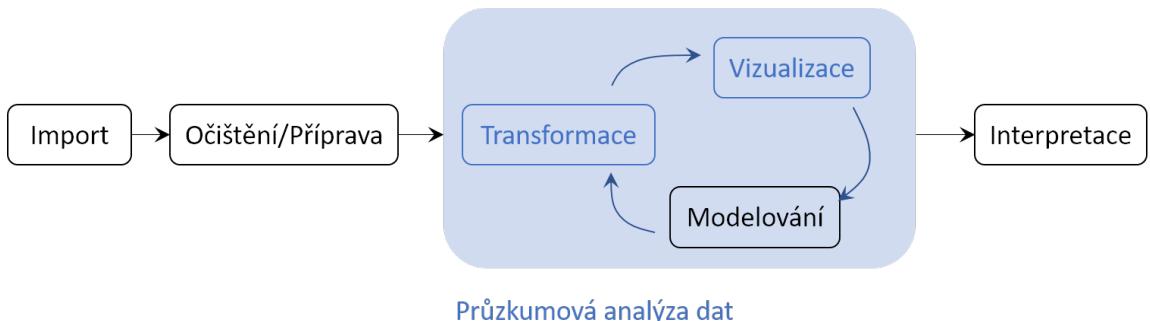
2.4.3 Číslicový histogram (*stem-and-leaf*)

Číslicový histogram, jinak známý jako *stem-and-leaf plot*, podobně jako histogram pomáhá vizualizovat tvar rozdělení. Jedná se spíše o historický typ grafu, který byl populární v osmdesátých letech, kvůli obtížnějšímu vykreslování velkých datasetu. Vstupní údaje jsou rozdelené vertikální linií na dva sloupce. Pravý sloupec obsahuje listy (*leaf*) - poslední číslice po desetinné čárce a levý sloupec obsahuje stonek (*stem*) - číslice před desetinnou čárkou. Každý stonek je uveden pouze jednou i pokud neobsahuje žádné listy. Listy se uvádějí od nejmenšího po největší. [27] Proto v příkladu uvedeném níže je v prvním řádku stonkem číslice -2 a listy jsou číslice 9 a 2. Víme tak, že v datasetu se vyskytli čísla -2.9 a -2.2. Tento typ grafu v prostředí R se vykresluje pomocí funkce `stem()`:

```
stem(x)
```

```
##  
## The decimal point is at the |  
##  
## -2 | 92  
## -1 | 888755333332211100  
## -0 | 998888776666665555443332111100  
## 0 | 001112222233334444456777788888999  
## 1 | 0233445689  
## 2 | 0012
```

3 Průzkumová analýza dat



Obrázek 14: 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 jejich kvality. EDA je důležitou částí datové analýzy a měla by být jedním z jejích prvních kroků.

Zařazení průzkumové analýzy dat do procesu datové analýzy je zobrazeno v diagramu 14. Prvním krokem datové analýzy je **import** dat. Obecně to v tomto případě 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 vhodné je **očistit** neboli **přípravit**. Tím je myšleno ukládání dat v konzistentní a systematické formě, odpovídající sémantice původního datasetu. Zkrátka očištěná data jsou taková data, ve kterých sloupce odpovídají proměnným a řádky odpovídají pozorováním. Takováto příprava dat usnadňuje další práci s nimi.

Jakmile jsou data očištěna, je obvyklým krokem jejich **transformace**. Transformací se rozumí omezení pozorování (například dle zájmového území či povodí), vytváření nových proměnných na základě již existujících, agregace (např. z denního do měsíčního kroku), výpočet souhrnných statistik (středních hodnot, kvantilů atd.), odstranění odlehlych pozorování a normalizace. Poté, co jsou data očištěna a obsahují veškeré potřebné proměnné, je možné na ně aplikovat dva nejdůležitější nástroje k zjištění informací: vizualizaci a modelování. Jakákoliv analýza tyto nástroje opakovaně využívá, ačkoliv samozřejmě mají své výhody a nevýhody.

Vizualizace je schopná odhalit neočekávané chování dat a napovědět 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 výsledků je subjektivní, tudíž závisí na analytikovi.

Modelování je v ramci průzkumové analýzy dat doplňkem 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 ověřeny před jejich aplikací, na rozdíl od vizualizace, která žádné předpoklady nevyžaduje.[29]

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

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

- vyhledávání vybočujících (odlehlých) pozorování
- nahraď 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 být pečlivě zkoumány. Dále je důležité posoudit, jak je jimi výsledek analýzy ovlivněn, případně zdali je předpoklady metody připouštějí.

Hledání odlehlých, vybočujících, pozorování a jiných anomálií pro jednotlivé veličiny lze provést graficky například pomocí boxplotu (viz. sekce 2.3.2), bodových grafů (2.1) nebo číslicových histogramů (2.4.3). Dají se také vypočítat pomocí různých statistik, například metodou *jackknife*, která je 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*” [26] a jedná se o speciální případ metody *bootstrap* (více o metodě B. Efron a R. Tibshirani v “*An Introduction to the Bootstrap*” [7]).

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 použít je střední hodnota \bar{x} . Pro původní dataset obsahující n pozorování lze střední hodnotu odhadnout dle vzorce $\bar{x} = \frac{1}{n} \sum_{i=1}^n \bar{x}_i$. Střední hodnota Jackknife podsouborů se vyhodnotí následovně:

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

Porovnání lze provést 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, tj. přítomnost odlehlych pozorování. Metoda může být také použita k odhadu skutečné, neovlivněné střední hodnoty datasetu. [15]

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. Proto je v řadě případů vhodné použít mahalanobisovou vzdálenost. Je definovaná matice $\mathbf{X}(n \times p)$, obsahující n objektů \mathbf{x}_i a p proměnných. Euklidovská vzdálenost mezi vektorem i -tého řádku $\mathbf{x}_i(1 \times p)$ této matice a vektoru středních hodnot $\bar{\mathbf{x}}(1 \times p)$ se spočítá jako

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

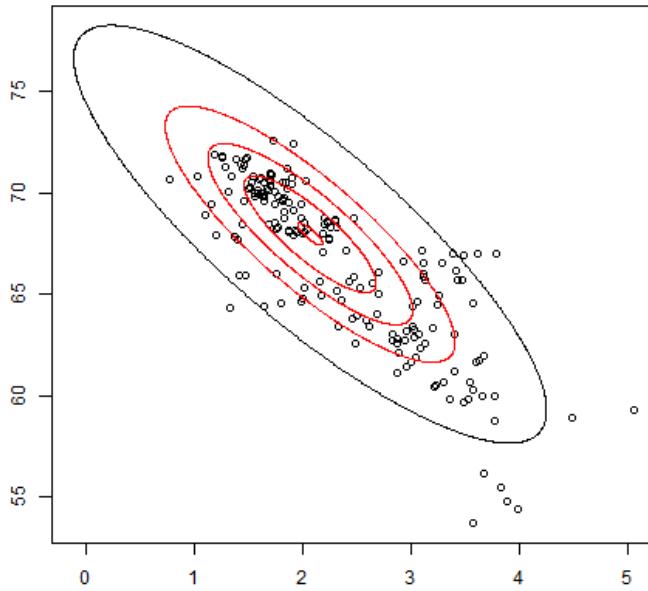
zatímco mahalanobisova vzdálenost se spočítá jako

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

kde \mathbf{C}_x je kovarianční matice. [6]

Na obrázku 15 jsou znázorněny elipsy mahalanobisových vzdáleností, kde každá elipsa představuje vzdálenost od průměru. Z tohoto je zřejmé, 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 vzdáleností umožňuje identifikaci odlehlych pozorování.

Metoda byla představena P.C. Mahalanobisem v roce 1936 ve článku “On the Generalized Distance in Statistics” [14]. Mahalanobisové vzdálenosti se používají nejenom k nalezení odlehlych 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 15: 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 matice \mathbf{H} , která je produktem metody nejmenších čtverců a je definována

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T.$$

Model lineární regrese může být zapsán následovně:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

kde vektor vysvětlované proměnné je \mathbf{y} , matice vysvětlujících proměnných je \mathbf{X} , vektor regresních koeficientů, který je odhadován, je $\boldsymbol{\beta}$ a vektor náhodné složky je $\boldsymbol{\varepsilon}$. Metoda nejmenších čtverců poskytuje řešení regresních rovnic:

$$\boldsymbol{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Lze dosadit:

$$\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}.$$

Výsledný vektor má tvar $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$, kde \mathbf{H} je projekční matice. [3]

3.2 Náhrada chybějících pozorování

Problém chybějících pozorování spočívá v neschopnosti jejich zpracovávání některými metodami. 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 kovarianci datasetu, a proto se nedoporučuje v případě vyššího podílu chybějících údajů. Další možnou metodou je náhrada náhodným číslem generovaným z příslušného rozdělení (parametry jsou odhaduty z výběru). V tomto případě se respektuje variabilita datasetu, ale nerespektuje se jeho kovariance. Chybějící údaje lze také odvodit pomocí známých hodnot na základě pomocné jednoduché lineární regresní funkce. Tato metoda respektuje nejenom variabilitu vzorku, ale i jeho korelační strukturu. [18]

3.3 Transformace dat

Jedním z cílů transformace dat je dosažení srovnatelnosti proměnných: sjednocení měřítka, variace a typu proměnných. Hlavním využitím je splnění podmínek vyžadovaných metodami, například podmínky normality, kde je snaha převést data na normální rozdělení, snížení vlivu rušivých proměnných (odlehlych hodnot) atd. [25] Rozdělujeme transformaci lineární (centrování, normování) a nelineární (plynoucí z typu a charakteru dat).

Lineární transformace zachovává lineární vztahy mezi proměnnými. Jedním z příkladů takovéto 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 tudíž rovnají nule. Toto lze zapsat následovně:

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

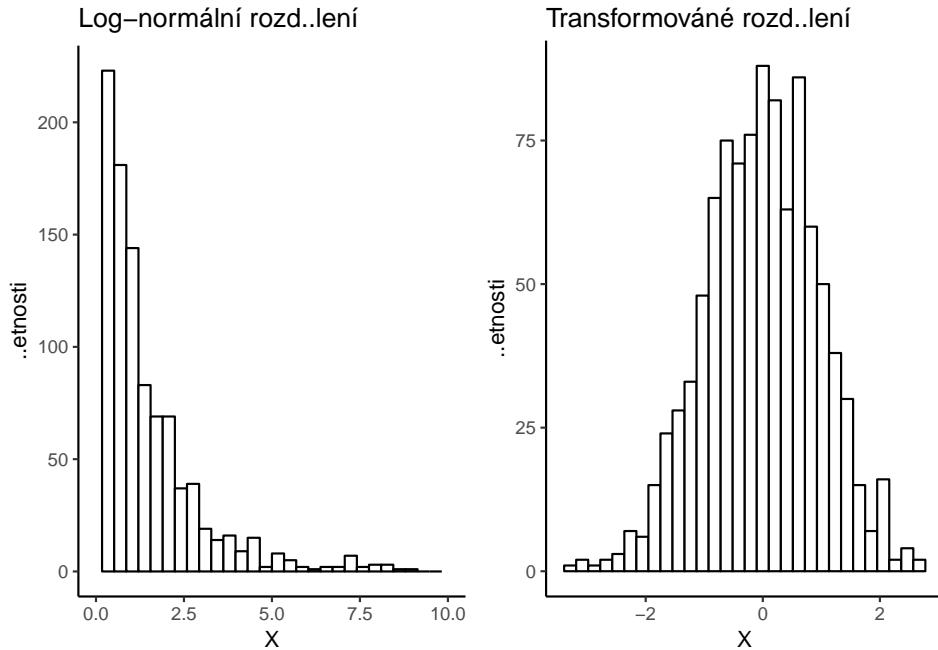
Vektor průměrů \bar{v} je nulový, kovariance a korelace proměnných zůstává nezměněna. [9] Další často využívanou metodou je metoda normalizace dat. Tato metoda transformuje měřítka vzorků pro možnost jejich porovnání (eliminuje jednotky měření), po úpravě střední hodnota vzorku tedy odpovídá nule a odchylka 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{z} je nulový a kovariance vektoru nových proměnných se shoduje s korelací původního vektoru. [1]

Nelineární transformace vyplývá z typu dat a mění (snižuje či zvyšuje) lineární vztahy mezi proměnnými a to znamená, že nezachovává korelací mezi nimi. Pokud data

mají charakter absolutní četnosti, používá se odmocninová transformace $X' = \sqrt{X}$, pokud odpovídají log-normálnímu rozdělení, používá se logaritmická transformace $X' = \log_{10} X$ atd. Logaritmus náhodné veličiny s log-normálním rozdělením má normální rozdělení (viz obrázek 16). Logaritmická transformace může být použita pouze u nezáporných rozdělení. [31] [12]

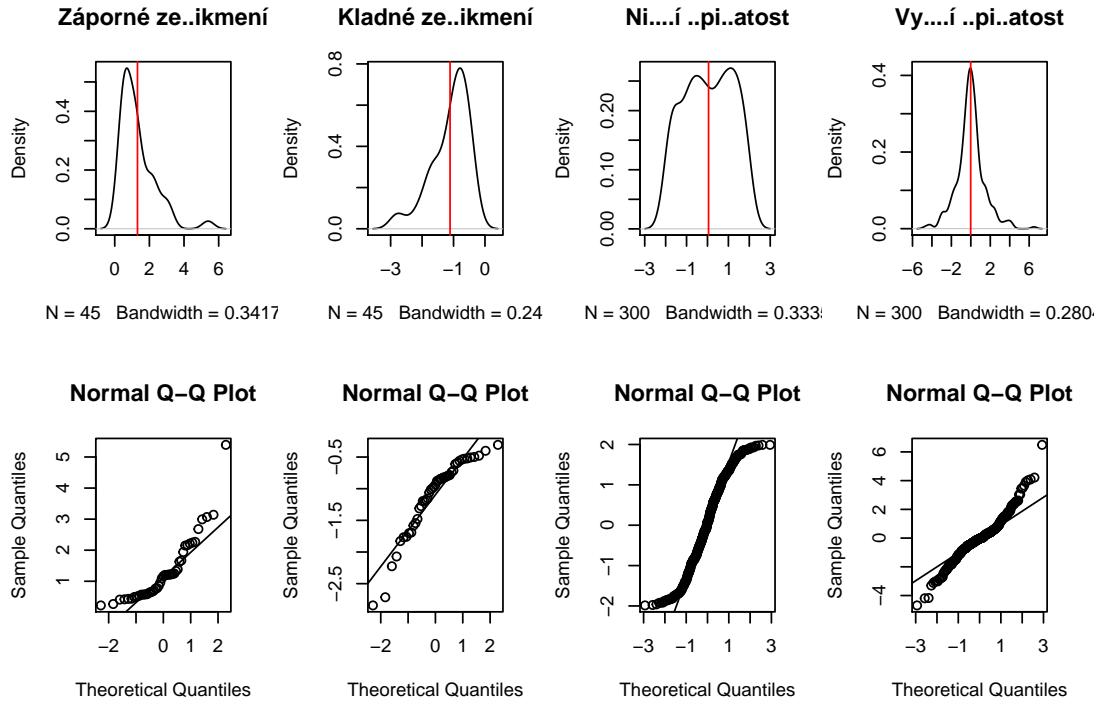


Obrázek 16: Log-normální rozdělení transformováné na normální rozdělení

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.

Zjistit zda-li vzorek pochází z normálního rozdělení lze grafickým posouzením nebo pomocí testů normality. Mezi nástroje grafického posouzení normality se řadí 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 17. Dále existuje řada testů normality, zde jsou popsány testy Shapiro-Wilk (SW) a Jarqua-Bera (JB).



Obrázek 17: Vztah hustoty rozdělení a Q-Q grafu pro různá narušení normality

Shapiro-Wilk test byl poprvé představen v roce 1965 S. S. Shapirem a M. Wilkem [23]. Metoda dokáže pracovat se vzorky velikosti 12 až 5000 pozorování. 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:

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

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 hypotéza se schoduje s SW testem. Používá se pro větší datasety nad 2000 pozorování. [17]

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

Praktická část

Seznam obrázků

1	Kombinace různých vuzuálních technik, Playfair 1801	10
2	Dobytek odeslaný z celé Francie ke spotřebě v Paříži, Minard 1858	11
3	Mapa světové migrace, Minard 1858	11
4	Postup Napoleonských vojsk v letech 1812-13, Minard 1869	12
5	Ukázky vizualizaci ze začatku 20. století, Brinton 1919	13
6	Porovnání základních typů grafů: (a) - bodový, (b) - liniový, (c) - složený, (d) - histogram	16
7	Hustota pravděpodobnosti normálního rozdělení (a) a její distribuční funkce (b)	17
8	Q-Q Graf (a) a P-P Graf (b)	18
9	Boxplot	19
10	Ukázka jednoduchého sloupcového grafu	20
11	Histogram s odhadem hustoty pravděpodobnosti	22
12	Skládaný sloupcový graf transformovaný do polárního souřadnicového systému	22
13	Ukázka jednoduchého koláčového grafu	23
14	Posloupnost datové analýzy	24
15	Mahalanobisovy vzdálenosti	27
16	Log-normální rozdělení transformované na normální rozdělení	29
17	Vztah hustoty rozdělení a Q-Q grafu pro různá narušení normality	30

Seznam tabulek

1	Základní atributy parametru ‘type’	16
2	Funkce pro práci s rozděleními	17

Literatura

- [1] Abdi, H. and Williams, L. 2010. Normalizing data. Encyclopedia of research design. *Thousand Oaks, CA: Sage.* (2010).
- [2] Brinton, W.C. 1919. *Graphic methods for presenting facts.* The Engineering Magazine Company, New York.
- [3] 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).
- [4] Chang, W. 2012. *R graphics cookbook: Practical recipes for visualizing data.* O'Reilly Media.
- [5] Cleveland, W.S. 1994. *The elements of graphing data.* Hobart Press.
- [6] De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D.L. 2000. The mahalanobis distance. *Chemometrics and intelligent laboratory systems.* 50, 1 (2000).
- [7] Efron, B. and Tibshirani, R. 1994. *An introduction to the bootstrap.* Taylor & Francis Ltd.
- [8] Friendly, M. 2006. A brief history of data visualization. *Handbook of computational statistics: Data visualization.* C. Chen, W. Härdle, and A. Unwin, eds. Springer-Verlag.
- [9] Hebák, P., Hustopecký, J., Jarošová, E. and Pecáková, I. 2007. *Vícerozměrné statistické metody (1).* Informatorium, Praha.
- [10] Histogram - wikipedia: <https://en.wikipedia.org/wiki/Histogram>.
- [11] How william cleveland turned data visualization into a science: <https://priceconomics.com/how-william-cleveland-turned-data-visualization/>.
- [12] Kutner, M.H., Nachtsheim, C.J., Neter, J. and Li, W. 2004. *Applied linear statistical models.* McGraw-Hill/Irwin.
- [13] Maciejewski, R. 2011. *Data representations, transformations, and statistics for visual reasoning.* Morgan & Claypool Publishers.
- [14] Mahalanobis, P.C. 1936. On the generalised distance in statistics. *Proceedings National Institute of Science.* 2, 1 (1936).
- [15] McIntosh, A. 2016. The jackknife estimation method. *arXiv.* (2016).
- [16] Novovičová, J. 2006. *Pravděpodobnost a matematická statistika.* Praha: Vydavatelství ČVUT,
- [17] 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).

- [18] Pecáková, I. 2014. Problém chybějících dat v dotazníkových šetřeních. *Politická ekonomie*. 2014, (Jan. 2014).
- [19] P–p plot - wikipedia: https://en.wikipedia.org/wiki/P%E2%80%93P_plot.
- [20] R: Generic x-y plotting: <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/plot.html>.
- [21] R: The r graphics package: <https://stat.ethz.ch/R-manual/R-devel/library/graphics/html/graphics-package.html>.
- [22] Rahlf, T. 2017. *Data visualisation with r: 100 examples*. Springer.
- [23] 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>.
- [24] Teator, P. 2011. *R cookbook: Proven recipes for data analysis, statistics, and graphics*. O'Reilly Media.
- [25] Transformation of data: <http://statisticalconcepts.blogspot.cz/2010/02/transformation-of-data-validity-of.html>.
- [26] Tukey, J.W. 1958. Bias and confidence in not quite large samples. *Annals of Mathematical Statistics*. 29, (1958).
- [27] Tukey, J.W. 1977. *Exploratory data analysis*. ADDISON WESLEY PUB CO INC.
- [28] Tukey, J.W. 1962. The future of data analysis. *Annals of Mathematical Statistics*. 33, 1 (Mar. 1962), 1–67. DOI:<https://doi.org/10.1214/aoms/1177704711>.
- [29] Wickham, H. and Grolemund, G. 2017. *R for data science*. O'Reilly Media.
- [30] Wilkinson, L., Wills, D., Rope, D., Norton, A. and Dubbs, R. 2006. *The grammar of graphics*. Springer New York.
- [31] Zumel, N. and Mount, J. 2014. *Practical data science with r*. Manning.