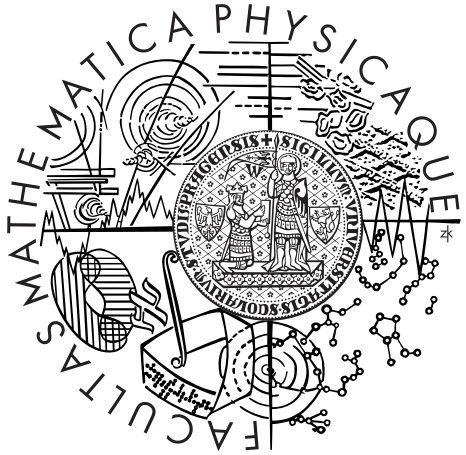
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

BAKALÁŘSKÁ PRÁCE



Anastasia Tyuleneva

Grafické znázornění směrových dat

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce:	doc. Mgr. Zdeněk Hlávka, Ph.D.
Studijní program:	Matematika
Studijní obor:	Finanční matematika

Praha 2014

Ráda bych poděkovala vedoucímu práce doc. Mgr. Zdeňku Hlávkovi, Ph.D. za ochotu spolupracovat, za čas, který mi věnoval a hlavně za odborné rady, kterými přispěl k vypracování této bakalářské práce. Rovněž bych chtěla poděkovat své rodině a kamarádům za veškerou podporu a pomoc.

Prohlašuji, že jsem tuto bakalářskou práci vypracovala samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova v Praze má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... Podpis

Název práce: Grafické znázornění směrových dat

Autor: Anastasia Tyuleneva

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: doc. Mgr. Zdeněk Hlávka, Ph.D., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Cílem této práce je rozšíření znalostí o možnostech přehledného zobrazení směrových dat pomocí různých druhů krabicového grafu (boxplot). V práci je popsán klasický boxplot, mimo jiné je detailně rozebrán vztah mezi výběrovými a teoretickými kvantily. V teoretické části je popsán samotný krabicový graf pro teoretická a výběrová data, zejména pak jednotlivé součásti tohoto grafu, což tvoří základ pro následující části práce. Pak bude následovat konstrukce směrového boxplotu pro dvourozměrná směrová data a odvození jeho vlastností pomocí von Misesova rozdělení. Poslední kapitola této bakalářské práce obsahuje krátký popis způsobu konstrukce vícerozměrného boxplotu neboli bagplotu pro třírozměrné Fisherovo rozdělení.

Klíčová slova: krabicový diagram, boxplot, bagplot, směrová data, von Misesovo rozdělení, Fisherovo rozdělení

Title: Graphical methods for directional data

Author: Anastasia Tyuleneva

Department: Department of Probability and Mathematical Statistics

Supervisor: doc. Mgr. Zdeněk Hlávka, Ph.D.,

Department of Probability and Mathematical Statistics

Abstract: This work is focused on the improvement of knowledge about the graphical directional data visualizing using different kinds of boxplot. The method and restrictions of the boxplot construction are depicted in this work, especially the need of knowledge of the relationship of sample $\alpha-$quantiles to the theoretical quantiles. The first part is focused on the most importat information of the theoretical background, boxplot and its parts description. In the second part the construction of the circular boxplot for two-dimensional directional data and representation of their properties using von Mises distribution is described. The final part is consisted of a brief description of the method of construction of multidimensional boxplot or bagplot for three-dimensional Fisher distribution.

Keywords: boxplot, bagplot, directional data, von Mises distribution, Fisher distribution

Obsah

Úvod

Krabicový graf neboli boxplot je jeden ze způsobů grafického zobrazení numerických dat, který obsahuje informace o maximální a minimální hodnotě souboru zkoumaných dat, mediánu, horním a dolním kvartilu tohoto souboru a některé další informace. Krabicový graf poprvé použil John Tukey v roce 1970 [Tukey, 1978]. V současné době je jedním z nejrychlejších způsobů zkoumání jednoho nebo více souborů dat. Boxplot je jednoduší než histogram, ale má své výhody: kompaktnost, jednoduchost a přehlednost. Zobrazení tohoto grafu nevyžaduje velké množství místa a umožňuje jednoduše porovnávat četnost dat v různých souborech. Tento typ grafu je názorný a jednoduchý k interpretaci, a právě proto se často používá v publikacích k zobrazení dat.

Krabicové grafy zobrazují rozdíly mezi datovými soubory bez jakýchkoli předpokladů normálního rozdělení dat, jsou neparametrické. Graf se sestává z krabičky a „vousů“. Je orientován vertikálně, resp. horizontálně, potřebné údaje zjišťujeme na vertikální ose y, resp. horizontální ose x. Minimální a maximální hodnoty jsou dané začátkem dolního a koncem horního vousu, pokud neexistují nějaké extrémní nebo odlehlé hodnoty, které budou zobrazeny samostatně. Dolní a horní kvartily jsou zaznamenány pomocí spodního a horního okraje krabičky a medián je reprezentován úsečkou uvnitř krabičky. V intervalu daném dolním a horním kvartilem se nachází 50% hodnot, z ostatních 50% hodnot, pokud neexistují odlehlé hodnoty, 25% připadá na interval mezi začátkem dolního vousu a spodní hranou krabičky a taky na interval mezi horní hranou krabičky a koncem horního vousu.

Práce není zaměřena pouze na teoretické vysvětlení pojmu boxplot a na reprezentaci některých jednorozměrných rozdělení nezávislých náhodných veličin pomocí tohoto grafu, ale také na zobrazení dvourozměrných a třírozměrných směrových dat. Konkrétní vlastnosti krabicového grafu pro vícerozměrná směrová data budou odvozeny pro dvourozměrné von Misesovo rozdělení a také bude následovat krátký popis způsobu konstrukce boxplotu pro třírozměrné Fisherovo rozdělení.

Kapitola 1

Kvantily a boxplot

Kvantilová funkce

Nechť náhodná veličina X je definována na pravděpodobnostním prostoru $(\Omega, \mathcal{A}, \mathsf{P})$ a má distribuční funkcí $F_X(x)$.

Definice 1. *Mějme náhodnou veličinu X a distribuční funkci F_X. Pak* ***kvantilovou funkcí*** *nazveme funkci*

$$F_X^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F_X(x) \geq \alpha, \textit{ pro } 0 < \alpha < 1\}. \tag{1.1}$$

Poznámka 1. Distribuční funkce plně charakterizuje rozdělení pravděpodobnosti náhodné veličiny. Často je třeba najít takový bod x splňující $P(X \leq x_\alpha) \geq \alpha$ a $P(X < x_\alpha) \leq \alpha$, pro $\alpha \in (0{,}1)$. Problém je s body, kde funkce F_X má skok, a také s body, kde F_X neroste čili inverzní funkce by nebyla jednoznačná.

Kvantil

Definice 2. *Hodnota kvantilové funkce v bodě α, tj. $F_X^{-1}(\alpha)$, je α-kvantil a značí se x_α. Pro kvantil spojitého rozdělení platí:*

$$F_X(x_\alpha) = \int_{-\infty}^{x_\alpha} f_X(x)dx = \alpha.$$

Jinak řečeno, pro spojitou náhodnou veličinu X, která má distribuční funkci F_X, je α-kvantil x_α taková hodnota, pro niž platí, že výskyt hodnot menších než x_α nastane pouze s pravděpodobností α, tj. pro kterou je distribuční funkce $F_X(x_\alpha)$ rovna pravděpodobnosti α

$$\mathbb{P}(X \leq x_\alpha) = F_X(x_\alpha) = \alpha.$$

Poznámka 2. Kvantil x_α standardizované normální veličiny $U \sim N(0, 1)$ se značí u_α.

Existuje několik význačných hodnot kvantilů, které mají svá jména, viz tabulka 1.1.

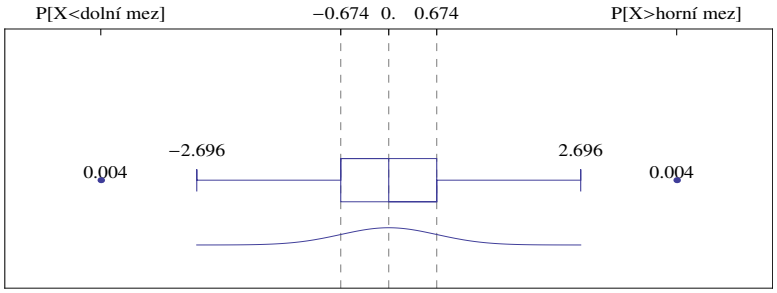
Kvantil	Název
$x_{0.50}$	medián
$x_{0.25}$	dolní kvartil
$x_{0.75}$	horní kvartil
$x_{0.10}$	1. decil
$x_{0.90}$	9. decil
$x_{0.01}$	1. percentil
$x_{0.99}$	99. percentil

Tabulka 1.1: Teoretické kvantily

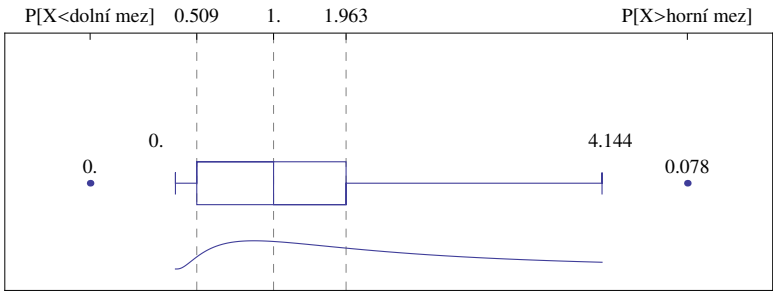
Teoretické boxploty

Pomocí teoretických kvantilů můžeme pro zvolené rozdělení spočítat a nakreslit teoretický boxplot. Následující obrázky 1.1, 1.2 a 1.3 ukazují, jak vypadá graf boxplotu pro konkrétní rozdělení.

Kvantil $LogNormalni(0,1)$ rozdělení se počítá pomocí kvantilu rozdělení $N(0,1)$. Obecně je $\alpha-$kvantil $LogNormalni(\mu,\sigma^2)$ $x_\alpha = \exp\{\mu + u_\alpha\sigma\}$, kde $\mu + u_\alpha\sigma$ je α-kvantil normálního rozdělení.



Obrázek 1.1: Krabicový graf pro $N(0,1)$ rozdělení.



Obrázek 1.2: Krabicový graf pro *LogNormalni*(0,1) rozdělení.

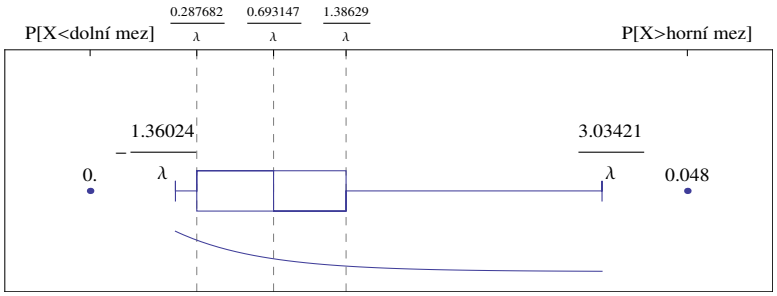
Příklad 1. Najdeme dolní a horní kvartil normálního rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$.

$$x_{0.25} = \mu + u_{0.25}\sigma = 0 + u_{0.25} = -0.674$$
$$x_{0.75} = \mu + u_{0.75}\sigma = 0 + u_{0.75} = 0.674$$

Dolní a horní kvartil lognormálního rozdělení s parametry $\mu = 0$ a $\sigma^2 = 1$ se rovná

$$x_{0.25} = \exp\{\mu + u_{0.25}\sigma\} = \exp\{-0.674\} = 0.509$$
$$x_{0.75} = \exp\{\mu + u_{0.75}\sigma\} = \exp\{0.674\} = 1.963.$$

Graf 1.3 exponenciálního rozdělení $Exp(\lambda)$ ukazuje, že hodnota kvantilu závisí na hodnotě parametru rozdělení. Pokud se změní parametry rozdělení, stejně se budou měnit i hodnoty kvantilů.



Obrázek 1.3: Krabicový graf pro $Exp(\lambda)$ rozdělení.

Přesto, že exponenciální a lognormální rozdělení je definované pro hodnoty $x > 0$, jejich teoretická dolní mez je záporná, a to kvůli tomu, že dolní mez je

$$x_{0.25} \leq 1.5(x_{0.75} - x_{0.25}).$$

Navíc pro lognormální a exponenciální rozdělení se dolní mez rovná 0, protože tato rozdělení nejsou definovaná pro $x < 0$ a $\lambda > 0$. Ale pravděpodobnost toho, že se mezi nezávislými náhodnými veličinami se stejným rozdělením vyskytne taková hodnota x, která je větší než horní mez, je malá. Pro $N(0,1)$ rozdělení je 0.004, viz obrázek 1.1, pro $LogNormalni(0,1)$ je 0.08, viz obrázek 1.2 a pro $Exp(\lambda)$ pravděpodobnost je 0.05, viz obrázek 1.3.

Kapitola 2

Krabicový graf. Výběrový kvantil

2.1 Výběrový kvantil

Nechť $X_1, \ldots, X_{\mathrm{n}}$ je náhodný výběr, tj. n-tice nezávislých náhodných veličin, které mají stejné rozdělení, tj. mají stejnou distribuční funkci F_X. Datový soubor získaný náhodným výběrem lze popsat pomocí číselných charakteristik, které nazýváme výběrové charakteristiky. Dělíme je na **míry polohy** a **míry variability**.

Výběrový kvantil $\tilde{x}_\alpha$ patří k mírám polohy a je obecně definován jako hodnota rozdělující výběrový soubor na dvě části. První část obsahuje $n\alpha$ hodnot menších než kvantil a druhá část $n(1-\alpha)$ hodnot, které jsou rovny nebo větší než hodnota kvantilu.

Podobně rozeznáváme **výběrové percentily** $(\tilde{x}_{0.01}, \ldots, \tilde{x}_{0.99})$, **výběrové decily** $(\tilde{x}_{0.1}, \ldots, \tilde{x}_{0.9})$ a kvantily: **dolní kvartil** $\tilde{x}_{0.25}$, **medián** $\tilde{x}_{0.5}$, **horní kvartil** $\tilde{x}_{0.75}$.

Při určování výběrových kvantilů nejprve uspořádáme hodnoty souboru dat podle velikosti $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. Hledaný kvantil bude buď konkrétní hodnota ze souboru, nebo vážený průměr dvou určitých hodnot. Pořadové číslo pozorování, jejíž hodnota je hledaným kvantilem značené jako z_α, se počítá pomocí vzorce:

$$\mathrm{n}\alpha < z_\alpha < \mathrm{n}\alpha + 1\,.$$

Poznámka 3. (Medián) **Medián** $\tilde{x}_{0.5}$ souboru hodnot $x_1, x_2, \ldots, x_n$ je prostřední hodnota ze všech hodnot souboru seřazených podle velikosti.

Míra variability.

Definice 3. ***Mezikvartilové rozpětí*** *se nazývá rozdíl horního a dolního výběrového kvartilu, tj. rozdíl 75% a 25% kvartilu:*

$$\tilde{R}_q = \tilde{x}_{0.75} - \tilde{x}_{0.25}\,,$$

nebo decilové rozpětí

$$\tilde{R}_d = \tilde{x}_{0.90} - \tilde{x}_{0.10},$$

u rozsáhlejších souborů ještě analogické percentilové rozpětí.

2.2 Odlehlé a extrémní hodnoty

Při konstrukci boxplotu rozlišujeme *odlehlé* a *extrémní* hodnoty. Odlehlá měření silně zkreslují odhady polohy, takže velmi ztěžují provedení další statistické analýzy. Tyto hodnoty se co do velikosti značně liší od ostatních dat a lze je rozpoznat v diagnostickém grafu jako boxplot.

Odlehlé hodnoty leží buď v intervalu od $\tilde{x}_{0.25} - 3\tilde{R}_q$ do $\tilde{x}_{0.25} - 1.5\tilde{R}_q$, nebo v intervalu od $\tilde{x}_{0.75} + 1.5\tilde{R}_q$ do $\tilde{x}_{0.75} + 3\tilde{R}_q$, a extrémní jsou hodnoty menší než $\tilde{x}_{0.25} - 3\tilde{R}_q$ nebo větší než $\tilde{x}_{0.75} + 3\tilde{R}_q$.

Rozhodujícími pro výpočet jsou tedy 1.5 násobek a trojnásobek $\tilde{R}_q$. Můžeme tyto K-násobky mezikvartilového rozpětí měnit, ale tak, aby pravděpodobnost, že žádný prvek z daného rozdělení nebude větší než hodnota $\tilde{x}_{0.75}+\mathrm{K}(\tilde{x}_{0.75}-\tilde{x}_{0.25})$ nebo menší než hodnota $\tilde{x}_{0.25} - \mathrm{K}(\tilde{x}_{0.75} - \tilde{x}_{0.25})$, byla dostatečně velká.

2.3 Konvergence

Pro posloupnost $X_1, \ldots, X_\mathrm{n}$ nezávislých náhodných veličin se stejnou distribuční funkcí F_X je její výběrový α-kvantil definován jako α-kvantil empirické distribuční funkce F_n, resp. $F_n^{-1}(\alpha)$. Dále výběrový kvantil budeme značit $\tilde{x}_{\alpha n}$ nebo $\tilde{x}_\alpha$.

Věta 1. *Nechť* $0 < \alpha < 1$*, pak za předpokladu, že* x_α *je jediné řešení*

$F_X(x_-) \leq \alpha \leq F_X(x)$*, potom* $\tilde{x}_{\alpha n} \xrightarrow{P} x_\alpha$*.*

Důkaz. Podrobný důkaz je popsán v [Serfling, 1980, věta 2.3.1].

❑

Věta 1 říká, že výběrový kvantil $\tilde{x}_\alpha$ je konsistentním odhadem teoretického kvantilu x_α, pokud $F_X(x_\alpha) = \alpha$ a F_X existuje v levém okolí x_α.

Věta 2. *Nechť* $0 < \alpha < 1$ *a za předpokladu, že* F_X *je spojitá v* x_α*, pak platí*

1. $$\lim_{n\longrightarrow\infty} \mathbb{P}\left(\sqrt{n}(\tilde{x}_{\alpha n} - x_\alpha) \leq 0\right) = \Phi(0) = \frac{1}{2}\,. \tag{2.1}$$

2. *Pokud existuje derivace zleva* $F_X^{'}(x_{\alpha-}) > 0$ *pro* $t < 0$ *a zprava* $F_X^{'}(x_{\alpha+}) > 0$ *pro* $t > 0$*, pak:*

$$\lim_{n\longrightarrow\infty} \mathbb{P}\left(\frac{\sqrt{n}(\tilde{x}_{\alpha n} - x_\alpha)}{\sqrt{\alpha(1-\alpha)}/F_X'(x_{\alpha\mp})} \leq t\right) = \Phi(t)\,. \tag{2.2}$$

Důkaz. Podrobný důkaz je popsán v [Serfling, 1980, věta 2.3.3].

❑

Důsledek 1. *Nechť* $0 < \alpha < 1$ *a* F_X *je diferencovatelná v* x_α *a zároveň* $F_X^{'}(x_\alpha) > 0$*, potom* $\tilde{x}_{\alpha n}$ *má přibližně normální rozdělení*

$$N\left(x_\alpha, \frac{\alpha(1-\alpha)}{[F_X'(x_\alpha)]^2 n}\right)\,.$$

Důsledek 2. *Nechť* $0 < \alpha < 1$ *a* F_X *má hustotu* f *v okolí kvantilu* x_α *a* f *je kladná a spojitá v* x_α*, potom* $\tilde{x}_{\alpha n}$ *má přibližně normální rozdělení*

$$N\left(x_\alpha, \frac{\alpha(1-\alpha)}{f^2(x_\alpha)n}\right).$$

Důsledky 1 a 2 popisují případy, ve kterých $\tilde{x}_\alpha$ je asymptotický normální odhad kvantilu. Jinak řečeno, $\tilde{x}_\alpha$ při vhodném normalizovaném tvaru bude konvergovat v distribuci k normálnímu normovanému rozdělení $N(0,1)$.

Příklad 2. Pro normální normované rozdělení ukažme praktické realizace vlastnosti výběrových kvantilů, viz tabulka 2.1, konkrétně, s rostoucím n se hodnota výběrového kvantilu blíží k hodnotě teoretického kvantilu.

Kvantil	**teoretický**	**n=20**	**n=50**	**n=100**	**n=200**
$\alpha = 0.50$	0	0.041	−0.021	0.013	−0.093
$\alpha = 0.25$	−0.674	−0.652	−0.789	−0.647	−0.679
$\alpha = 0.75$	0.674	0.644	0.622	0.649	0.679
R_q	1.349	1.296	1.411	1.295	1.357
$\tilde{x}_{0.25} - 1.5R_q$	−2.698	−2.596	−2.905	−2.589	−2.715
$\tilde{x}_{0.75} + 1.5R_q$	2.698	2.587	2.738	2.592	2.715

Tabulka 2.1: Porovnání teoretických kvantilů s výběrovými kvantily $N(0,1)$ rozdělení

Na výběrový kvantil se můžeme dívat jako na odhad teoretického kvantilu. Můžeme najít interval, konkrétně předpovědní interval, ve kterém by se mohla s velkou pravděpodobností výběrová hodnota vyskytnout.

Poznámka 4. V praxi obvykle konstruujeme spíš interval spolehlivosti, t.j. interval, který se známou pravděpodobností překryje teoretickou hodnotu.

Na rozdíl od intervalu spolehlivosti, předpovědní interval se počítá pro známé rozdělení se známými parametry a pro předem danou pravděpodobnost toho, že náhodná hodnota, t.j. výběrový kvantil, padne do tohoto intervalu.

Díky tomu, že výběrový kvantil je asymptotickou normální veličinou, pak pro vícerozměrnou variantu důsledku 2 platí následující věta.

Věta 3. *Nechť* $0 < \alpha_1 < \cdots < \alpha_k < 1$ *a* F_X *má hustotu* f *v okolí kvantilů* $x_{\alpha_1}, \cdots . x_{\alpha_k}$ *a* f *je větší než nula a spojitá v* $x_{\alpha_1}, \cdots . x_{\alpha_k}$*. Pak* $(\tilde{x}_{\alpha_1}, \cdots . \tilde{x}_{\alpha_k})$ *má asymptotické normální rozdělení s vektorem středních hodnot* $(x_{\alpha_1}, \cdots . x_{\alpha_k})$ *a kovariancí* $\frac{\sigma_{ij}}{n}$*, kde*

$$\sigma_{ij} = \frac{\alpha_i(1-\alpha_j)}{f(x_{\alpha_i})f(x_{\alpha_j})} \text{ pro } i \leq j \tag{2.3}$$

a $\sigma_{ij} = \sigma_{ji}$ *pro* $i > j$*.*

Důkaz. Jedna možnost důkazu je popsána v [Serfling, 1980, věta 2.3.4] nebo jiná v [Serfling, 1980, věta 2.5.1] .

❑

Z důsledku věty 2 známe rozdělení náhodných veličin

$$\tilde{x}_{0.25} \sim N\left(x_{0.25}, \frac{0.25 * 0.75}{f^2(x_{0.25})n}\right) \text{ a } \tilde{x}_{0.75} \sim N\left(x_{0.75}, \frac{0.25 * 0.75}{f^2(x_{0.75})n}\right).$$

Při konstruování předpovědního intervalu nás budou zajímat takové meze, pro které by platilo

$$\mathbb{P}\left(\tilde{x}_{0.25} \in \left(x_{0.25} \mp u_{1-p/2}\sqrt{\frac{0.25 * 0.75}{f^2(x_{0.25})n}}\right)\right) = 1 - p\,, \tag{2.4}$$

a

$$\mathbb{P}\left(\tilde{x}_{0.75} \in \left(x_{0.75} \mp u_{1-p/2}\sqrt{\frac{0.75 * 0.25}{f^2(x_{0.75})n}}\right)\right) = 1 - p\,, \tag{2.5}$$

kde $0 \leq p \leq 1$.

Konstruování předpovědního intervalu pro $\tilde{x}_{0.25} - 1.5\tilde{R}_q$ a $\tilde{x}_{0.75} + 1.5\tilde{R}_q$ nebude tak jednoduché, ale s využitím věty 3 si lze horní (resp. dolní) mez představit jako lineární kombinace náhodných vektorů, tzn.

$$(2.5,\ -1.5)\begin{pmatrix} \tilde{x}_{0.25} \\ \tilde{x}_{0.75} \end{pmatrix}$$

a rozdělení těchto veličin už je

$$\mathbf{a}\begin{pmatrix} \tilde{x}_{0.25} \\ \tilde{x}_{0.75} \end{pmatrix} \sim N\left(\mathbf{a}\begin{pmatrix} x_{0.25} \\ x_{0.75} \end{pmatrix}, \frac{\mathbf{a}\ \sum\ \mathbf{a}^T}{n}\right),$$

kde $\mathbf{a}$ je řádkový vektor konstantních koeficientů, $(x_{0.25}, x_{0.75})$ teoretické kvantily, $\sum$ je kovarianční matice (2.3), n rozsah hodnot.

Na závěr této kapitoly ukážeme nejen praktické odvození konvergence výběrových hodnot k teoretickým kvantilům pro $N(0,1)$ rozdělení, viz tabulka 2.1, ale také předpovědní intervaly pro rozdělení z příkladu 1, kapitoly I.

Příklad 3. Najdeme 95% předpovědní interval horní meze pro náhodný výběr z $N(0,1)$ rozdělení a $n = 200$. Podle důsledku věty 2 a věty 3 spočítáme parametr

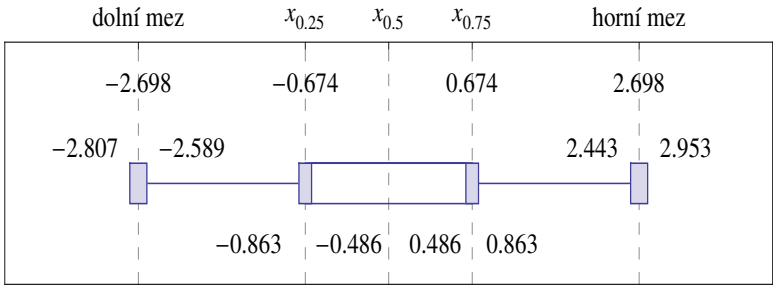
$$\mu = (2.5, -1.5)\begin{pmatrix} x_{0.75} \\ x_{0.25} \end{pmatrix} = 2.698$$

a

$$\sigma^2 = \frac{\mathbf{a}\ \sum\ \mathbf{a}^T}{n} = \frac{(2.5, -1.5)\sum(2.5, -1.5)^T}{200} = 0.017.$$

Asymptotický předpovědní interval střední hodnoty se známým rozptylem normálního rozdělení je

$$\mu - u_{0.975}\sigma = 2.698 - u_{0.975} * \sqrt{0.017} = 2.443$$
$$\mu + u_{0.975}\sigma = 2.698 + u_{0.975} * \sqrt{0.017} = 2.953.$$



Obrázek 2.1: Krabicový graf pro $N(0,1)$ rozdělení s vyznačenými předpovědními intervaly pro kvartily, dolní a horní meze, pro $n = 200$.

Předpovědní interval pro	$\tilde{x}_{0.25} - 1.5\tilde{R}_q$	$\tilde{x}_{0.25}$	$\tilde{x}_{0.75}$	$\tilde{x}_{0.75} + 1.5\tilde{R}_q$
Teoretická hodnota	−1.361	0.288	1.386	3.034
Horní hranice	−1.298	0.368	1.626	3.063
Dolní hranice	−1.422	0.208	1.146	3.005

Tabulka 2.2: Předpovědní interval pro $Exp(1)$ rozdělení a $n = 200$

Předpovědní interval pro	$\tilde{x}_{0.25} - 1.5\tilde{R}_q$	$\tilde{x}_{0.25}$	$\tilde{x}_{0.75}$	$\tilde{x}_{0.75} + 1.5\tilde{R}_q$
Teoretická hodnota	−1.671	0.509	1.963	4.143
Horní hranice	−1.529	0.606	2.334	4.183
Dolní hranice	−1.813	0.413	1.592	4.104

Tabulka 2.3: Předpovědní interval pro $LN(0,1)$ rozdělení a $n = 200$

Tabulky 2.2 a 2.3 obsahují jednotlivé předpovědní intervaly pro konkrétní hodnoty kvantilů $Exp(1)$ a $LN(0,1)$ rozdělení. Výběrový hodnoty α-kvantilů s 95% pravděpodobností budou ležet uvnitř odhadnutých intervalů.

Dostali jsme se k závěru této kapitoly a právě k tomu, co nám říká teoretická, praktická a grafická část. Všechny výběrové $\alpha-$kvantily z nějakého dostatečně velkého výběru nezávislých náhodných veličin se stejným rozdělením, které dále budeme používat při práci s vícerozměrnými daty, konvergují k teoretickým $\alpha-$kvantilům a navíc mají asymptotické normální rozdělení s příslušnými parametry.

Kapitola 3

Směrový krabicový graf

Ve dvoudimenzionálním prostoru při konstruování směrového krabicového grafu se používají směrová data, tj. data v úhlových stupních nebo radiánech. [Fisher, 1996] říkal, že směrový graf existuje od roku 1858 a je více známý jako růžicový diagram nebo větrná růžice. Obyčejně standardní krabicový graf není jednoznačně definován pro směrová data, proto se v praxi používá směrový krabicový graf, resp. směrový boxplot, který slouží k nalezení odlehlých hodnot výběrových směrových veličin.

3.1 Směrová data

Obvykle jsou při výpočtu směrová data vyjadřovaná pomocí směrového úhlu θ, pro který je třeba zvolit počáteční bod a směr otáčení ve směru hodinových ručiček nebo proti němu, přičemž všechny výsledky musí být ekvivalentní, nezávislé na výběru. Každý směr je reprezentován jako jednotkový vektor υ v rovině, který odpovídá jedinému uhlu θ, $0 < \theta < 2\pi$. Vektor υ lze popsat několika různými způsoby:

1. Pomocí kartézské soustavy souřadnic.
   Nechť $S^1 := \{\mathbf{x} \in \mathrm{R}^2 : \|\mathbf{x}\| = 1\}$ je jednotková kružnice v rovině s centrem v nule, pak každý bod na této kružnici lze zapsat ve tvaru $\mathbf{x} = (x_1, x_2)'$, kde

   $$x_1 = \cos\theta \text{ a } x_2 = \sin\theta, \text{ pro } \theta \in (0, 2\pi),$$

   x_1 a x_2 jsou hodnoty kartézské soustavy souřadnic.

2. Pomocí polární soustavy souřadnic.
   Každý bod v rovině lze zapsat v polární soustavě souřadnic $(r,\varphi)'_p$, přičemž r udává vzdálenost bodu od počátku souřadnic a $\varphi, \varphi \in (0, 2\pi)$, je úhel spojnice tohoto bodu a počátku od zvolené osy ležící v rovině. Navíc libovolný bod ve tvaru souřadnic kartézské soustavy lze převést na tvar polární soustavy $(r,\varphi)'_p$, kde $r = 1$ a stejně i opačně z polární na kartézskou.

3. Pomocí komplexní roviny.
   Nechť $C := \{z \in \mathrm{C} : \|z\| = 1\}$ je jednotková kružnice v komplexní rovině, pak každý bod na této kružnici lze zapsat pomocí komplexních čísel $z \in \mathrm{C}$. Protože $\|z\| = 1$, pak všechna $z \in \mathrm{C}$ lze zapsat ve tvaru:

$$z = e^{i\theta} = \cos\theta + i\sin\theta, \text{ pro } \theta \in (0, 2\pi).$$

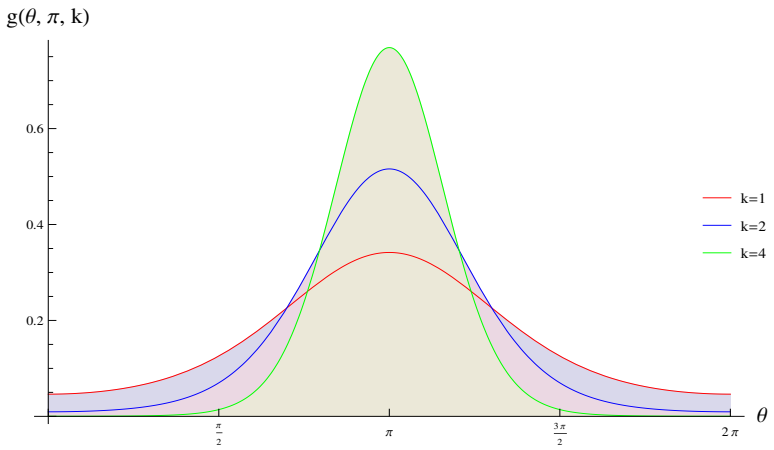
Cirkulární teorie pravděpodobnosti se liší od lineární teorie pravděpodobnosti, se kterou jsme zvyklí pracovat, rozdíly se vyskytují ve většině základních pojmů. Dokonce i definice náhodné proměnné musí být přizpůsobena struktuře kruhu.

Definice 4. *Dvoudimenzionální náhodný vektor* $\boldsymbol{X}$ *mající kruhové rozdělení v* R^2 *nazýváme jednotkový náhodný vektor, jestliže nabývá hodnot pouze na oblouku jednotkového kruhu v rovině se středem v počátku.*

Jiné definice a některé vlastnosti směrových dat viz [Malá, 2012, definice 2.1.1.-2.2.1].

3.2 Von Misesovo rozdělení

Nejdůležitější cirkulární pravděpodobnostní model je von Misesovo rozdělení $M_2(\boldsymbol{\mu}, k)$ se střední hodnotou $\boldsymbol{\mu}$ a mírou koncentrace k. Von Misesovo rozdělení poprvé použil Richard von Mises v roce 1918. Někdy se v cirkulární statistice von Misesovo rozdělení nazývá cirkulární (kruhové) normální rozdělení $CN(\boldsymbol{\mu}, k)$ [Jammalamadaka a Sengupta, 2001, 2.2.4] a je analogem Gaussova normálního rozdělení v lineární statistice.



Obrázek 3.1: Hustoty von Misesova rozdělení $M(\boldsymbol{\mu}, k)$ pro $\boldsymbol{\mu} = \pi$ a $k = \{1, 2, 4\}$.

Von Misesovo rozdělení pochází z p-dimenzionálního von Mises-Fisherova rozdělení $M_p(\boldsymbol{\mu}, k)$ s hustotou

$$f(\mathbf{x}; \boldsymbol{\mu}, k) = c_p(k) \exp\{k\, \boldsymbol{\mu}' \mathbf{x}\}, \quad \mathbf{x} \in S^{p-1}, \tag{3.1}$$

kde

$$c_p(k) = k^{p/2-1} \frac{1}{(2\pi)^{p/2} I_{p/2-1}(k)}, \tag{3.2}$$

$k \geq 0$ a $I_{p/2-1}(k)$ je modifikovaná Besselova funkce proměnné k prvního druhu a řádu $p/2 - 1$ [Abramowitz a Stegun, 1965, 9.6.19].

Pro $p = 2$ von Mises-Fisherovo rozdělení bude mít tvar klasického von Misesova rozdělení $M_2(\mu, k)$ s hustotou

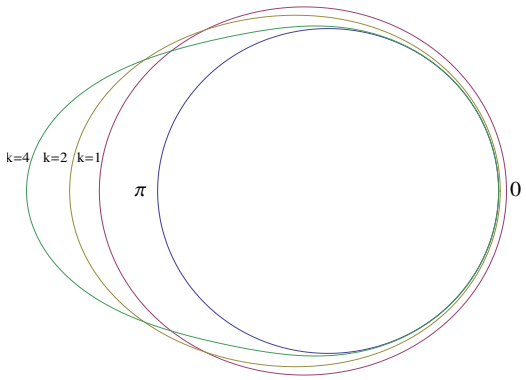
$$f(\mathbf{x};\boldsymbol{\mu},k)=\frac{1}{2\pi I_0(k)}\exp\{k\boldsymbol{\mu}'\mathbf{x}\}\quad \mathbf{x}\in S^1.$$

Poznámka 5. Modifikovanou Besselovu funkci prvního druhu v závislosti na p lze vyjádřit v integrálním tvaru

$$I_p(k)=\frac{1}{2\pi}\int_0^{2\pi}\cos(p\theta)\exp\{k\cos(\theta)\}\,d\theta,$$

viz [Jammalamadaka a Sengupta, 2001, 2.2.4].

Von Misesovo rozdělení je většinou definováno pro nezávislé náhodné úhly θ, $0\le\theta\le 2\pi$.



Obrázek 3.2: Hustoty von Misesova rozdělení $M(\boldsymbol{\mu},k)$ pro $\boldsymbol{\mu}=\pi$ a $k=\{0,1,2,4\}$ na kruhu.

Definice 5. *Necht' θ je náhodný úhel, říkáme že má von Misesovo rozdělení s hustotou*

$$q(\theta,\mu,k)=\frac{1}{2\pi I_0(k)}\exp\{k\cos(\theta-\mu)\},\ 0<\theta\le 2\pi, \tag{3.3}$$

kde $0\le\mu<2\pi$ a $k\ge 0$ jsou parametry.

Dále musíme rozlišovat μ úhel a $\boldsymbol{\mu}$ vektor. Pro $p=2$ je $\boldsymbol{\mu}=(\cos\mu,\sin\mu)'$. Navíc dáme přednost polárním souřadnicím před kartézskými pro reprezentaci $\mathbf{x}$ a $\boldsymbol{\mu}$, tj.

$$\boldsymbol{\mu}'\mathbf{x}=(\cos\mu,\,\sin\mu)\,(\cos\theta\ \sin\theta)'=\cos(\theta-\mu).$$

Parametry

Protože kosinus dosahuje svého maxima v bodě nula, hustota (3.3) von Misesova rozdělení bude mít maximum v $\theta=\mu$, jinak řečeno μ je modus s hodnotou

$$f(\mu)=\frac{e^k}{2\pi I_0(k)}. \tag{3.4}$$

Naopak, když $\cos\pi = -1$, pak kosinus dosahuje svého minima a minimální hodnota hustoty von Misesova rozdělení je

$$f(\mu \pm \pi) = \frac{e^k}{2\pi I_0(-k)}. \tag{3.5}$$

Z podílu (3.4) a (3.5) dojdeme k

$$\frac{f(\mu)}{f(\mu \pm \pi)} = e^{2k}. \tag{3.6}$$

Rovnost (3.6) ukazuje, že čím větší je hodnota k, tím větší bude podíl $f(\mu)$ a $f(\mu \pm \pi)$, proto parametr k měří koncentraci vůči střední hodnotě μ.

3.3 Teoretické kvantily směrových dat

Teoretický směrový medián

Medián pozorování $\theta_1, \ldots, \theta_n$ představuje bod $\theta_{0.50}$, pro který platí, že polovina všech dat leží v intervalu $arc\,[\theta_{0.50}, \theta_{0.50} + \pi]$, přičemž $f(\theta_{0.50}) \geq f(\theta_{0.50} + \pi)$.

Definice 6. *Směrový medián $\theta_{0.50}$ je jedním z řešení rovnice*

$$\int_{\theta_{0.50}}^{\theta_{0.50}+\pi} f(u)\,du = \frac{1}{2}, \tag{3.7}$$

kde u je náhodný úhel s hustotou $f(u)$.

Příklad 4. Ověříme (3.7) pro $M_2(0{,}1)$. Medián von Misesova rozdělení s parametry μ a k se rovná střední hodnotě tohoto rozdělení, tj. parametru μ. V případě $M_2(0{,}1)$ medián je $\mu = \theta_{0.50} = 0$. Podle (3.7) dostáváme rovnici

$$\int_{\theta_{0.50}}^{\theta_{0.50}+\pi} f(u)\,du = \int_{\mu}^{\mu+\pi} \frac{1}{2\pi I_0(k)} \exp\{k\cos(u-\mu)\}\,du = \int_{0}^{\pi} \frac{1}{2\pi I_0(1)} \exp\{\cos(u)\}\,du =$$
$$\frac{1}{2\pi I_0(1)} \int_{0}^{\pi} \exp\{\cos(u)\}\,du = \frac{1}{2\pi I_0(1)}\ \pi I_0(1) = \frac{1}{2},$$

kde $I_0(1)$ je podle poznámky 5 modifikovaná Besselova funkce prvního druhu s $p = 0$ a $k = 1$.

Dolní a horní směrové kvartily

Dolní $\theta_{0.25}$ a horní $\theta_{0.75}$ kvartily lze definovat stejně jako medián z definice 6, tj. $\theta_{0.25}$ a $\theta_{0.75}$ pro soubor směrových dat je řešením rovnic

$$\int_{\theta_{0.50}-\theta_{0.25}}^{\theta_{0.50}} f(u)\,du = \frac{1}{4} \text{ a } \int_{\theta_{0.50}}^{\theta_{0.50}+\theta_{0.75}} f(u)\,du = \frac{1}{4}. \tag{3.8}$$

Příklad 5. Stejně jako v příkladě 4 ověříme (3.8) pro $M_2(0{,}1)$. V případě $M_2(0{,}1)$ dolní kvartil je -0.809 a horní kvartil je 0.809. Protože $M_2(0{,}1)$ je symetrické rozdělení kolem parametru $\mu = 0$, stačí odvodit platnost (3.8) jenom pro horní kvartil.

$$\int_{\theta_{0.50}}^{\theta_{0.50}+\theta_{0.75}} f(u)\,du = \int_{\mu}^{\mu+\theta_{0.75}} \frac{1}{2\pi I_0(k)} \exp\{k\cos(u-\mu)\}\,du =$$

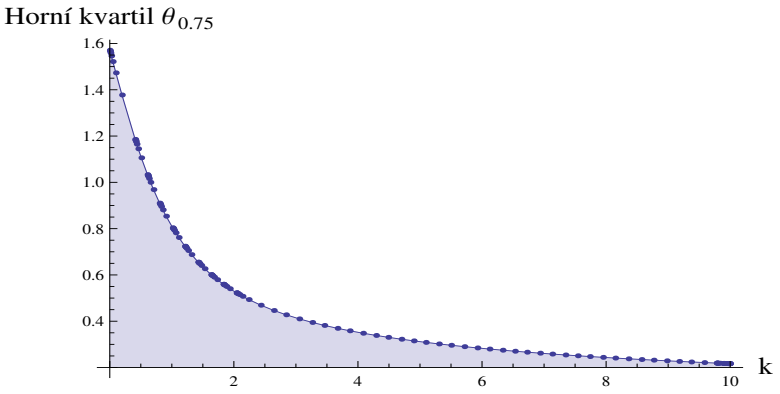
$$\int_{0}^{\theta_{0.75}} \frac{1}{2\pi I_0(1)} \exp\{\cos(u)\}\,du = \frac{1}{2\pi I_0(1)} \int_{0}^{\theta_{0.75}} \exp\{\cos(u)\}\,du = \frac{1}{2\pi I_0(1)}\,\frac{\pi I_0(1)}{2} = \frac{1}{4}.$$

Obecně pro von Misesovo rozdělení $M_2(0,k)$ platí

$$\int_{\theta_{0.50}}^{\theta_{0.50}+\theta_{0.75}} f(u)\,du = \int_{0}^{\theta_{0.75}} \frac{1}{2\pi I_0(k)} \exp\{k\cos(u)\}\,du = \frac{1}{2\pi I_0(k)}\,\frac{\pi I_0(k)}{2} = \frac{1}{4},$$

kde $I_0(k)$ je modifikovaná Besselova funkce prvního druhu s $p = 0$ podle poznámky 5.

Velikost hodnoty kvantilu je ovlivněná parametrem míry koncentrace. Konkrétně pro tento příklad, vztah horního kvartilu $\theta_{0.75}$ a parametru k lze zobrazit pomoci grafu, viz obrázek 3.3.



Obrázek 3.3: Horní kvartil $\theta_{0.75}$ z $M_2(0,k)$ rozdělení, pro $k \in \{0, 10\}$.

3.4 CR_q a K-násobek pro směrový boxplot

Při konstruovaní směrového boxplotu nejobtížnější momenty výpočtu vyvolá hodnota směrového mezikvartilového rozpětí CR_q a její vztah k míře koncentrace k, plus odhad konstanty K-násobku. Protože směrový boxplot leží na kružnici, může vzniknout problém překrytí dolní a horní meze. Z tohoto důvodu lze obtížně posoudit existenci odlehlých hodnot a směrový boxplot nebude poskytovat přehledný obraz grafu.

Vztah CR_q a míry koncentrace k

Mezikvartilové rozpětí CR_q lze odhadnout pomocí distribuční funkce libovolného rozdělení a definovaný odhad bude vztah $CR_q = x_{0.75} - x_{0.25}$, kde $x_{0.75}$ a $x_{0.25}$ jsou kvartily tohoto rozdělení. Ale v případě, kdy neznáme distribuční

funkce nebo během vyšetřování dat, neparametrické metody budou užitečné pro výpočet mediánu a směrového mezikvartilového rozpětí. Přesto je těžké najít podobnou funkci pro CR_q, protože neexistuje takové normované von Misesovo rozdělení, které by bylo analogické k normovanému normálnímu rozdělení. Vztah mezi CR_q a k lze odvodit s pomocí odhadu z von Misesova rozdělení pro velkou hodnotu míry koncentrace k.

Věta 4. *Nechť* $k \to \infty$,

$$\beta = \sqrt{k}(\theta - \mu) \overset{d}{\Rightarrow} N(0,1),$$

kde $0 < \theta \leq 2\pi$ *a* $0 \leq \mu < 2\pi$.

Důkaz. Podrobný důkaz je popsán v [Jammalamadaka a Sengupta, 2001, věta 2.2].

❑

Důsledek 3. *Pro dost velké* k *můžeme* θ *aproximovat normálním rozdělením se střední hodnotou* μ *a rozptylem* $1/k$.

Podle [Fox, 1997] pro náhodný výběr z $N(\mu,\sigma^2)$ rozdělení mezikvartilové rozpětí R_q lze odhadnout pomocí hodnoty 1.349σ. To znamená, že pro velkou hodnotu k platí

$$CR_q \doteq 1.349/\sqrt{k}, \tag{3.9}$$

viz [Dempster a kol., 2011] .

Za předpokladu, že θ je náhodná veličina z $M_2(\mu, k)$ rozdělení, pak podle věty 4 ji lze odhadnout z $N(\mu, 1/k)$ rozdělení a $CR_q \doteq R_q$. Protože směrové mezikvartilové rozpětí se přibližně rovná lineárnímu mezikvartilovému rozpětí, lze odvodit následující vztahy

$$CR_q \doteq R_q = x_{0.75} - x_{0.25} = \mu + u_{0.75}\,\sigma - \mu + u_{0.25}\,\sigma = \sigma(u_{0.75} + u_{0.25}) = 1.349\,\sigma,$$

tj.

$$CR_q \doteq 1.349\,\sigma. \tag{3.10}$$

Pro dost velké k platí z věty 4 $\sigma^2 = 1/k$, po úpravě $\sigma = 1/\sqrt{k}$. Pokud dosadíme $\sigma = 1/\sqrt{k}$ do (3.10), dostaneme (3.9).

Z odvození je jasné, že parametr normálního rozdělení σ závisí na míře koncentrace k. Navíc je zřejmý vztah von Misesova a normálního rozdělení, jejich parametrů, kvantilů a mezikvartilového rozpětí. Z tohoto důvodu se vztah mezi těmito rozděleními projeví i v zobrazení boxplotu.

K-násobek CR_q

Při konstruování boxplotu pro jednorozměrný náhodný výběr se obyčejně používá K-násobek rovný 1.5 nebo 3. Ale pro vícerozměrná data není rozumné používat stejný násobek, protože kvůli omezenosti rozsahu kruhu existuje možnost překrytí dolní a horní meze. Tato situace vetšinou nastává při práci s velkou hodnotou K-násobku a malou hodnotou míry koncentrace k. Existují různé možnosti odhadu parametru K, některé jsou popsané v [Hoaglin a kol., 1986].

Věta 5. *Pro velkou hodnotu míry koncentrace k a náhodný výběr z $M_2(\mu, k)$ rozdělení, velikosti $n \geq 10$ platí, že dolní a horní mez směrového boxplotu se budou překrývat, pokud*

$$K > \pi\sqrt{k}/1.349 - 0.5, \tag{3.11}$$

kde K je K-násobek směrového mezikvartilového rozpětí.

Důkaz. K problému překrývání pro velký výběr n a velké k dochází v momentě, když $\tilde{\theta}_{0.25}$+K$*CR_q > \pi$. Pro symetrický výběr či medián je 0, $\tilde{\theta}_{0.25} = CR_q/2$. Proto

$$(0.5\text{+K}) * CR_q > \pi.$$

Podle (3.9) $CR_q \doteq 1.349/\sqrt{k}$, pak

$$1.349(0.5\text{+K})/\sqrt{k} > \pi.$$

Po úpravě dostaneme, že překrytí dolní a horní meze nastane právě, když

$$\text{K} > \pi\sqrt{k}/1.349 - 0.5.$$

❑

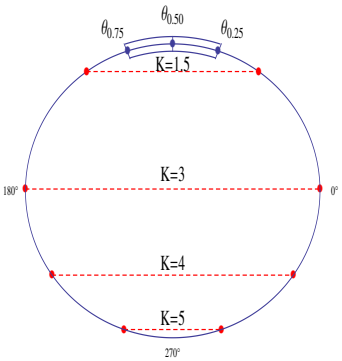
Příklad 6. Ukažme na příkladě teoretického směrového krabicového grafu pro $M_2(\pi/2, 5)$, jak velikost konstanty K ovlivňuje hodnotu dolní a horní meze, čímž evokuje možnost jejich překrytí, viz obrázek 3.4.

Spočítáme hodnotu K-násobku podle věty 5

$$K < \pi\sqrt{5}/1.359 - 0.5 = 4.7,$$

tj. $\text{K}_{max} = 4.7$, kde K_{max} je největší možná dosažitelná hodnota.

S rostoucím K se zvyšuje pravděpodobnost překrytí. Přičemž k překrytí mezí dochází dost často bez ohledu na to, že jsou to teoretické hodnoty von Misesova rozdělení.



Obrázek 3.4: Teoretický směrový boxplot pro $M_2(\pi/2, 5)$ rozdělení pro různé konstanty K-násobky K= $\{1.5, 3, 4, 5\}$

Volba konstanty K podle doporučení [Abuzaid a kol., 2012] se dělá ne jenom z odhadu K, ale ještě podle velikosti míry koncentrace k.

Poznámka 6. Pro velký parametr k, resp. $k > 3$, von Misesova rozdělení se používá konstanta K z intervalu $2.0 < K < 2.7$, pro malé k $1.0 < K < 2.0$.

3.5 Výběrové kvantily směrových dat

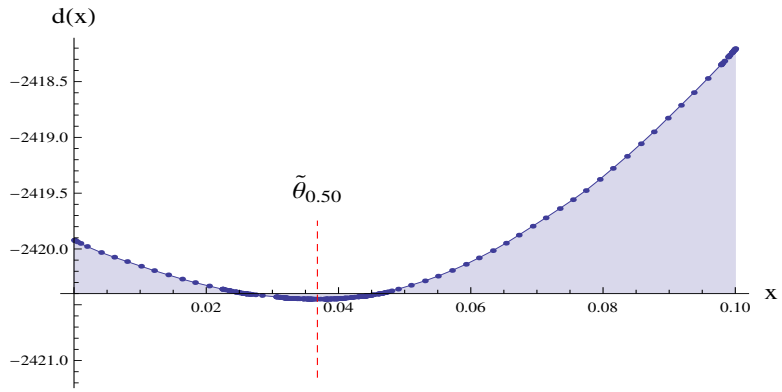
Výběrové kvantily

Medián směrových dat je hodnota, která dělí pozorovaný soubor dat na dvě stejné skupiny a je braná jako pozorovaní, které se značí $\tilde{\theta}_{0.50}$ a je to hodnota minimalizující součet obloukových vzdáleností, tj.

$$\tilde{\theta}_{0.50} = \underset{x\in(0,2\pi)}{argmin}\ d(x) = \underset{x\in(0,2\pi)}{argmin}\left(\pi - \sum_{i=1}^{n} |\pi - |\theta_i - x||\right), \tag{3.12}$$

kde θ_i jsou jednotlivá pozorování pro $i = 1, \ldots, n$ [Fisher, 1996].

Příklad 7. Výpočet směrového mediánu podle (3.12) ukažme pomocí obrázku 3.5 pro simulovaná data z $M_2(0, 1.8)$ rozdělení a $n = 1000$, kde medián směrových dat je pozorování $\tilde{\theta}_{0.50}$, které je zobrazeno jako červená přerušovací čára minimalizující hodnotu funkce $d(x)$, konkrétně pro tento obrázek 3.5 pro $x \in$ (0,0.1). Směrový medián se rovná $\tilde{\theta}_{0.50} = 0.037$.



Obrázek 3.5: Graf funkce $d(x)$ pro $x \in$ (0,0.1) data simulovaná z rozdělení $M_2(0, 1.8), n = 1000$, kde medián směrových dat je červená přerušovací čára.

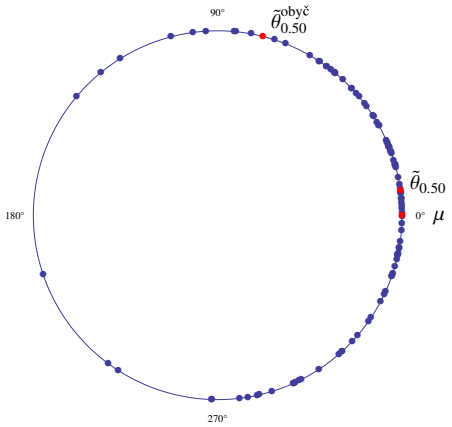
Poznámka 7. Jednou z důležitých vlastností směrového mediánu je to, že pokud zvětšíme (resp. zmenšíme) hodnotu ze základního souboru dat o libovolnou konstantu c, pak se hodnota mediánu také zvětší (resp. zmenší) o tuto konstantu.

Odvození obsahu poznámky 7 podle [Abuzaid a kol., 2012] bude vypadat takto

$$d\left(\tilde{\theta}_{0.50} + c\right) = \pi - \sum_{i=1}^{n} |\pi - |\theta_i + c - \tilde{\theta}_{0.50} - c|| = \pi - \sum_{i=1}^{n} |\pi - |\theta_i - \tilde{\theta}_{0.50}|| = d\left(\tilde{\theta}_{0.50}\right).$$

Hodnoty dolního a horního kvartilu je možné odhadnout pomocí rozdělení základního souboru směrových dat s pomocí výběrového směrového mediánu do dvou skupin, přičemž dolní kvartil $\tilde{\theta}_{0.25}$ bude považován za výběrový medián první skupiny a horní kvartil $\tilde{\theta}_{0.75}$ za výběrový medián druhé skupiny. Pokud $\tilde{\theta}_{0.25}$ bude větší než $\tilde{\theta}_{0.75}$, jednoduše je vyměníme nebo použijeme ekvivalenci směrových hodnot vůči rotaci. Pro výpočet hodnot $\tilde{\theta}_{0.25}$ a $\tilde{\theta}_{0.75}$ budeme nadále předpokládat platnost vztahu $\tilde{\theta}_{0.25} - \tilde{\theta}_{0.50} \in [\pi, 2\pi]$ a $\tilde{\theta}_{0.75} - \tilde{\theta}_{0.50} \in [0, \pi]$.

Rozdíl mezi směrovým a obyčejným mediánem je dobře vidět na obrázku 3.6 pro simulovaná data z rozdělení $M_2(0, 1.8)$, a pro $n = 100$. Směrový medián $\tilde{\theta}_{0.50}$ je skoro odpovídá teoretickému mediánu μ, viz obrázek 3.6, když obyčejný výběrový medián $\tilde{\theta}^{\text{obyč}}_{0.50}$ udává úplně špatný výsledek.



Obrázek 3.6: Porovnání směrového $\tilde{\theta}_{0.50}$, obyčejného $\tilde{\theta}^{\text{obyč}}_{0.50}$ výběrových mediánů a teoretického μ pro rozdělení $M_2(0, 1.8), n = 100$.

Směrové mezikvartilové rozpětí

Stejně jako v lineární teorii pravděpodobnosti, tak i v cirkulární existuje pojem mezikvartilové rozpětí, jinak řečeno směrové mezikvartilové rozpětí CR_q, které se používá při konstruování směrového boxplotu. Mezikvartilové rozpětí $\tilde{CR}_q$ směrových výběrových hodnot se počítá stejně jako rozpětí $\tilde{R}_q$ lineárních výběrových hodnot

$$\tilde{CR}_q = \tilde{\theta}_{0.75} - \tilde{\theta}_{0.25}. \tag{3.13}$$

Ale v případě rotace výběrových kvantilů $\tilde{\theta}_{0.25}$ a $\tilde{\theta}_{0.75}$ lze $\tilde{CR}_q$ spočítat jako

$$\tilde{CR}_q = 360° - \tilde{\theta}_{0.75} + \tilde{\theta}_{0.25}. \tag{3.14}$$

Navíc, dolní mez se rovná $\tilde{\theta}_{0.25}-\text{K}*\tilde{CR}_q$ a horní mez $\tilde{\theta}_{0.75}+\text{K}*\tilde{CR}_q$, kde K je násobek směrového mezikvartilového rozpětí, které závisí na koncentraci dat v souboru. V případě, kdy dojde k rotaci $\tilde{\theta}_{0.25}$ a $\tilde{\theta}_{0.75}$, dolní mez je $\tilde{\theta}_{0.25}+\text{K}*\tilde{CR}_q$ a horní mez je $\tilde{\theta}_{0.75}-\text{K}*\tilde{CR}_q$. Při velké koncentraci směrové kvantily a střední hodnota mohou ležet ve stejném bodě a to kvůli tomu, že $\tilde{CR}_q = 0$.

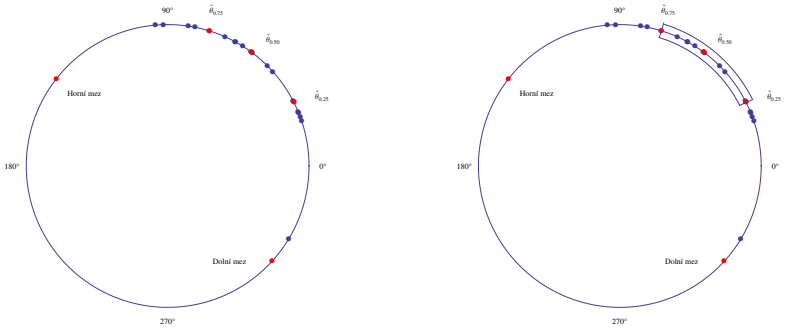
Příklad 8. Sestrojíme kruhový krabicový graf pro náhodný výběr z Von Misesova rozdělení s parametry $\mu = \pi/4$ a $k = 4$ o rozsahu $n = 20$, viz obrázek 3.7.

Výběrový medián $\tilde{\theta}_{0.50}$ pro tento výběr se rovná $\tilde{\theta}_{0.50} = 53.5°$, dolní kvartil je $\tilde{\theta}_{0.25} = 26.7°$ a horní kvartil je $\tilde{\theta}_{0.75} = 72.8°$. Podle (3.14) vypočítáme cirkulární mezikvantilové rozpětí

$$\tilde{CR}_q = \tilde{\theta}_{0.75} - \tilde{\theta}_{0.25} = 72.8° - 26.7° = 46.1°.$$

Dopočítáme dolní a horní mez pro klasickou variantu, kde K= 1.5, tj. dolní mez= $\tilde{\theta}_{0.25} - 1.5 * \tilde{CR}_q = 317.5°$ a horní mez= $\tilde{\theta}_{0.75} + 1.5 * \tilde{CR}_q = 142°$.

Na závěr této kapitoly na základě dat otvíracích dob jednadvaceti burz cenných papírů, sestrojíme časový směrový boxplot.



Obrázek 3.7: Struktura cirkulárního boxplotu pro náhodný výběr z $M_2(\pi/4, 4)$ rozdělení, $n = 20$.

Příklad 9. Otvírací doby burz cenných papírů podle světového času (UTC) jsou

14:30, 14:30, 0:00, 8:00, 8:00, 1:15, 1:30, 14:30, 7:00, 23:50, 3:45,
3:45, 8:00, 13:00, 0:00, 1:30, 8:00, 7:00, 6:00, 1:00, 1:00

Odhadneme rozdělení dat a příslušné parametry. Protože čas si lze představovat jako body na jednotlivé kružnici v časové soustavě souřadnic, nejlepší bude použití odhadu pro von Misesovo rozdělení či lze parametry jednoduše odhadnout pomocí libovolného softwaru, například *MATHEMATICA*. Pro práci se souborem vstupních dat je z časových hodnot převedeme na desetinné čísla a pak na radiány, podle vzorců

$$\text{radián}= \left(\frac{\check{c}as}{60} * \frac{360^\circ}{24} \right) * \frac{\pi}{180^\circ}.$$

Odhadnuté rozdělení je $M_2(1.19496, 0.837109)$, kde střední časová hodnota $\widehat{\mu} = 1.19496$, tj. 4:34 a odhad míry koncentrace $\widehat{k} = 0.837109$. Časový medián vstupních dat je $\tilde{\theta}_{0.50} = 6$, tj. 6:00 hodin. Dolní kvartil $\tilde{\theta}_{0.25}$ se rovná $\tilde{\theta}_{0.25} = 1{:}30$, tj. 1 hodina a 30 minut a horní kvartil $\tilde{\theta}_{0.75}$ se rovná 8 hodin. Mezikvartilové rozpětí $\tilde{CR}_q$ odvodíme jako rozdíl dolního a horního kvartilu

$$\tilde{CR}_q = \tilde{\theta}_{0.75} - \tilde{\theta}_{0.25} = 8{:}00 - 1{:}30 = 6{:}30.$$

Z vypočtených údajů již lze sestrojit krabici časového boxplotu, ale hlavní problém se týká konstanty K, která je potřebná pro výpočet dolní a horní meze. Podle věty 5 odhadneme K-násobek,

$$\begin{aligned} \text{K} &< \pi/\tilde{CR}_q - 0.5, \\ \text{K} &< 1.4. \end{aligned}$$

Jinak řečeno, $\text{K}_{max} = 1.4$ je největší možná hodnota, kterou má smysl volit.

Kvůli tomu, že konstantu K můžeme volit libovolně tak, aby patřila do intervalu $K \in (0, 1.4)$, pomocí tabulky 3.1 porovnáme hodnoty dolní a horní meze pro různé násobky.

Při konstruování boxplotu bychom měli volit $\text{K}= \{0.2, 0.5, 1.0, 1.2, 1.3\}$, protože pro daný odhad parametru $\widehat{k}$ a daný rozsah výběru n nedojde k překrytí

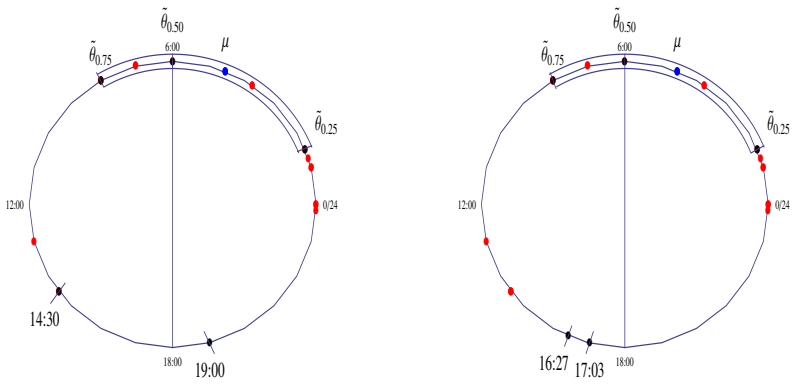
K	Dolní mez	Horní mez	Počet odlehlých hodnot
0.2	0:12	9:18	7
0.5	22:15	11:15	4
1.0	19:00	14:30	0
1.2	17:42	15:48	0
1.3	17:03	16:27	0
K_{max}	16:45	16:45	0
2.0	12:30	21:00	0

Tabulka 3.1: Dolní a horní meze pro různé K-násobky pro otvírací časy burz cenných papírů

meze. Avšak čím menší násobek zvolíme, tím více bude odlehlých hodnot. Proto podle poznámky 6 parametr $\widehat{k}$ není velká hodnota a nejlepší volba K-násobku je K= $\{1.0, 1.2, 1.3\}$. Právě v těchto případech nedojde k překrytí a počet odlehlých hodnot je nulový.

V bodě K_{max} dolní a horní mez mají stejný čas a to je kvůli tomu, že právě v tomto bodě dochází k částečnému překrytí.

Pro časový boxplot z dat otvírací doby jednadvaceti burz cenných papírů použijeme dvou různých násobků $K_1 = 1.0$ a $K_2 = 1.3$, viz obrázek 3.8.



Obrázek 3.8: Časový boxplot z dat otvírací doby burz cenných papírů pro $K_1 = 1.0$ a $K_2 = 1.3$

Kapitola 4

Vícerozměrný směrový krabicový graf

Pro vícerozměrná směrová data se krabicový graf změní. Při práci s mnohorozměrnými daty se používá místo kvantilů (které jsou definovány pouze pro jednorozměrné náhodné veličiny) pojem hloubky. Klíčovým pojmem bude poloprostorová hloubka, na které se vzhledem k vícerozměrnému datovému souboru nachází body.

Na základě tohoto pojetí lze odvodit dvoudimenzionální boxplot, známý jako bagplot. Pro práci s vícerozměrnými směrovými daty popišme třírozměrné Fisherovo rozdělení, které je speciálním případem von Mises-Fisherova rozdělení.

4.1 Bagplot

Vícerozměrný směrový krabicový graf neboli bagplot na rozdíl od boxplotů, které byly popsané v kapitole 1 a kapitole 3, bude mít jiný tvar konstrukce. Bagplot sestrojíme ze třech částí: *bag*, *fence*, *loop*. Bagplot zobecňuje boxplot na dvoudimenzionálním prostoru následujícím způsobem, viz tabulka 4.1.

Boxplot	**Bagplot**
Krabice	Bag
Medián	Hloubkový medián
Maximální „vous“	Fence
„vousy“	Loop
Odlehlé hodnoty	Odlehlé hodnoty

Tabulka 4.1: Porovnání boxplotu a bagplotu

Polygon *bag* obsahuje 50 % bodů a na grafu vypadá jako tmavý prostor, *fence* je mez, která odděluje hodnoty nacházející se uvnitř polygonů od odlehlých hodnot. Část *loop* označuje body, které leží mezi částí bag a fence, na grafu se značí jako trochu světlejší prostor na rozdíl od bag. Kromě těchto částí lze

v bagplotu vydělit hloubkový medián, což je bod s nejvyšší poloprostorovou hloubkou. Medián se nachází uprostřed a obyčejně se značí jako křížek.

Pro velmi „ploché“ vícerozměrné hodnoty se bagplot stává boxplotem. Přičemž světlejší prostor částí loop odpovídá „vousům“ krabicového grafu.

Stejně jako boxplot, bagplot zviditelní vlastnosti dat. Například hloubkový medián, rozměr bagu, korelaci (orientace bagu), šikmost (tvar bagu a loopu) a chvosty (body za hranice loopu a odlehlých hodnot).

4.2 Třírozměrné Fisherovo rozdělení

Fisherovo rozdělení stejně jako von Misesovo pochází z p-dimenzionálního von Mises-Fisherovo rozdělení. Pro $p = 3$ von Mises-Fisherovo rozdělení bude mít tvar Fisherova rozdělení $F(\boldsymbol{\mu}, k)$. Fisherovo rozdělení bylo poprvé použito v [Langevin, 1905], ale bylo pojmenované po Ronaldu Fisherovi, který podrobně studoval toto rozdělení včetně jeho parametrů střední hodnoty $\boldsymbol{\mu}$ a míry koncentrace k. Fisherovo rozdělení pro $p = 3$ a $k > 0$ má hustotu

$$f(\mathbf{x}; \boldsymbol{\mu}, k) = \frac{k}{4\pi \sinh k} \exp\{k\boldsymbol{\mu}'\mathbf{x}\}, \quad \mathbf{x} \in S^2. \tag{4.1}$$

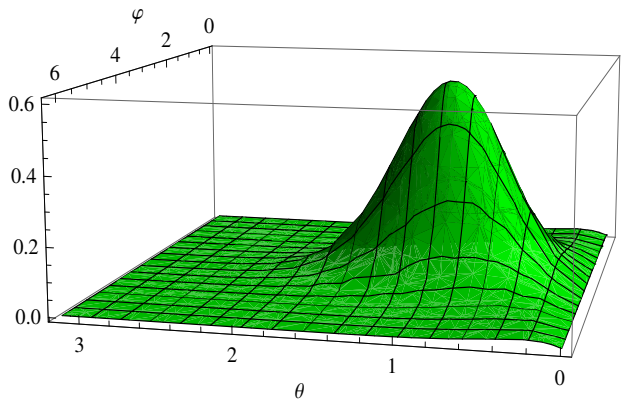
Na rozdíl od von Misesova rozdělení je hustota $F(\boldsymbol{\mu}, k)$ obyčejně definovaná pro náhodné sférické úhly, ne pro náhodné jednotkové vektory, viz obrázek 4.1.

Pro sférický náhodný vektor (Θ, Φ) a podle základní věty o transformaci [Klaška, 2006, věta 9.3-9.12] hustota Fisherova rozdělení bude mít tvar

$$g(\theta,\varphi; \alpha, \beta, k) = \frac{k}{4\pi \sinh k} \exp\{k\left(\cos\theta\cos\alpha + \sin\theta\sin\alpha\cos(\varphi - \beta)\right)\} \sin\theta, \tag{4.2}$$

kde $0 \leq \theta \leq \pi$ a $0 \leq \varphi \leq 2\pi$ a kvůli tomu, že $\boldsymbol{\mu}$ je vektor, pro $p = 3$ platí

$$\boldsymbol{\mu} = (\sin\alpha\cos\beta, \sin\alpha\sin\beta, \cos\alpha)'.$$



Obrázek 4.1: Hustota Fisherova rozdělení pro $\alpha = \pi/4$, $\beta = \pi$, $k = 5$

Hustota Fisherova rozdělení je zřejmě symetrická kolem parametru $\boldsymbol{\mu}$, můžeme tedy prohlásit, že $\boldsymbol{\mu}$ je vícerozměrný medián. Oblast obsahující polovinu

hustoty (4.1) je oddělena od jiných částí pomocí kvartilové křivky. Jinak řečeno, kvartilová křivka je hranice bagu u teoretického bagplotu.

Příklad 10. Jako ilustrace použijeme Fisherovo rozdělení $F(\boldsymbol{\mu}, k)$ pro $k = 1$ a spočítáme jeho kvartilovou křivku. S pomocí věty o transformaci [Klaška, 2006, věta 9.3-9.12] přivedeme hustotu (4.1) na tvar hustoty (4.2), kde $\alpha = 0, \beta = 0$. Při dosazení parametrů do hustoty (4.2) dostaneme

$$\begin{aligned} g(\theta, \varphi; 0, 0, 1) &= \frac{1}{4\pi \sinh 1} \exp\{1\,(\cos\theta \cos 0 + \sin\theta \sin 0 \cos(\varphi - 0))\} \sin\theta \\ &= \frac{1}{4\pi \sinh(1)} \exp\{\cos\theta\} \sin\theta, \end{aligned}$$

kde $0 \leq \theta \leq \pi$ a $0 \leq \varphi \leq 2\pi$.

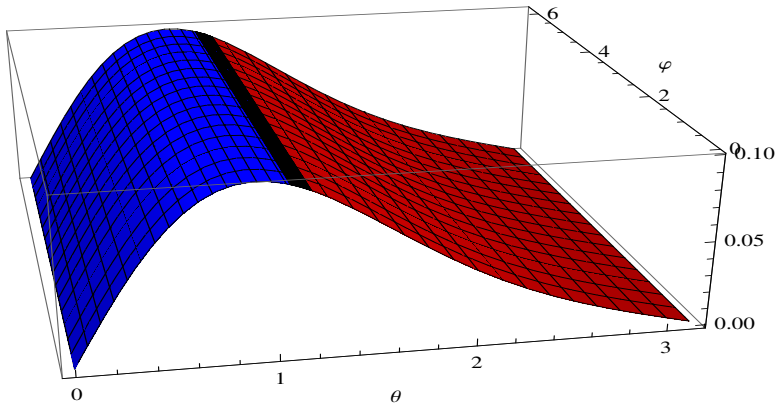
Problém výpočtu sférické kvartilové křivky převádíme na výpočet kvantilu marginálního rozdělení θ, tj.

$$\int_0^t \left[\int_0^{2\pi} g(\theta, \varphi; 0, 0, 1)\, d\varphi \right] d\theta = \int_0^t 2\pi g_\theta(\theta; 0, 0, 1)\, d\theta = \frac{1}{2},$$

kde t je konstanta, která určuje kvartilovou křivku.

Hlavní problém výpočtu kvartilů se týká odvození konstanty t, která závisí na míře koncentrace k. Konkrétně pro tento případ je kvartilová křivka určena $t = 1.122$.

Grafy 4.2 a 4.3 zobrazují hustotu (4.2) Fisherova rozdělení s parametry $\alpha = 0, \beta = 0$ a mírou koncentrace $k = 1$, kde černá oblast je kvartilová křivka (resp. hranice bagu), modrá barva odpovídá mezikvartilovému rozpětí, tj. bagu a červená je zbytek.

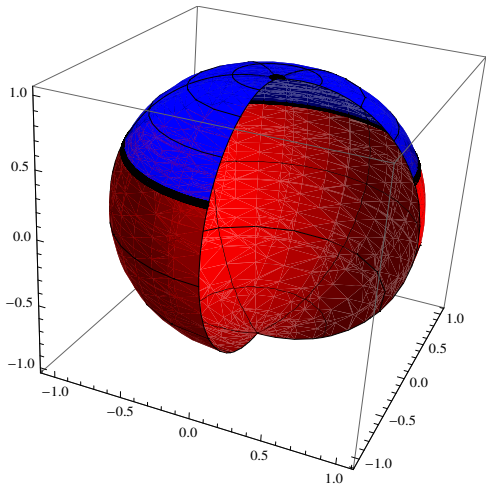


Obrázek 4.2: Hustota Fisherova rozdělení pro $\alpha = 0$, $\beta = 0$, $k = 1$

Konstrukce bagplotu

Existují různé způsoby konstrukce bagplotu a algoritmy výpočtu hloubkového mediánu, viz [Rousseeuw a kol., 1999].

Definice 7. *Hloubkové umístění* $ldepth(\theta, \boldsymbol{Z})$ *v poloprostoru některých bodů* $\theta \in \mathbb{R}^2$ *z vícerozměrných náhodných dat* $\boldsymbol{Z} = \{z_1, z_2, \ldots, z_n\}$ *je nejmenší počet* z_i*, který je obsažen v libovolném uzavřeném poloprostoru s hraniční čárou přes* θ*.*



Obrázek 4.3: Ilustrace hustoty $f(\mathbf{x}; \boldsymbol{\mu}, k)$ Fisherova rozdělení ve sférické soustavě souřadnic pro $\boldsymbol{\mu} = (0, 0, 1)$ a $k = 1$.

Poznámka 8. Hloubkový medián vícerozměrných dat $\mathbf{Z}$ je definován jako hodnota θ, $\theta \in \mathbb{R}^2$ z nejvyšší $ldepth(\theta, \mathbf{Z})$, pokud existuje jenom jedna taková θ. Jinak: medián je těžištěm definičního oboru hodnot. Algoritmus pro hloubkový medián je podrobněji popsán v [Rousseeuw a kol., 1998]

Na rozdíl od teoretického zobrazení náhodných veličin pomocí Fisherova rozdělení, pro výběrová sférická data je obtížné zjistit hodnotu hloubkového mediánu a kvartilů. Přehledný obraz skutečných dat lze představit pomocí bagplotu, například v prostředí *R* nebo *MATLAB*. Ale úprava směrových dat pro aplikace funkce bagplot() není jednoduchá. Podrobné algoritmy výpočtu mediánu jsou popsané v [Rousseeuw a kol., 1998], ale zobecnění na sférická data už není částí teto práce.

Kapitola 5

Závěr

Jak bylo řečeno v úvodu, práce popisuje možnosti zobrazení teoretických a výběrových hodnot pro jednorozměrné a pro vícerozměrné směrové rozdělení, konkrétně von Misesovo a Fisherovo rozdělení, pomocí různých druhů krabicového grafu, konkrétně pomocí obyčejného krabicového grafu, směrového a časového boxplotu a bagplotu.

Na základě kapitoly 1 bylo odvozeno, že krabicový graf se podobá grafu hustot teoretických rozdělení a obraz boxplotu je ovlivněn velikostí kvantilů, které závisí na parametrech rozdělení. Pro výběrová data byly popsány důležité pojmy, jako výběrové kvantily, mezikvartilové rozpětí, dolní a horní mez. Velmi vážný moment kapitoly 2 se týkal konvergence všech výběrových $\alpha-$kvantilů z dostatečně velkého výběru nezávislých náhodných veličin k teoretickým $\alpha-$kvantilům. Navíc bylo zjištěno, že výběrové kvantily mají asymptotické normální rozdělení s příslušnými parametry. V kapitole 3 jsme se seznámili směrovými daty, jejich vlastnostmi a pro ně typickým rozdělením, jako von Misesovo. Přitom byl odvozen blízký vztah von Misesova a normálního rozdělení včetně jejich parametrů. Pro konstrukci směrového boxplotu jsme nalezli způsob, jak odhadnout konstantu K tak, aby nedošlo k překrytí dolní a horní meze, což by mohlo poškodit obraz grafu. Sama konstrukce směrového boxplotu od obyčejného se liší tím, že data leží na jednotkovém kruhu a odvodit medián, například i směrové mezikvartilové rozpětí, těchto dat, není tak jednoduché. Ale v této práci se nám to podařilo. Přičemž pro konkrétní příklad jsme rozšířili pojem směrového boxplotu na časový boxplot, pomocí konvertování času na radián, a radiánu na čas. V poslední kapitole jsme se podívali na vícerozměrná data a na některé jejich důležité vlastnosti, včetně Fisherova rozdělení, jako reprezentanta tohoto druhu dat. Jako způsob zobrazení vícerozměrných dat byl zvolen bagplot a stručně byl popsán postup jeho konstrukce.

Tato práce je zaměřená na rozšíření znalostí o možnosti přehledného zobrazení směrových dat pomocí různých druhů boxplotu a je základem k pochopení některých statistických pojmů a jejich dalšího rozvoje. Práce obsahuje hodně příkladů, což ji dělá srozumitelnou pro libovolného čtenáře.

Literatura

Abramowitz, M. a Stegun, I. A. (1965). *Handbook of Mathematical Functions*. Dover, New York. ISBN 0-486-61272-4.

Abuzaid, A. H., Mohamed, I. B. a Hussin, A. G. (2012). Boxplot for circular variables. *Computational Statistics*, **27**, 381–392.

Dempster, A. P., Laird, N. M. a Rubin, D. B. (2011). Boxplot for circular variables. *Computational Statistics*, **39**(1), 1–38.

Fisher, N. I. (1996). *Statistical analysis of circular data*. Cambridge University Press, London. ISBN 9780521568906.

Fox, J. (1997). *Applied Regression Analysis, Linear Models, and Related Methods*. SAGE Publications, Singapore. ISBN 9780803945401.

Hoaglin, D., Iglewicz, B. a Tukey, J. (1986). Performance of some resistant rules for outlier labeling. *Journal of American Statistical Association*, **81**, 991–999.

Jammalamadaka, S. a Sengupta, A. (2001). *Topics in circular statistics*. World Scientific Press, Singapore. ISBN 9810237782.

Klaška, J. (2006). *Diferenciální a integrální počet funkcí více proměnných*. Učební text předmětu Matematika II, ÚM FSI, Brno.

Langevin, P. (1905). Magnétisme et théeorie des électrons. *Annales de chimie et de physique*, **5**, 41–127.

Malá, O. (2012). *Fisher-Bingham distribution*. bakalářská práce, MFF UK, Praha.

Rousseeuw, P. J., Ruts, I. a Tukey, J. (1998). Constructing the bivariate tukey median. *Statistica Sinica*, **8**, 827–839.

Rousseeuw, P. J., Ruts, I. a Tukey, J. (1999). The bagplot: a bivariate boxplot. *Journal of the American Statistical Association*, **53**(4), 382–387.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. xiv, 371 s. John Wiley Sons, Inc., New York: Wiley. ISBN 0-471-21927-4.

Tukey, J. W. (1978). Variations of box plots. *The American Statistician*, **32** (1), 12–16.

Seznam obrázků

Seznam tabulek