VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV TELEKOMUNIKACÍ

DEPARTMENT OF TELECOMMUNICATIONS

APLIKACE METOD DETEKCE ODLEHLÝCH HODNOT V OBLASTI OBJEKTIVNÍ ANALÝZY PARKINSONOVY NEMOCI

APPLICATION OF OUTLIERS DETECTION METHODS IN THE FIELD OF OBJECTIVE ANALYSIS OF PARKINSON'S DIASEASE

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE — Daniel Sadílek

AUTHOR

VEDOUCÍ PRÁCE — Ing. Zoltán Galáž

SUPERVISOR

BRNO 2017

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Teleinformatika**
Ústav telekomunikací

Student: Daniel Sadílek ***ID:*** 164387
Ročník: 3 ***Akademický rok:*** 2016/17

NÁZEV TÉMATU:

Aplikace metod detekce odlehlých hodnot v oblasti objektivní analýzy Parkinsonovy nemoci

POKYNY PRO VYPRACOVÁNÍ:

V rámci bakalářské práce budou nastudovány a implementovány metody detekce odlehlých hodnot. Dále bude vytvořeno grafické uživatelské rozhraní aplikace, která bude tyto metody využívat. Vybrané metody budou demostrovány na příkladech parametrů kvantifikujících poškození řeči pacientů s Parkinsonovou nemocí.

DOPORUČENÁ LITERATURA:

[1] BARNETT, Vic a Toby LEWIS. Outliers in statistical data. 3rd ed. Chichester: John Wiley & Sons, 1998. Wiley series in probability and mathematical statistics. ISBN 0-471-93094-6.

[2] HAWKINS, Douglas M. Identification of outliers. New York: Chapman and Hall, 1980. ISBN 041221900X.

Termín zadání: 1.2.2017 ***Termín odevzdání:*** 8.6.2017

Vedoucí práce: Ing. Zoltán Galáž
Konzultant:

doc. Ing. Jiří Mišurec, CSc.
předseda oborové rady

ABSTRAKT

Bakalářská práce „Aplikace metod detekce odlehlých hodnot v oblasti objektivní analýzy Parkinsonovy nemoci“ se zabývá detekcí odlehlých hodnot v datových souborech pacientů s Parkinsonovou nemocí, která je velmi důležitá pro další zpracování dat, kde by jinak docházelo ke zkreslení a znehodnocení výsledků. Vybrané metody byly nastudovány, implementovány a odzkoušeny v programovém prostředí MATLAB, společně s vytvořením grafického rozhraní.

KLÍČOVÁ SLOVA

Analýza, detekce, histogram, jádrový odhad hustoty pravděpodobnosti, krabicový graf, odlehlá hodnota, Parkinsonova nemoc.

ABSTRACT

The bachelor thesis „Application of outliers detection methods in the field of objective analysis of Parkinson's disease“ deals with the detection of outliers in the files of patients with Parkinson disease, which are essential in further data processing, where otherwise distortion and debasement of data could occur. The selected methods were studied, implemented and tested in the MATLAB software with creating graphical user interface.

KEYWORDS

Analysis, detection, histogram, kernel density estimation, box plot, outlier, Parkinson‘s disease.

SADÍLEK, D. *APLIKACE METOD DETEKCE ODLEHLÝCH HODNOT V OBLASTI OBJEKTIVNÍ ANALÝZY PARKINSONOVY NEMOCI.* BRNO: VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ, FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ, 2017. 32 S. VEDOUCÍ BAKALÁŘSKÉ PRÁCE ING. ZOLTÁN GALÁŽ.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma „Aplikace metod detekce odlehlých hodnot v oblasti objektivní analýzy Parkinsonovy nemoci“ jsem vypracoval samostatně pod vedením vedoucího bakalářské práce, s použitím odborné literatury a dalších informačních zdrojů, které jsou citovány v práci a uvedeny v seznamu literatury na konci práce.

Zároveň jako autor bakalářské práce prohlašuji, že v souvislosti s vytvořením této bakalářské práce nebyla porušena autorská práva třetích osob, zejména nebylo nedovoleným způsobem zasaženo do cizích autorských práv osobnostních, nebo majetkových. Jsem si plně vědom následků porušení ustanovení §11 a následujících autorského zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

(podpis autora)

PODĚKOVÁNÍ

Tímto bych chtěl poděkovat vedoucímu práce panu Ing. Zoltánu Galážovi za odborné vedení, poskytnutí důležitých dokumentů a užitečné rady týkající se dané problematiky.

V Brně dne ....

(podpis autora)

OBSAH

SEZNAM OBRÁZKŮ

SEZNAM TABULEK

ÚVOD

V dnešní době je potřeba zpracovávat stále větší objemy dat a tím získavat důležité informace. Je potřeba, aby tato data byla co nejpřesnější a nevyskytovaly se v nich chyby a odchylky, zapříčiněné náhodnými jevy nebo lidským faktorem. V oblasti zdravotnictví je na tyto požadavky kladen obzvláště velký důraz, protože by jinak mohlo dojít k chybnému, či nepřesnému zpracování dat a tím i ohrození pacienta. Proto je nutné prověřit a posoudit kvalitu naměřených dat.

Kvalitou naměřených dat, se také myslí detekce odlehlých hodnot, jenž se mohou vyskytovat v reálných datových souborech. Nalezená odlehlá pozorování jsou důsledkem náhodných chyb, nedodržení zásad měření, rušení okolí nebo lidských chyb. Ne vždy, jsou tyhle hodnoty způsobené chybou, ale mohou pocházet z jiného rozdělení než normálního, či se může jednat o zásadní extrémní hodnoty pro další analýzu dat. Je ale potřeba zdůraznit, že označit hodnotu za odlehlou a případně ji vyřadit ze souboru, aby neovlivnila další analýzu, je zapotřebí jistých zkušeností v daném oboru a nespoléhat se pouze na výsledky detekčních testů.

Tato bakalářská práce slouží k seznámení se s vybranými metodami pro detekci odlehlých pozorování a následně se v druhé části bude věnovat implementaci v programovém prostředí MATLAB, včetně grafického uživatelského rozhraní, kde také proběhne odzkoušení na reálných jednorozměrných datech pacientů s Parkinsonovou nemocí, pocházející z databáze PARCZ.

1 DETEKCE ODLEHLÝCH POZOROVÁNÍ

Pod pojem odlehlá pozorování ([1], [7], [16]) neboli často používaný anglický výraz „outliers“, se rozumí hodnoty, které se nápadně liší od většiny ostatních ze zkoumaného souboru a mohou tím ovlivnit výslednou kvalitu příslušných statistických analýz.

Je velmi obtížné definovat pojem „hodnoty, které se nápadně liší“, protože je velice složité rozhodnout, zda odlehlé pozorování je chybou nebo pouze extrémní hodnotou, nesoucí důležitou informaci, která do zkoumaného souboru dat patří a její vyřazení by mohlo ovlivnit pozdější analýzu dat. Naopak ponechání chybné hodnoty způsobí kontaminaci datového souboru šumem. Je tedy pouze na analytikovi, jestli pozorování označí jako chybné či extrémní a jakou metodu k detekci použije.

Mimo náhodné chyby dochází k chybám zapříčiněných lidským faktorem, jako třeba porušení podmínek, za kterých mělo měření probíhat, nesprávným měřením, špatně nastavenými měřidly nebo také selháním měřící techniky

1.1 Grafiké metody

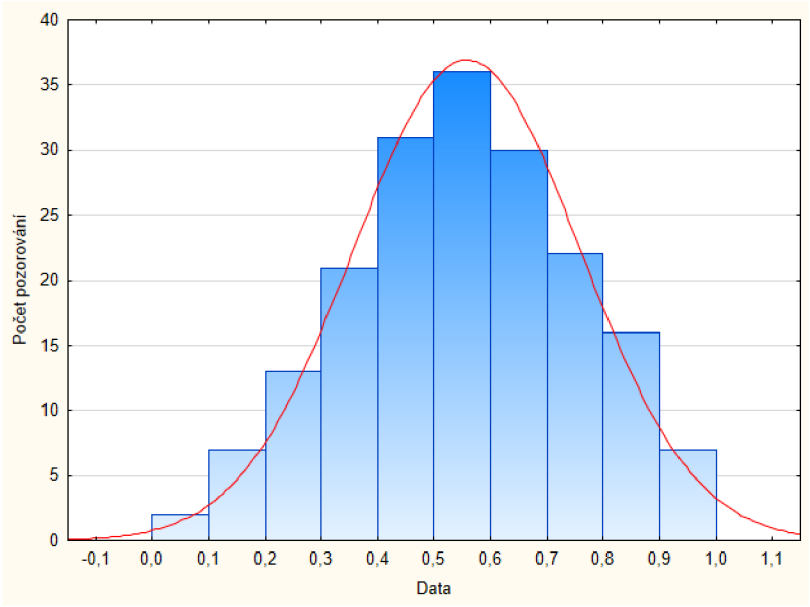
V této práci budou požity jednorozměrné metody, jelikož budou aplikovány na jednorozměrná data. Tyto metody lze rozdělit na parametrické a neparametrické. V případě, že je předem známé statistické rozdělení, využívá se parametrických metod, které většinou předpokládají normální neboli Gaussovo rozdělení.

1.1.1 Pravidlo 3-sigma

Na začátku je potřeba zmínit histogram, ze kterého se bude často vycházet. Histogram, sloupcový graf neboli graf četností (Obr. 1.1), slouží ke grafickému znázornění zkoumaných dat. Osa x je rozdělena na několik dělících intervalů (tříd) a osa y udává počet neboli četnost hodnot, které do těchto intervalů spadají. Důležité je správně určit počet tříd a jejich šířku, jinak by mohlo dojít k vykreslení nic neříkajícího grafu. K tomu slouží tzv. Sturgesovo pravidlo:

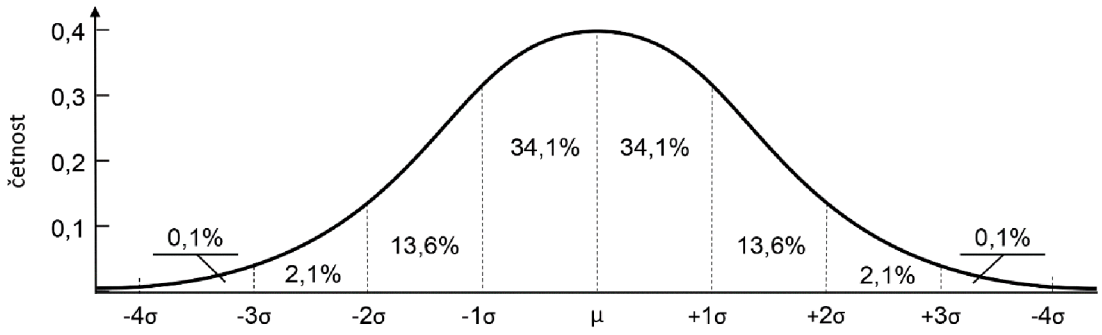
$$k = 1 + 3{,}3\log_{10}(n), \tag{1.1}$$

kde k je počet intervalů a n je počet pozorování.



Obr. 1.1: Histogram s Gaussovou křivkou.

Při normálním rozdělení se často používá pravidlo 3-sigma (3σ), které udává, že ve vzdálenosti 3 směrodatných odchylek od průměru (μ – 3σ, μ + 3σ) se nachází 99,73% hodnot ze zkoumaného souboru dat ([15]). Hodnoty ležící za hranicí 3σ mohou být považovány za odlehlé hodnoty. Grafické znázornění směrodatných odchylek je zobrazeno na Obr. 1.2.



Obr. 1.2: Zobrazení pravidla 3-sigma na Gaussově křivce ([9]).

1.1.2 Box-plot

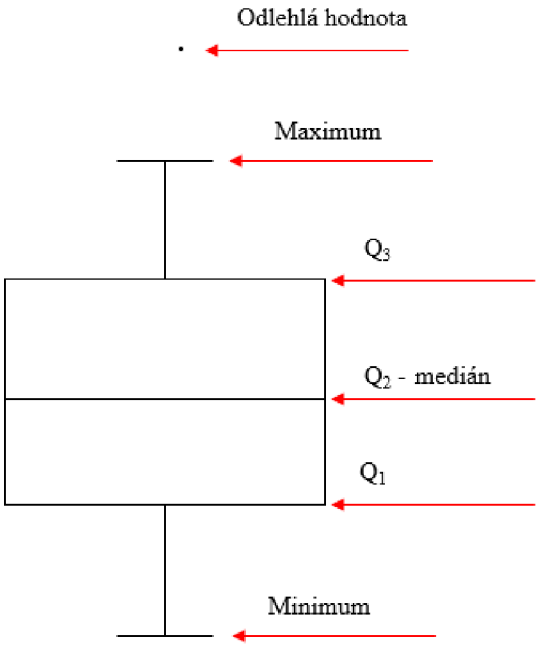
Box-plot ([1], [2], [7]), zvaný krabicový diagram (viz Obr. 1.3) zobrazující graficky soubor dat pomocí jeho kvartilů. Kvartil je typ kvantilu. Jsou to tři body, rozdělující seřazená data podle počtu na čtyři stejně velké části. Střední část box-plotu je, jak už název napovídá, složena z „krabice“ nebo chceme-li boxu. Box-plot může být zobrazen horizontálně nebo vertikálně.

Budeme-li uvažovat vertikální zobrazení, tak spodní hranu krabicové části nazýváme 1. kvartilem (Q_1, x_{25}), což znamená, že 25 % hodnot je nižších, než hodnota toho dolního kvartilu a zbylých 75% je větších nebo rovno. Horní hrana krabicové části se nazývá 3. kvartil (Q_3, x_{75}), který udává, že 75 % hodnot leží pod tímto kvartilem a 25 % je větších nebo rovno. Samotnou „krabici“ rozděluje 50 % kvartil, který, jak už vyplývá z označení, rozděluje setříděná data na dvě poloviny a je nazýván mediánem (Q_2, x_{50}).

Co se týče takzvaných vousů vycházejících z konců krabice, je více možností, které délka těchto čar představuje. Jednou z možností je označení minimální a maximální hodnoty ze souboru dat. Druhou a námi využívanou metodou zobrazení vousů u krabicového diagramu bude zobrazení posledního maxima (prvního minima), které splňuje vzdálenost maximálně 1,5-násobku IQR neboli mezikvartilového rozpětí od krajních kvartilů. IQR má velikost rozdílu 3. a 1. kvartilu (Q_3-Q_1).

Délka výše zmíněných vousů tedy bude sahat do maxim, potažmo minim splňujících podmínku Q_1 – 1,5xIQR a Q_3 + 1,5xIQR. Body vykreslené nad těmito hradbami, nazývané také jako vnitřní hradby, se považují za námi hledané odlehlé hodnoty.

Někdy se objevují i takzvané vnější hradby, které jsou vzdáleny od 1. a 3. kvartilu Q_1 –3xIQR a Q_3 + 3xIQR. Hodnoty mezi vnitřními a vnějšími hradbami jsou odlehlá pozorování a za vnějšími hradbami se prezentují jako extrémní hodnoty.

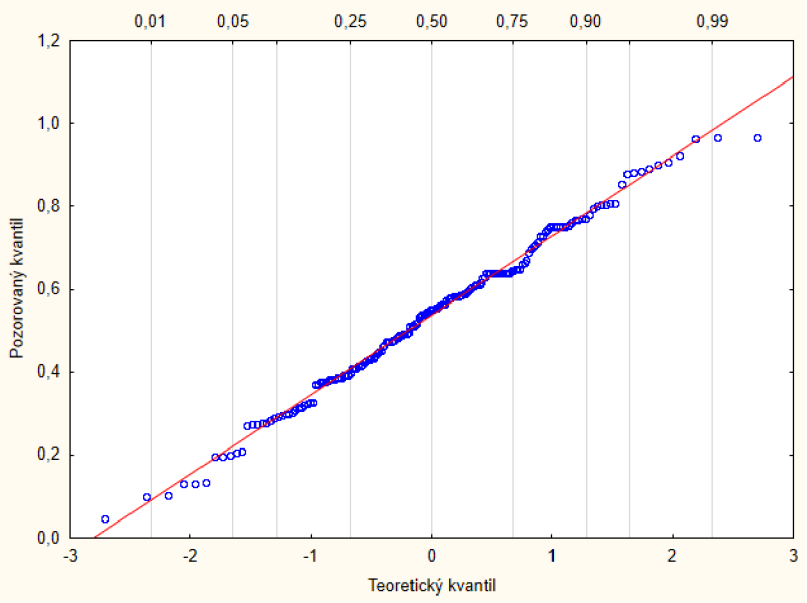


Obr. 1.3: Popis diagramu box-plot

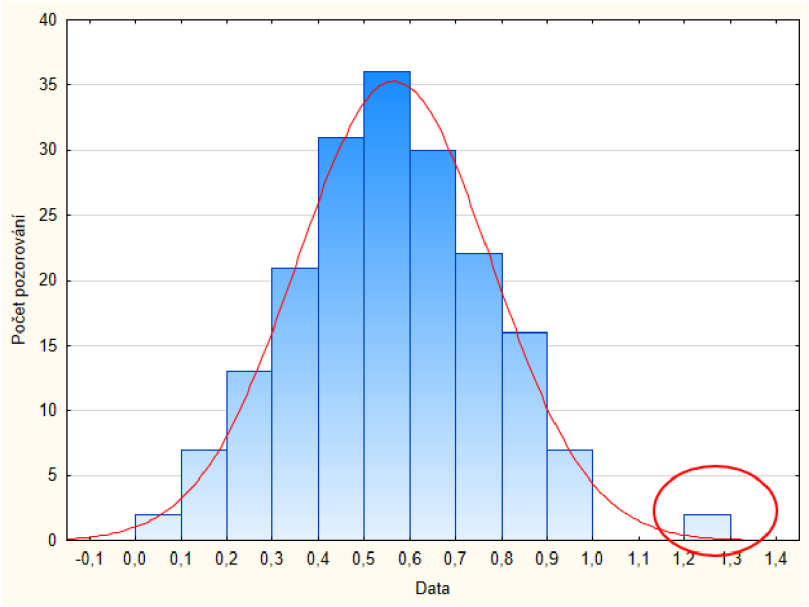
1.1.3 Q-Q graf

Tato grafická metoda neboli kvantil-kvantil graf ([12], [13]), neslouží přímo k detekci odlehlých hodnot, nýbrž k ověření, z jakého pravděpodobnostního rozdělení data pochází, ale z grafu lze také odhadnout, které naměřené hodnoty lze považovat za odlehlá pozorování. Nejjednodušší cesta pro posouzení, z jakého rozdělení data jsou, je zajisté histogram, ale u něj je potřeba určit vhodný počet dělících intervalů. Pokud je špatně určen počet dělení, nemůžeme zhodnotit výsledek, zda data opravdu patří do určitého rozdělení.

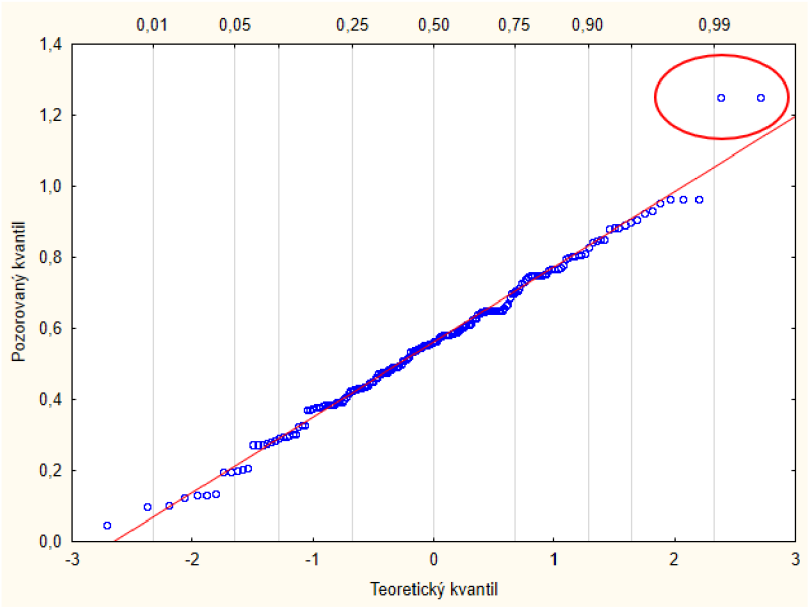
A proto je efektivnější využít grafických metod, které jsou založeny na principu porovnávání kvantilů teoretického rozdělení (nejčastěji normálního) s kvantily ze zkoumaného datového souboru. Neboli porovnání naměřených hodnot s očekávanými, u normálního rozdělení se očekává zobrazení bodů co nejblíže k přímce, jako na Obr 1.4. Na Obr. 1.6 je patrné, že data pochází z normálního rozdělení, ale obsahují odlehlá pozorování, stejně jako histogram na Obr. 1.5. Z umístění bodů vůči přímce, je možné si udělat představu o rozdělení dat. Například z Obr. 1.7 je patrné, že data pochází z jiného, než normálního rozdělení.



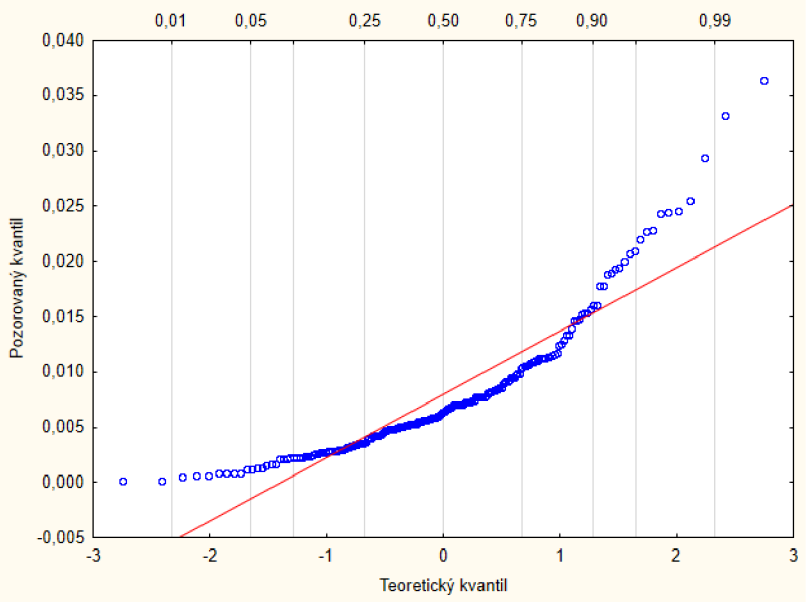
Obr. 1.4: Q-Q graf odpovídající normálnímu rozdělení



Obr. 1.5: Histogram odpovídající normálnímu rozdělení s odlehlými hodnotami



Obr. 1.6: Q-Q graf odpovídající normálnímu rozdělení s odlehlými hodnotami



Obr. 1.7: Q-Q graf neodpovídající normálnímu rozdělení

1.1.4 Jádrový odhad funkce hustoty pravděpodobnosti

KDE (kernel density estimation) neboli jádrový odhad hustoty pravděpodobnosti (viz Obr. 2.3) ([3], [6], [10]), patří mezi neparametrické metody a je úzce spojen s histogramem, ale disponuje vlastnostmi jako hladkost nebo spojitost. Není potřeba znát dopředu, z jakého rozdělení se bude hustota dohadovat. Jedná se o klouzavý vážený průměr, který je závislý na tvaru zvolené jádrové funkce, a především na šířce vyhlazovacího okna h, která určuje hladkost odhadu. Spojitý odhad hustoty pro n prvků je dán vztahem ([10])

$$\hat{f}_h(\mathrm{x}) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right), \tag{1.2}$$

kde K představuje jádro (kernel) a h je kladné číslo, které značí šířku vyhlazovacího okna. Velmi často používaným jádrem K je Gaussovo jádro, které je definováno jako ([10])

$$K_h(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2h^2}}. \tag{1.3}$$

Co se týče vyhlazovacího parametru h, tak si vystačíme s odhadem „pokus-omyl“, protože existuje několik metod pro výpočet tohoto parametru, které mohou být výpočetně náročné a zároveň je tohle téma natolik obsáhlé a složité, že by přesahovalo rámec a zaměření této práce.

1.2 Statistické testy

Po nastudování dané problematiky byly zvolené následující statistické testy, které jsou v některých případech velmi úzce spojené s grafickými metodami popsanými v předchozích kapitolách, a proto budou jen vzpomenuty.

1.2.1 Pravidlo 3-sigma

Pravidlo 3-sigma (viz kapitola 1.1.1) označuje za odlehlé hodnty ty, které jsou ve vzdálenosti větší než 3 směrodatných odchylek od průměru.

1.2.2 Pravidlo kvartilů

Toto pravidlo je založeno na tom, že ve vzdálenosti větší než 1,5-násobek mezikvartilového rozpětí (IQR) od krajních kvartilů, je možné prohlásit hodnoty za potencionálně odlehlé (viz kapitola 1.1.2).

1.2.3 Grubbsův test

Grubbsův test je určen pro soubory s větším počtem dat ([11], [4]). Tato data je potřeba vzestupně seřadit $x_1 \leq x_2 \leq \ldots \leq x_n$ aby bylo možné odhalit odlehlosti v minimální (x_{min}) nebo maximální (x_{max}) hodnotě souboru pomocí následujících vztahů.

$$G_{min} = \frac{\bar{x} - x_{min}}{s} \quad \text{a} \quad G_{max} = \frac{x_{max} - \bar{x}}{s} \tag{1.4}$$

Kde $\bar{x}$ značí aritmetický průměr a s výběrovou směrodatnou odchylku.

Vypočtené G_{min} a G_{max} je nutné porovnat s kritickými hodnotami s požadovanou hladinou významnosti a pokud je větší, znamená to, že se dají považovat za odlehlé hodnoty. .

1.2.4 GESD

GESD je upravenou verzí Grubbsova testu a dokáže zamezit takzvanému „maskování“, které vzniká v případě, kdy data obsahují více outlierů blízko sebe a tím se zvětší směrodatná odchylka, která má za následek snížení hodnot G_{min} (G_{max}), což může vést k neoznačení hodnot jako odlehlých ([4]).

Statistické metody vzpomenuty výše, jsou již implementovány v programovém prostředí Matlab R2017a a bylo jich využito.

2 PRAKTICKÁ ČÁST

Praktická realizace vybraných metod pro detekci odlehlých hodnot je realizována v programovém prostředí MATLAB ([5], [17]). Tyto metody jsou aplikovány na reálná data pacientů s Parkinsonovou nemocí. Výstupem zpracovaných dat jsou tabulky, statisticky popisující klinická data (viz kapitola 2.1) a řečová cvičení, pro lepší představu o tom, jaká data a v jakém rozsahu se v souboru nachází. Dalším výstupem jsou grafy, které umožní graficky odhadnout přítomnost odlehlých hodnot. Vše bylo po otestování funkcí a skriptů implementováno do grafického rozhranní (viz kapitola 2.3).

2.1 Databáze

Reálná data pochází z české řečové databáze pacientů s Parkinsonovou nemocí PARCZ (Czech Parkinsonian Speech Database), ([8]). Databáze vznikla ve spolupráci s pracovištěm I. Neurologické kliniky Lékařské fakulty Masarykovy univerzity a Fakultní nemocnice u sv. Anny v Brně.

Databáze obsahuje celkem 91 řečových úloh, které byly absolvovány jak pacienty s Parkinsonovou nemocí, tak i věkově vázanými kontrolními řečníky, u kterých nebyla diagnostikována žádná neurologická onemocnění či řečové poruchy. Tyto úlohy byly zaznamenávány v tiché místnosti, v níž okolní hluk nepřesáhl úroveň 30 dB (měřeno akustickým analyzátorem NTI Acoustilyzer AL1). Dále jsou v databázi informace o klinickém stavu pacientů. Konkrétně:

- věk pacienta
- doba trvání PN v letech
- UPDRS III – unifikovaná škála hodnotící motorické aspekty PN
- UPDRS IV – unifikovaná škála hodnotící problémy terapie PN
- FOG-Q – škála hodnotící poruchy chůze
- NMSS – škála hodnotící nemotorické aspekty PN
- RBDSQ – škála hodnotící poruchy spánku
- BDI – škála hodnotící depresi při PN
- ACE-R – Addenbrookský kognitivní test – revidovaný
- MMSE – test kognitivních funkcí

2.2 Testování grafických metod

Celé testování a odzkoušení na datech z **feature_matrix.xlsx**, je rozděleno do devíti samostatných skriptů, s označením (s01-s09)*.m, dvou samostatných funkcí, boxPlot.m a kernel_densityPlot.m, jenž jsou k nahlédnutí v příloze A a B.

Skript **s01_prepare_data.m** slouží, jak je patrné z názvu, k přípravě dat pro další části. Nejprve jsou načtena všechna data z excelu **feature_matrix.xlsx** a poté jsou data roztříděna do proměnných:

- clinical_labels – klinické informace o pacientech
- feature_labels – obsahují hodnoty řečových parametrů
- clinical_data – matice dat popsaných v clinical_labels
- feature_data – matice řečových cvičení popsaných ve feature_labels
- feature_data_female – matice řečových cvičení pro ženy
- feature_data_male – matice řečových cvičení pro muže

Takhle roztříděná data jsou uložená do *.mat souborů, ale ještě předtím proběhne normalizace řečových dat na maximum sloupce, kdy jsou všechny hodnoty ve sloupci poděleny maximální hodnotou daného sloupce a standardizace řečových dat směrodatnou odchylkou, čímž se převedou na stejné měřítko ([15]).

V skriptu **s02_stats_clinical_features.m** jsou z klinických dat vytvořeny tabulky popisných statistik viz Tab. 2.1, které se uloží do excelu **statistics-clinical_data.xlsx** jak pro obě pohlaví dohromady, tak pro každé zvlášť (obdobně tomu bude i u dalších skriptů). Obsahují informace jako minimum, maximum, aritmetický průměr, medián, směrodatná odchylka, 1. a 3. kvartil.

Tab. 2.1: Statistická tabulka klinických dat pro obě pohlaví

	min	**Q1**	**median**	**Q3**	**max**	**mean**	**std**
Věk pacienta (v letech)	49	63	67	73	86	67,26	8,13
Délka trvání PN (v letech)	1	4	7	11	21	7,53	4,25
UPDRS III	3	13	25	33	55	24,01	12,30
UPDRS IV	0	0	2	5	9	2,91	2,70
FOG-Q	0	1	5,5	10	20	6,40	5,67
NMSS	2	18	33	52	94	35,30	20,90
RBDSQ	0	1	3	5	13	3,66	3,15
ACE-R	60	83	87	93	99	86,74	8,08
MMSE	16	28	28,5	29	30	27,90	2,55
BDI	0	6	9	13	27	10,46	6,18

Skript **s03_stats_speech_features** je stejný jako předchozí skript, s tím rozdílem, že se zpracují řečová cvičení a uloží do nového souboru **statistics-feature_data.xlsx**.

s04_histogram_clinical_features.m a také **s05_histogram_speech_features.m** slouží k vykreslení histogramů pro klinická data i řečová cvičení. Výsledné histogramy jsou uloženy do *.png souborů, stejně jako u ostatních skriptů, viz příloha C.

Dále, **s06_boxplot_clinical_features.m** a také **s07_boxplot_speech_features.m** volají funkci **boxPlot.m**, která společně s již implementovanou funkcí MATLABU boxplot.m slouží k vykreslení box-plot grafů. Zmíněná funkce má následující vstupní parametry:

- options.normalize – rozhodnutí o normalizaci (true/false)
- options.font_type – typ písma
- options.font_size – velikost písma
- options.xlabel – popis osy x
- options.ylabel – popis osy y
- options.labels – popisky jednotlivých boxplotů v grafu
- options.title – název grafu

s**08_kdeplot_clinical_features.m** a **s09_kdeplot_speech_features.m** volají funkci **kernel_densityPlot.m**, která společně s implementovanou funkcí ksdensity.m slouží k vykreslení KDE grafů (jádrový odhad hustoty pravděpodobnosti). Vstupní parametry této funkce jsou následující:

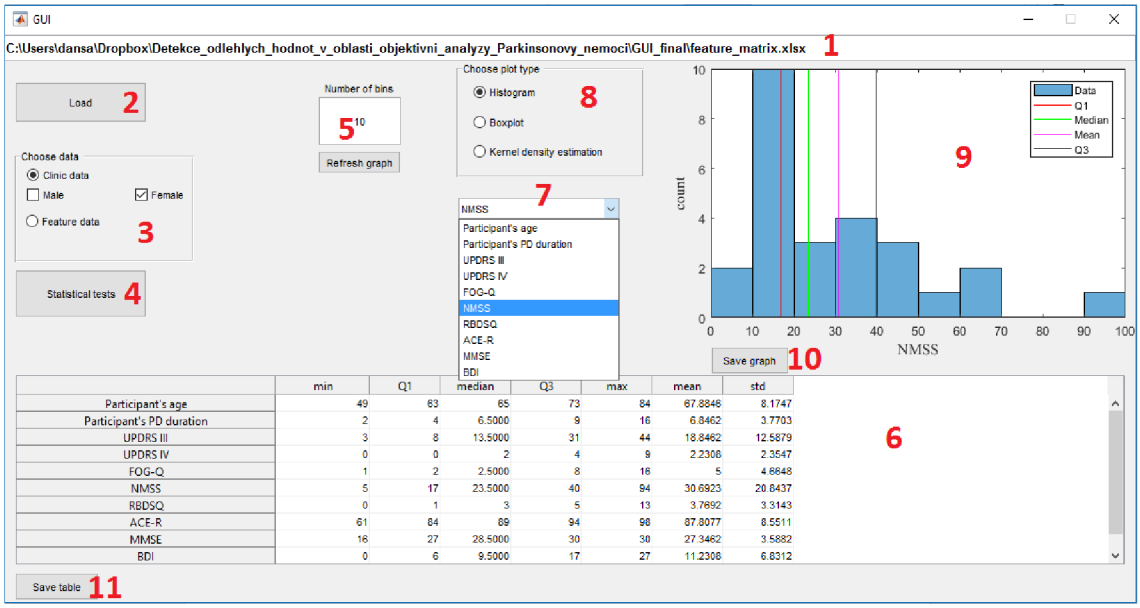
- options.normalize – rozhodnutí o normalizaci (true/false)
- options.plot_values – rozhodnutí o vykreslení aktuálních dat(true/false)
- options.plot_legend – rozhodnutí o vykreslení legendy(true/false)
- options.kernel – výběr jádrové funkce
- options.npoints – počet bodů odhadu jádrové hustoty
- options.graph_colors – barvy použité pro vykreslení průběhů
- options.graph_lines – druhy čar pro vykreslení průběhů
- options.graph_symbols – symboly pro zobrazení do grafu
- options.legend_loc – umístění legendy
- options.font_type – typ písma
- options.font_size – velikost písma
- options.title – název grafu
- options.xlabel – popis osy x
- options.ylabel – popis osy y
- options.legend – popis legendy grafu

2.3 GUI

Grafické prostředí bylo realizováno v programovém prosředí Matlab R2017a pomocí interaktivního GUIDE (graphical user interface development environment). Vycházelo se z funkcí a skriptů popsaných v kapitole 2.2, které byly následně upraveny a rozšířeny pro potřebu správné funkčnosti aplikace.

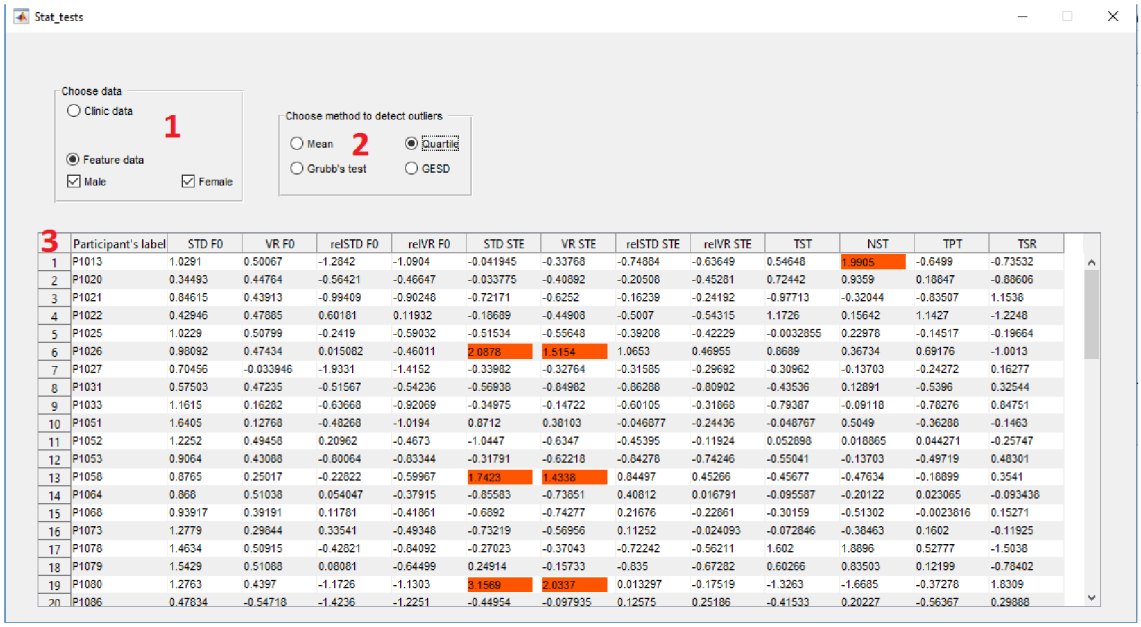
Samotná aplikace se skládá ze dvou oken. Hlavní okno spuštěné souborem GUI.m zobrazuje tabulku popisných statistik vybraných dat, společně s uživatelem vybraným typem grafu. Druhé okno spouštěné z prvního, pomocí tlačítka „Statistical tests“, slouží k zobrazení samotných zkoumaných dat, ve kterých je za pomocí zvolených metod, barevně zvýrazněn potencionální outlier.

Bližší pochopení funkčnosti je patrné z následujích snímků samotného grafického rozhraní, společně s popisem jednotlivých částí.



Obr. 2.1: Hlavní okno běžícího programu

1: Řádek pro výpis cesty k testovanému souboru dat
2: Tlačítko „Load“ sloužící k výběru a načtení souboru
3: Panel pro volbu klinických/řečových dat s následným výběrem pohlaví
4: Tlačítko „Statistical tests“ otevře další okno viz Obr. 2.2
5: Zadávání počtu tříd histogramu (default 10) s „refresh“ tlačítkem
6: Tabulka popisných statistik
7: Rozbalovací seznam pro výběr konkrétních dat k vykreslení zvoleného typu grafu
8: Panel pro výběr typu grafu (histogram, boxplot, KDE)
9: Vykreslení zvoleného grafu (u histogramu i se zvýrazněním kvartilů a průměru)
10: „Save“ tlačítko pro uložení aktuálního grafu (formát PNG)
11: „Save“ tlačítko pro uložení aktuální tabulky (6) do excelu



	Participant's label	STD F0	VR F0	relSTD F0	relVR F0	STD STE	VR STE	relSTD STE	relVR STE	TST	NST	TPT	TSR
1	P1013	1.0291	0.50067	-1.2842	-1.0904	-0.041945	-0.33768	-0.74884	-0.63649	0.54648	1.9905	-0.6499	-0.73532
2	P1020	0.34493	0.44764	-0.56421	-0.46647	-0.033775	-0.40882	-0.20508	-0.45281	0.72442	0.9359	0.18647	-0.88606
3	P1021	0.84615	0.43913	-0.99409	-0.90248	-0.72171	-0.6252	-0.16239	-0.24192	-0.97713	-0.32044	-0.83507	1.1536
4	P1022	0.42946	0.47885	0.60181	0.11932	-0.18689	-0.44908	-0.5007	-0.54315	1.1726	0.15642	1.1427	-1.2248
5	P1025	1.0229	0.50799	-0.2419	-0.59032	-0.51534	-0.55648	-0.39208	-0.42229	-0.0032855	0.22978	-0.14517	-0.19664
6	P1026	0.98092	0.47434	0.015082	-0.46011	2.0878	1.5154	1.0653	0.46955	0.8689	0.36734	0.69176	-1.0013
7	P1027	0.70456	-0.033946	-1.9331	-1.4152	-0.33882	-0.32764	-0.31585	-0.29692	-0.30962	-0.13703	-0.24272	0.16277
8	P1031	0.57503	0.47235	-0.51567	-0.54236	-0.56938	-0.84982	-0.86288	-0.80902	-0.43536	0.12891	-0.5396	0.32544
9	P1033	1.1615	0.16282	-0.63668	-0.92069	-0.34975	-0.14722	-0.60105	-0.31868	-0.79387	-0.09118	-0.76276	0.84751
10	P1051	1.6405	0.12768	-0.48268	-1.0194	0.8712	0.38103	-0.046877	-0.24436	-0.046767	0.5049	-0.36288	-0.1463
11	P1052	1.2252	0.49458	0.20962	-0.4673	-1.0447	-0.6347	-0.45395	-0.11924	0.052898	0.018865	0.044271	-0.25747
12	P1053	0.9064	0.43068	-0.80064	-0.83344	-0.31791	-0.62218	-0.84278	-0.74246	-0.55041	-0.13703	-0.49719	0.48301
13	P1056	0.8765	0.25017	-0.22822	-0.59967	1.7423	1.4338	0.84497	0.45266	-0.45677	-0.47634	-0.16899	0.3541
14	P1064	0.868	0.51038	0.054047	-0.37915	-0.85583	-0.73851	0.40812	0.016791	-0.095587	-0.20122	0.023065	-0.093436
15	P1068	0.93917	0.39191	0.11761	-0.41861	-0.6892	-0.74277	0.21676	-0.22661	-0.30159	-0.51302	-0.0023816	0.15271
16	P1073	1.2779	0.29844	0.33541	-0.49348	-0.73219	-0.56956	0.11252	-0.024093	-0.072846	-0.38463	0.1602	-0.11925
17	P1078	1.4634	0.50915	-0.42821	-0.84092	-0.27023	-0.37043	-0.72242	-0.56211	1.602	1.8696	0.52777	-1.5038
18	P1079	1.5429	0.51068	0.08081	-0.64499	0.24914	-0.15733	-0.835	-0.67282	0.60266	0.83503	0.12199	-0.78402
19	P1080	1.2763	0.4397	-1.1726	-1.1303	3.1569	2.0337	0.013297	-0.17519	-1.3263	-1.6685	-0.37278	1.8309
20	P1086	0.47834	-0.54718	-1.4236	-1.2251	-0.44954	-0.097935	0.12575	0.25186	-0.41533	0.20227	-0.56367	0.29888

Obr. 2.2: Okno se statistickými testy

1: Panel pro volbu klinických/řečových dat s následným výběrem pohlaví
2: Panel pro výběr statistické metody pro detekci odlehlých hodnot
3: Tabulka zvolených dat s vyznačenými možnými odlehlými pozorováními na základě zvolené metody pro detekci

3 ZÁVĚR

V této Bakalářské práci byly nastudovány metody, běžně používané při statistické analýze datového souboru. V oblasti objektivní IT analýzy Parkinsonovy nemoci je analýza statistických vlastností dat a detekce odlehlých pozorování před samotným zpracováním nutnou součástí následného klinicky validního postupu.

Metody, které byly v práci nastudovány byly následně implementovány v programovém prostředí MATAB R2017a. Mezi tyto metody patří analýza histogramu, tzv. „box plot“ (krabicový graf) a také tzv. „kde plot“ (graf jádrového odhadu hustoty pravděpodobnosti). Metoda Q-Q grafu byla popsána teoreticky. Její implementace byla po dohodě s vedoucím práce zrušena. Dále statistické testy pro detekci odlehlých hodnot, mezi které patří pravidlo 3-sigma, metoda kvartilů, Grubbsův test a GESD.

Tyto metody byly následně upraveny a zakomponovány do grafického uživatelského rozhraní. Tato aplikace byla odzkoušena na datech z databáze PARCZ. Do budoucna je možnost rozšíření aplikace o další metody pro detekci outlierů a propracovanější nakládání s daty. Například možnost zpětného upravení či odstranění dat původního souboru přímo z okna aplikace.

LITERATURA

[1] ACUNA, Edgar a Caroline RODRIGUEZ. A Meta analysis study of outlier detection methods in classification [online]. University of Puerto Rico at Mayaguez, 2004 [cit. 2016-11-25].Dostupné z: https://www.researchgate.net/profile/Edgar_Acuna/publication/228728761_A_meta_analysis_study_of_outlier_detection_methods_in_classification/links/00b7d525e85fae9659000000.pdf

[2] Box-and-Whisker Plots: Quartiles, Boxes, and Whiskers. Purplemath [online]. [cit. 2016-12-12]. Dostupné z: http://www.purplemath.com/modules/boxwhisk.htm

[3] DUONG, Tarn. An introduction to kernel density estimation. University of Western Australia, 2001. Dostupné také z: http://www.mvstat.net/tduong/research/seminars/seminar-2001-05/index.html

[4] How to: Identify outliers. GraphPad Software [online]. [cit. 2017-06-01]. Dostupné z: http://www.graphpad.com/guides/prism/7/statistics/index.htm?stat_how_to_removing_outliers.htm

[5] KOLÁČEK, Jan a Kateřina KONEČNÁ. Jak pracovat s MATLABem. Brno. Dostupné také z: https://www.math.muni.cz/~kolacek/vyuka/vypsyst/navod.pdf

[6] LATECKI, Jan, Aleksandar LAZAREVIC a Dragoljub POKRAJAC. Outlier Detection with Kernel Density Functions. 2007. Dostupné také z: https://cis.temple.edu/~latecki/Papers/mldm07.pdf

[7] LOUDOVÁ, Zuzana. Metody identifikace odlehlých pozorování. Plzeň, 2015. Diplomová práce. Západočeská univerzita v Plzni.

[8] MEKYSKA, Jiří. Analýza řečových promluv pro IT diagnostiku neurologických onemocnění. Brno, 2014. Dizertační práce. Vysoké učení technické v Brně.

[9] Normální rozdělení. WikiSkripta [online]. 2013 [cit. 2016-12-12]. Dostupné z: http://www.wikiskripta.eu/index.php/Norm%C3%A1ln%C3%AD_rozd%C4%9Blen%C3%AD

[10] ORAVA, Jan. Volba vyhlazovacího parametru při jádrových odhadech hustoty. Brno, 2008. Diplomová práce. Masarykova univerzita.

[11] Procedures for Detecting Outlying Observations in Samples, Technometrics 11 [online]. E. GRUBBS, Frank. 1969, s. 1-21 [cit. 2017-06-01].

[12] Quantile-Quantile (q-q) Plots Author(s). Online Statistics Education: An Interactive Multimedia Course of Study [online]. [cit. 2016-12-12]. Dostupné z: http://onlinestatbook.com/2/advanced_graphs/q-q_plots.html

[13] SCHULZ, Jaroslav. Diagnostické grafy v programu STATISTICA. Brno, 2007. Bakalářská práce. Masarykova univerzita.

[14] SEO, Songwon. A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets. University of Pittsburgh, 2006.

[15] Standardizace dat. Matematická biologie [online]. [cit. 2016-12-12]. Dostupné z: http://portal.matematickabiologie.cz/index.php?pg=analyza-a-hodnoceni-biologickych-dat--vicerozmerne-metody-pro-analyzu-dat--vicerozmerna-rozdeleni-pravdepodobnosti--transformace-dat--standardizace-dat

[16] THODE, Henry C. Testing for normality. 2002. New York: Marcel Dekker, s. 123-126. ISBN 0-8247-9613-6.

[17] ZAPLATÍLEK, Karel. MATLAB – průvodce začínajícího uživatele. 2011. Tribun EU. ISBN 978-80-263-0014-4.

SEZNAM SYMBOLŮ, VELIČIN A ZKRATEK

ACE-R	Addenbrookský kognitivní test – revidovaný
BDI	škála hodnotící depresi při PN
FOG-Q	škála hodnotící poruchy chůze
IQR	mezikvartilové rozpětí
KDE	jádrový odhad hustoty pravděpodobnosti
MMSE	test kognitivních funkcí
NMSS	škála hodnotící nemotorické aspekty PN
PARCZ	česká řečová databáze pacientů s Parkinsonovou nemocí
PN (PD)	Parkinsonova nemoc
Q-Q	kvantil-kvantil
Q_x	kvartil
RBDSQ	škála hodnotící poruchy spánku
std (σ)	směrodatná odchylka
UPDRS III	unifikovaná škála hodnotící motorické aspekty PN

$\hat{f}_h$	spojitý odhad hustoty pravděpodobnosti
h	vyhlazovací parametr
k	počet intervalů
K	jádro
n	počet prvků/pozorování
e	eulerovo číslo
s	výběrová směrodatná odchylka
x	bod na ose x
$\bar{x}$	aritmetický průměr
$x_{min(max)}$	minimální (maximální) hodnota seřazených dat
$G_{min(max)}$	vypočtené testovací kritérium
x_i	i-tý bod na ose x

SEZNAM PŘÍLOH

A ZDROJOVÝ KÓD FUNKCE BOX PLOT

```
function boxPlot(data, options)

%% Paths and variables
if ((nargin < 2) || isempty(options))
    options.normalize = true;
    options.font_type = 'Times New Roman';
    options.font_size = 10;
    options.xlabel    = 'data [-]';
    options.ylabel    = 'data values [-]';
    options.title     = '';
    options.whisker   = 1.5;
    num_labels        = 1:size(data, 2);
    options.labels    = cellstr(num2str(num_labels(:)));
else
    if (~isfield(options, 'normalize'))
        options.normalize = true;
    end
    if (~isfield(options, 'font_type'))
        options.font_type = 'Times New Roman';
    end
    if (~isfield(options, 'font_size'))
        options.font_size = 10;
    end
    if (~isfield(options, 'xlabel'))
        options.xlabel = 'data [-]';
    end
    if (~isfield(options, 'ylabel'))
        options.ylabel = 'data values [-]';
    end
    if (~isfield(options, 'title'))
        options.title = '';
    end
    if (~isfield(options, 'whisker'))
        options.whisker = 1.5;
    end
    if (~isfield(options, 'labels'))
        num_labels     = 1:size(data, 2);
        options.labels = cellstr(num2str(num_labels(:)));
    end
end

%% Set temporary variables (for: code readability)
NUM = size(data, 2);
OPT = options;

%% Normalize feature values
if (OPT.normalize)
    for dat = 1:NUM
        tmp = data(:, dat);
        data(:, dat) = (tmp - mean(tmp))/std(tmp);
    end
end

%% Plot the boxplot
boxplot(data, 'whisker', OPT.whisker, 'labels', OPT.labels);

%% Set the graph properties
xlabel(OPT.xlabel, 'FontSize', OPT.font_size, 'FontName', OPT.font_type);
ylabel(OPT.ylabel, 'FontSize', OPT.font_size, 'FontName', OPT.font_type);
title(OPT.title, 'FontSize', OPT.font_size, 'FontName', OPT.font_type);
```

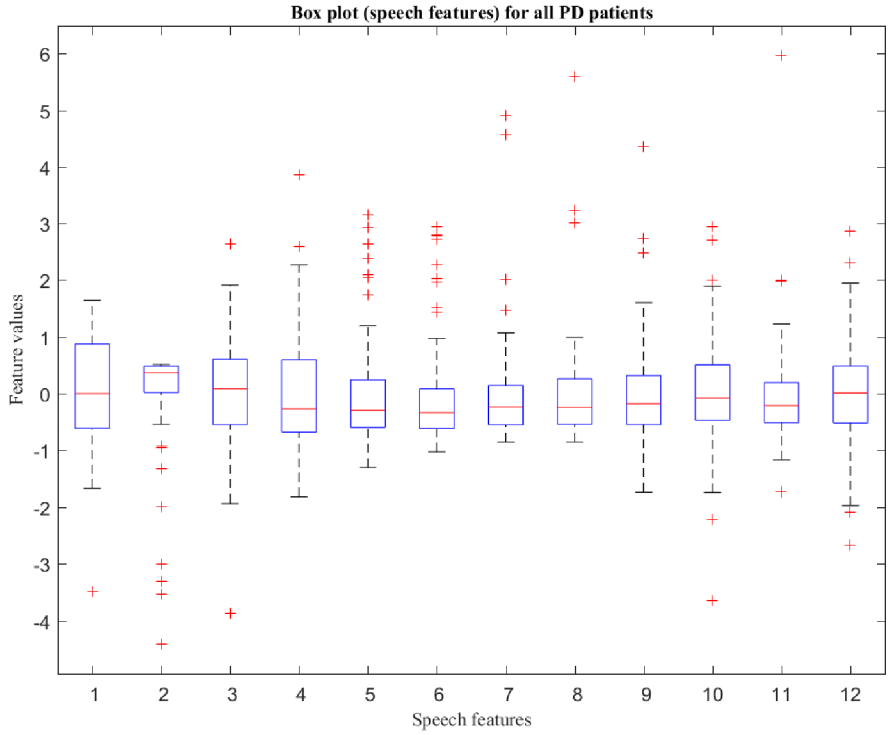
B ZDROJOVÝ KÓD FUNKCE KDE PLOT

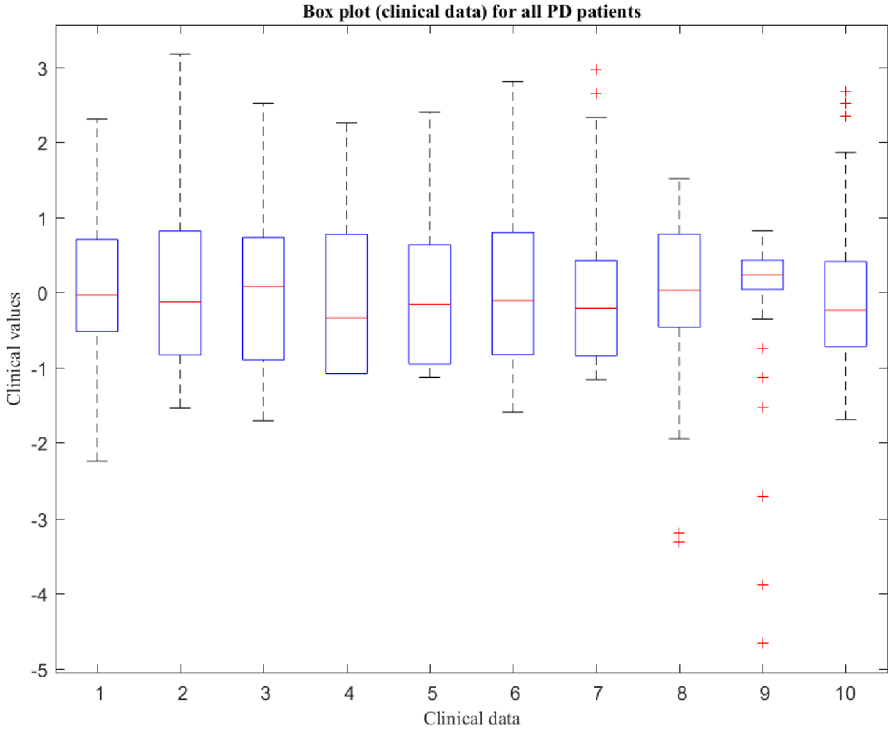
```
function kernel_densityPlot(data, options)

%% Paths and variables
if ((nargin < 2) || isempty(options))
    options.normalize      = true;
    options.plot_values    = true;
    options.plot_legend    = true;
    options.kernel         = 'normal';
    options.npoints        = 100;
    options.graph_colors   = {'b'; 'r'; 'c'; 'm'; 'k'; 'y'};
    options.graph_lines    = {'-'; '--'; '-.'; ':'; '-'; '--'};
    options.graph_symbols  = {'o'; '+'; '*'; 'x'; 's'; 'd'};
    options.legend_loc     = 'northeast';
    options.font_type      = 'Times New Roman';
    options.font_size      = 10;
    options.title          = '';
    options.xlabel         = 'data values [-]';
    options.ylabel         = 'probability density function [-]';
    options.legend         = [];
else
    if (~isfield(options, 'normalize'))
        options.normalize = true;
    end
    if (~isfield(options, 'plot_values'))
        options.plot_values = true;
    end
    if (~isfield(options, 'plot_legend'))
        options.plot_legend = true;
    end
    if (~isfield(options, 'kernel'))
        options.kernel = 'normal';
    end
    if (~isfield(options, 'npoints'))
        options.npoints = 100;
    end
    if (~isfield(options, 'graph_colors'))
        options.graph_colors = {'c'; 'm'; 'b'; 'r'; 'k'; 'y'};
    end
    if (~isfield(options, 'graph_lines'))
        options.graph_lines = {'-'; '--'; '-.'; ':'; '-'; '--'};
    end
    if (~isfield(options, 'graph_symbols'))
        options.graph_symbols = {'o'; '*'; '+'; 'x'; 's'; 'd'};
    end
    if (~isfield(options, 'legend_loc'))
        options.legend_loc = 'northeast';
    end
    if (~isfield(options, 'font_type'))
        options.font_type = 'Times New Roman';
    end
    if (~isfield(options, 'font_size'))
        options.font_size = 10;
    end
    if (~isfield(options, 'title'))
        options.title = '';
    end
    if (~isfield(options, 'xlabel'))
        options.xlabel = 'data values [-]';
    end
```

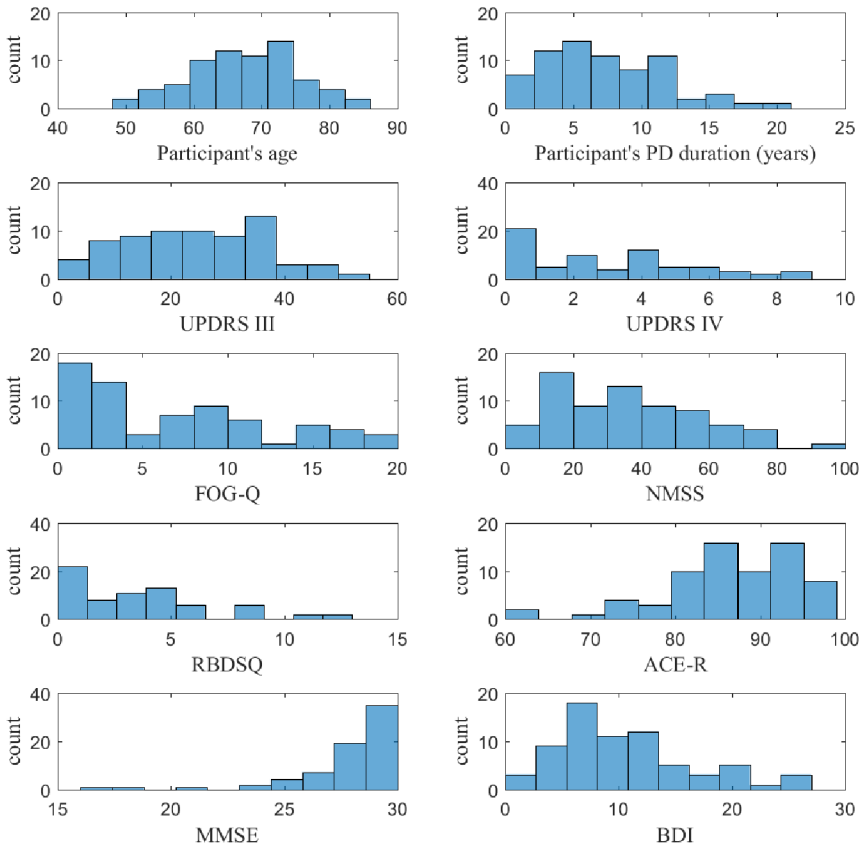
```
    if (~isfield(options, 'ylabel'))
        options.ylabel = 'probability density function [-]';
    end
    if (~isfield(options, 'legend'))
        options.legend = [];
    end
end

%% Check data consistency
if (size(data, 2) > length(options.graph_colors))
    message = ['maximum number of columns exceeded (set to ´´: ' ...
        num2str(length(options.graph_colors)) ')'];
    warning(message);

    % Limit the number of columns in data matrix
    data = data(:, 1:length(options.graph_colors));
end

%% Set temporary variables (for: code readability)
NUM = size(data, 2);
EST = struct();
OPT = options;
brd = ones(1, NUM);
F   = [];
D   = [];
h   = figure;

%% Process the data (per columns)
for dat = 1:NUM
    vec = data(:, dat);
    vec = cell2mat(vec);

    % Normalize feature values
    if (OPT.normalize)
        vec = (vec - mean(vec))/std(vec);
    end

    % Compute the probability density estimation
    [EST(dat).f, EST(dat).x] = ksdensity(vec, ...
        'kernel', OPT.kernel, 'npoints', OPT.npoints);
    brd(dat) = length(vec);
    D = [D; vec(:)];
    F = [F EST(dat).f];
end

%% Plot the probability density function estimation
for dat = 1:NUM
    col = [OPT.graph_colors{dat} OPT.graph_lines{dat}];
    plot(EST(dat).x, EST(dat).f, col);
    hold on;
end

%% Plot the actual data values
for dat = 1:NUM
    if (OPT.plot_values)
        if (dat == 1)
            low = 1;
        else
            low = brd(dat - 1) + 1;
        end

        high = sum(brd(1:dat));
        col  = [OPT.graph_colors{dat} OPT.graph_symbols{dat}];
```

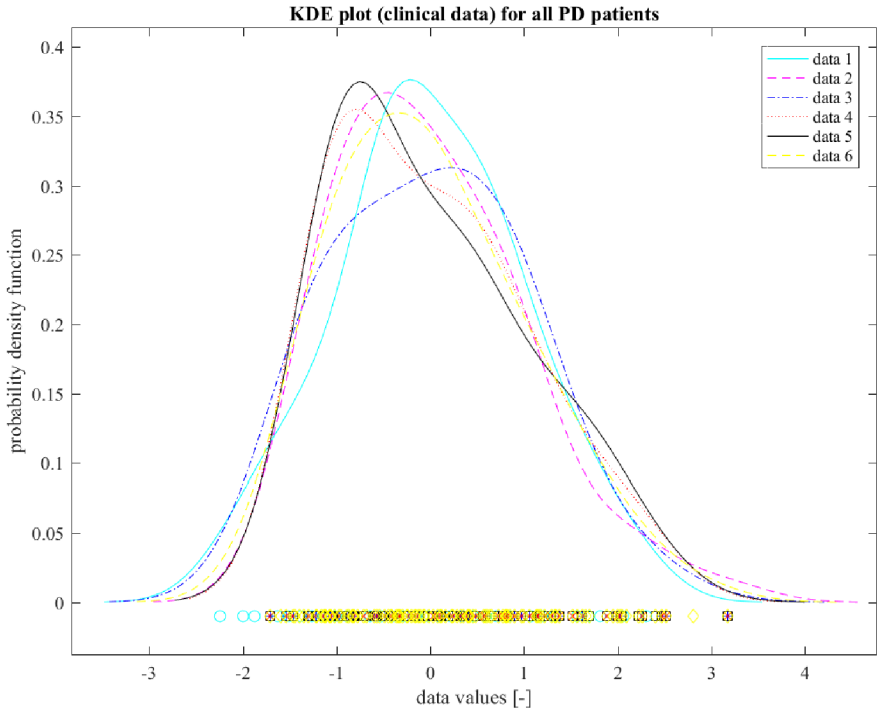
```
        plot(D(low:high), ones(length(D(low:high)), 1)*(-0.01), col);
        hold on;
    end
end

%% Set the legend
if (OPT.plot_legend)
    if (isempty(OPT.legend))
        L = cell(1, size(data, 2));
        for dat = 1:NUM
            L{dat} = ['data ' num2str(dat)];
        end
    else
        L = OPT.legend;
    end
end

%% Set the graph properties
x_start = min(D(:)) - max(D(:))/2;
y_start = -(max(max(F(:)))/10);
x_end   = max(D(:)) + max(D(:))/2;
y_end   = max(max(F(:))) + max(max(F(:)))/10;

xlim([x_start x_end]);
ylim([y_start y_end]);

xlabel(OPT.xlabel, 'FontSize', OPT.font_size, 'FontName', OPT.font_type);
ylabel(OPT.ylabel, 'FontSize', OPT.font_size, 'FontName', OPT.font_type);
title(OPT.title, 'FontSize', OPT.font_size, 'FontName', OPT.font_type);

if (OPT.plot_legend)
    legend(L, 'Location', OPT.legend_loc);
end

p = get(h, 'CurrentAxes');
set(p, 'FontSize', OPT.font_size, 'FontName', OPT.font_type);
hold off;
grid off;
```

C PŘÍKLADY ZOBRAZENÍ






count
Participant's age
Participant's PD duration (years)
UPDRS III
UPDRS IV
FOG-Q
NMSS
RBDSQ
ACE-R
MMSE
BDI


KDE plot (clinical data) for all PD patients
data 1
data 2
data 3
data 4
data 5
data 6
probability density function
data values [-]