

Průzkumová analýza dat a metoda hlavních komponent

4ST512 Vícerozměrná statistika

Lubomír Štěpánek

24. října 2017

Obsah

1	Zadání úlohy	1
2	Řešení úlohy	2
2.1	Metodologie a analýza dat	2
2.1.1	Odlehlá pozorování	2
2.1.2	Ověřování normality	2
2.1.3	Analýza hlavních komponent	3
2.2	Výsledky	3
2.3	Závěr	10
3	Apendix	10
4	Reference	19

1 Zadání úlohy

Východiskem je soubor `du1_30.sav`, který obsahuje šest vybraných biometrických měr u 254 žen.

- (i) Prozkoumejme data s ohledem na marginální i sdružené rozdělení veličin a identifikujme případné odchylky od normality a podezřelá pozorování.
- (ii) U podezřelých pozorování rozhodněme, zda jsou chybná, případně se pokusme odchylky vysvětlit.
- (iii) Dále se zabývejme „dimenzionalitou“ dat. Jsou rozměry mezi sebou nezávislé, nebo naopak další proměnné nepřináší unikátní informaci?
- (iv) Rozhodněme, zda je vhodné pomocí metody hlavních komponent analyzovat kovarianční, nebo korelační matici. Pokuste se interpretovat hlavní komponenty¹.

¹Hodnocena bude správnost posouzení jedno- a vícerozměrné normality proměnných (včetně vysvětlení odchylek), identifikace odlehlých pozorování, určení dimenzionality dat, interpretace komponent. Všechny výstupy musí být okomentovány a interpretovány, všechna rozhodnutí musí být zdůvodněna. Body mohou být odečteny i za nedodržení formálních požadavků (délka, srozumitelnost, čitelnost výstupů).

2 Řešení úlohy

2.1 Metodologie a analýza dat

Celá úloha byla řešena v prostředí R, které je určeno pro statistické výpočty a následné grafické náhledy [1]. Datový soubor `du1_30.sav` byl pomocí balíčku `foreign` nahrán do prostředí R.

Pro účely průzkumové analýzy data (EDA, exploratory data analysis) byly použity grafické nástroje jazyka R.

2.1.1 Odlehlá pozorování

Pro detekci možných odlehlých pozorování bylo použito zobrazení každé proměnné pomocí boxplotu a dále metoda *inner and outer fences* (*vnitřní a vnější hradby*) [2].

Je-li $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$, pak suspektně odlehlá v rámci výběru \mathbf{x} je ta hodnota x_i , pro kterou platí, že

$$x_i \notin \langle Q_1 - 1,5 \cdot (Q_3 - Q_1); Q_3 + 1,5 \cdot (Q_3 - Q_1) \rangle, \quad (1)$$

kde $i \in \{1, 2, \dots, n\}$, n je rozsah souboru, Q_1 a Q_3 je první a třetí výběrový kvartil nad vektorem dat \mathbf{x} , respektive. Graficky suspektní odlehlost zobrazují i krabicové diagramy pomocí samostatně zobrazených pozorování mimo krabici danou kvartilovým rozpětím. Bylo by možné použít i formální testy hypotéz na jednorozměrnou odlehlost v rámci výběru \mathbf{x} , například Dixonův test nebo Grubbsův test. Nicméně vždy jde o podpůrnou metodu, konečné rozhodnutí o odlehlosti vytipované hodnoty má spíše expertní charakter.

2.1.2 Ověřování normality

Pro ověřování jednorozměrné normality dat spojitých proměnných byly použity histogramy, dále byly vykresleny kvantil-kvantil diagramy porovnávající teoretické a ve výběrech pozorované kvantily normálního rozdělení. Nakonec byly provedeny formální testy normálnosti, a sice Kolmogorov-Smirnovův a Shapiro-Wilkův test.

V histogramech byl počet intervalů k , na které byl rozsah hodnot proměnné rozdělen, určen podle *Sturgesova pravidla*, tedy $k = \lceil \log_2 n \rceil + 1$, kde n je počet pozorování ve výběru proměnné.

Kolmogorovův-Smirnovův jednovýběrový test zkoumá nulovou hypotézu H_0 o tom, že výběr pochází z předpokládaného teoretického rozdělení, zde normálního, pomocí statistiky D_1 . Vstupem jednovýběrového testu je k tříd testovaného výběru a předpokládané teoretické rozdělení, které je rozděleno do stejněho počtu tříd. Nad každou třídou testovaného výběru spočteme četnosti n_{1i} a nad každou třídou teoretického rozdělení četnosti n_{2i} , kde $i \in \{1, 2, \dots, k\}$. Poté vyčíslíme kumulativní četnosti pro výběr $N_{1i} = \sum_{j=1}^i n_{1j}$ a pro testované rozdělení $N_{2i} = \sum_{j=1}^i n_{2j}$. Hodnocené kritérium D_1 je pak

$$D_1 = \frac{1}{n} \max_i |N_{1i} - N_{2i}|,$$

kde n je celkový počet prvků výběru. Hodnota kritéria D_1 se porovná s kritickou hodnotou $D_{1,\max}$ pro danou hladinu významnosti α ; ta je tablována, případně ji lze pro větší hodnoty $D_{1,\max}$ odhadnout podle numerických pravidel, více v [3]. Hladina významnosti obou testů byla pro oba výběry vypočtena numericky v prostředí R pomocí funkce `ks.test()`.

Shapiro-Wilkův test testuje nulovou hypotézu H_0 o tom, že statistický výběr $\mathbf{x} = (x_1, x_2, \dots, x_n)^T$ pochází z normálního rozložení. Testová statistika

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

kde $x_{(i)}$ je i -tá nejmenší hodnota výběru \mathbf{x} , \bar{x} je výběrový průměr a a_i jsou konstanty dány Shapirou-Wilkovou metodikou podle [4]. Kritické hodnoty pro statistiku W jsou pro danou hladinu významnosti α tablovány nebo dostupné ve vhodném software. V prostředí R je Shapirův-Wilkův test implementován ve funkci `shapiro.test()`.

Pro ověření vícerozměrné normality byl použit R-kový balíček **MVN**, který nabízí testy vícerozměrné normality (např. Henze-Zinklerův test, [5]) stejně jako matice kvantil-kvantil diagramů pro sdružená vícerozměrná rozložení více proměnných z datasetu. Nakonec byla spočítána i adjustovaná robustní Mahalanobisova distance dle [6], která pomohla identifikovat pozorování podezřelá z vícerozměrné odlehlosti.

2.1.3 Analýza hlavních komponent

Přestože se používá i jako nástroj extrakce proměnných, zde bude použita jako metoda redukce dimenzionality v systému proměnných, které jsou navzájem relativně hodně korelovány.

Principem analýzy hlavních komponent je přepsání vstupních dat v matici \mathbf{X} na data \mathbf{Y} tak, aby $\mathbf{Y} = \mathbf{XP}$, kde \mathbf{P} je matice vlastních vektorů kovarianční matice $\Sigma_{\mathbf{X}}$. Kovarianční matice $\Sigma_{\mathbf{X}}$ splňuje vztah $\Sigma_{\mathbf{X}} = \mathbf{P}\Lambda\mathbf{P}^T$, kde Λ je diagonální matice obsahující na diagonále vlastní čísla matice $\Sigma_{\mathbf{X}}$.

Seřadíme-li vlastní vektory v \mathbf{P} podle velikosti vlastních čísel, dostaneme složky v \mathbf{Y} setříděné podle rozptylu, čímž lze určit, kolik prvních vlastních vektorů (komponent) již suficientně vysvětlí požadovanou míru variability \mathbf{Y} . Podrobněji pak v [7].

2.2 Výsledky

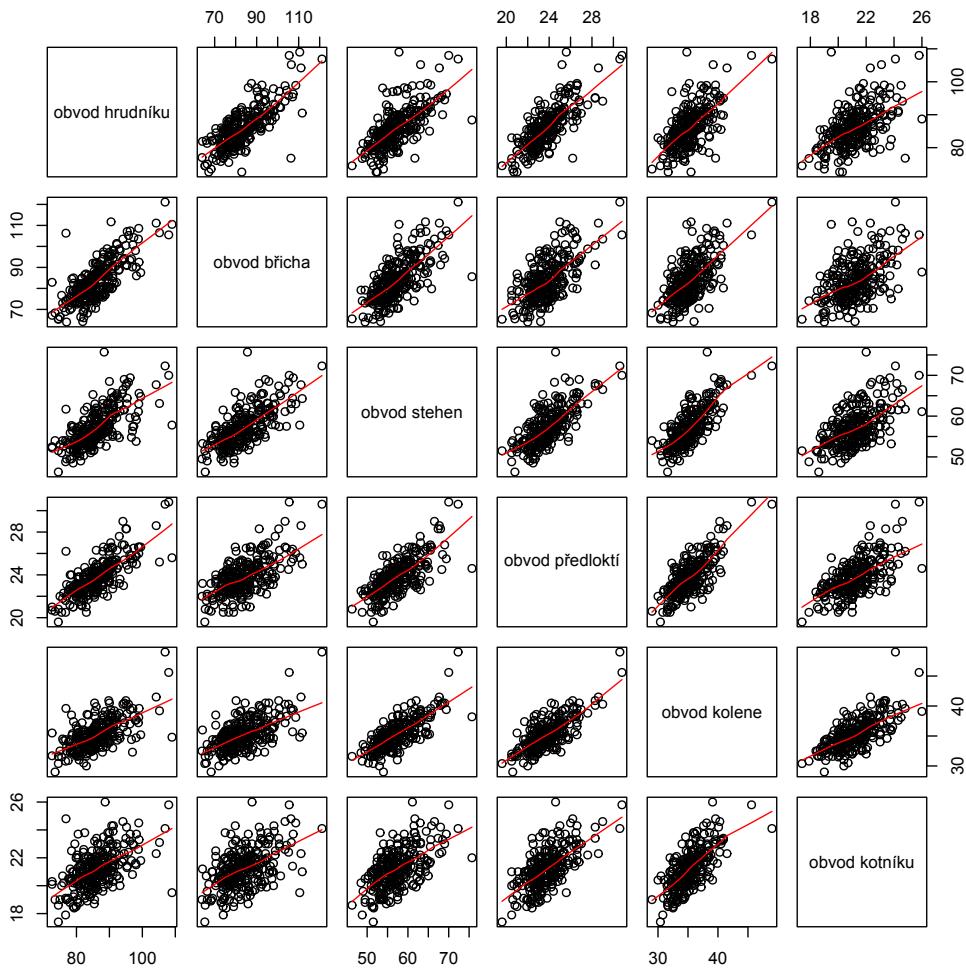
Dataset `du1_30.sav` obsahuje šest numerických spojitých proměnných, a sice *obvod hrudníku*, *obvod břicha*, *obvod stehen*, *obvod předloktí*, *obvod kolene*, *obvod kotníku*, a jednu identifikátorovou proměnnou *id*. Nejsou přítomny žádné chybějící hodnoty. Základní vztahy mezi numerickými proměnnými vidíme v rámci průzkumové analýzy dat na obrázku 1.

Průzkumová analýza dat. Z obrázku 1 se jeví, že mezi všemi šesti spojitými proměnnými je vztah typu přímé úměry, přesněji s růstem jedné proměnné rostou hodnoty každé další (bez ohledu na kauzalitu).

Odlehlá pozorování a ověřování normality. Z tabulky 1 nahlédneme, že předpoklad normálního rozdělení není u žádné proměnné dobře udržitelný; je to však dáno i převažujícími horními odlehlými hodnotami ve výběrech. Zároveň však z histogramů a především z boxplotů na obrázcích 2 až 7 vidíme, že žádná detekovaná odlehlá hodnota se zbylým hodnotám daného výběru nevymyká řádově. Lze tedy předpokládat, že detekované odlehlé hodnoty pomocí formule (1) nejsou ve skutečnosti hodnoty vzniklé náhodnou chybou (např. měření), ale jde o varianty na hraně normy, popisující skutečnou realitu. Pro další analýzy vyžadující normální rozdělení by bylo nutné zvážit vyloučení odlehlých hodnot či transformace proměnných např. logaritmováním, jak radí [8]. Zešikmení zprava (odporující dobré normalitě dat) je pak vždy vidět i na histogramech a na kvantil-kvantil diagramech jako tzv. lehké konce (*light tails*), viz obrázky 2 až 7.

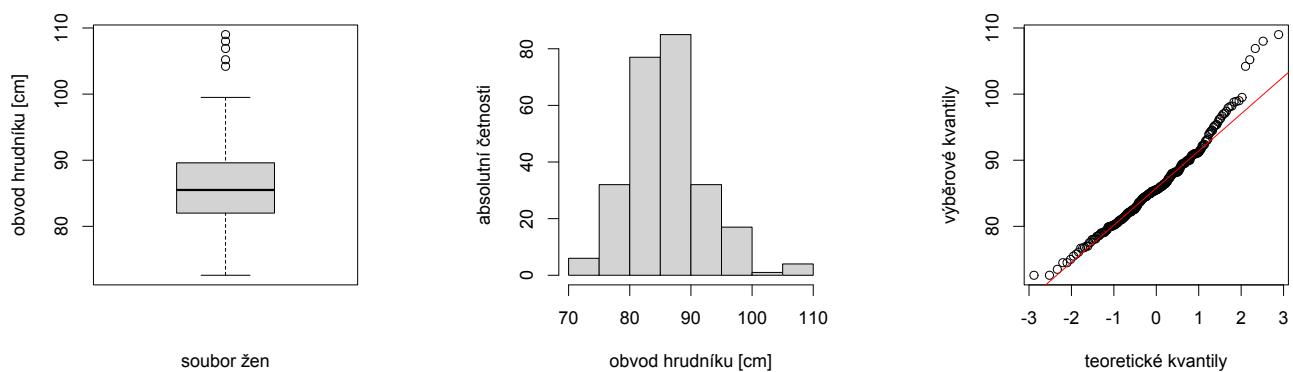
proměnná	<i>p</i> -hodnota Kolgomorova-Smirnova testu	<i>p</i> -hodnota Shapiro-Wilkova testu
obvod hrudníku	< 0,0001	< 0,0001
obvod břicha	< 0,0001	< 0,0001
obvod stehen	< 0,0001	< 0,0001
obvod předloktí	< 0,0001	< 0,0001
obvod kolene	< 0,0001	< 0,0001
obvod kotníku	< 0,0001	0,0911

Tabulka 1: Odhadování hladin významnosti (*p*-hodnot) testů normality pro jednotlivé proměnné



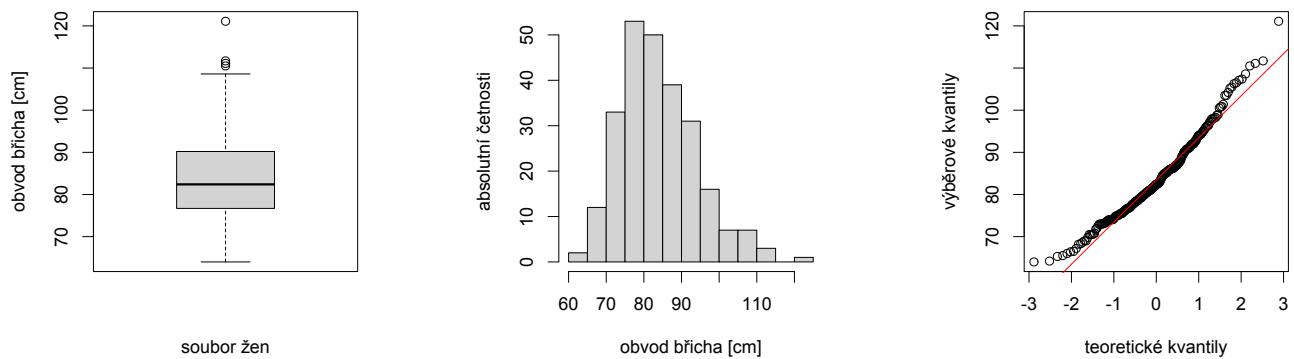
Obrázek 1: Matice bodových diagramů s naznačeným lokálně-váženým (LOESS) regresním trendem (červeně).

Na obrázku 2 je krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod hrudníku*. Lze nahlédnout, že proměnná obsahuje dle podmínky (1) několik suspektně odlehčlých pozorování (vzhledem ke zbylým pozorováním), a sice pozorování s indexy 8, 44, 154, 159 a 161 (horní odlehčlé). Protože jde o soubor žen, může jít o hodnoty obvodu hrudníku u žen s tzv. *makromastii*; jde o benigní stav, kdy je objem prsů nadměrně zvětšen po menopauze.



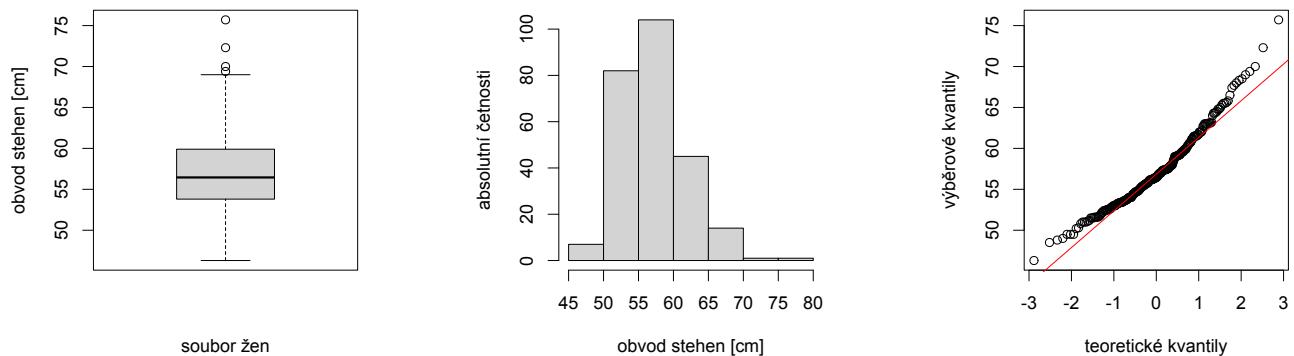
Obrázek 2: Krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod hrudníku*

Na obrázku 3 je krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod břicha*. Lze nahlédnout, že proměnná obsahuje dle podmínky (1) několik suspektně odlehlych pozorování (vzhledem ke zbylým pozorováním), a sice pozorování s indexy 8, 95, 154 a 161 (horní odlehly). Může se jednat například o ženy s gynoidním typem obezity v rámci metabolického syndromu.



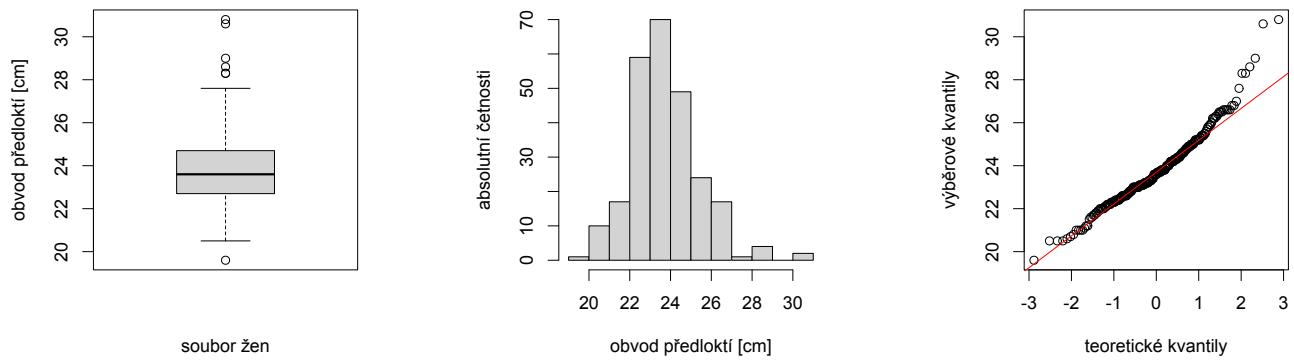
Obrázek 3: Krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod břicha*

Na obrázku 4 je krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod stehen*. Lze nahlédnout, že proměnná obsahuje dle podmínky (1) několik suspektně odlehlych pozorování (vzhledem ke zbylým pozorováním), a sice pozorování s indexy 2, 22, 44, 115, 154 (horní odlehly). Může se jednat opět například o ženy s gynoidním typem obezity (typ „hruška“) v rámci metabolického syndromu, kdy je somatický tuk ukládán v oblasti boků a stehen.



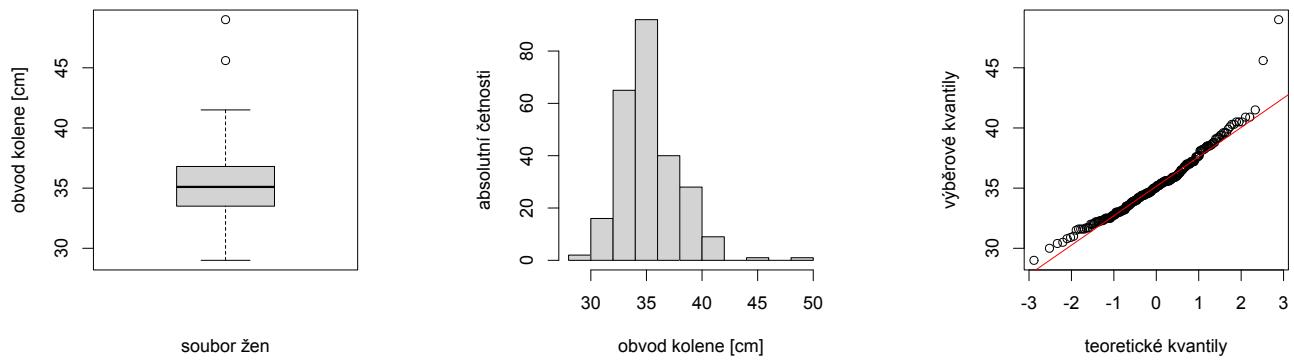
Obrázek 4: Krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod stehen*

Na obrázku 5 je krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod předloktí*. Lze nahlédnout, že proměnná obsahuje dle podmínky (1) několik suspektně odlehlych pozorování (vzhledem ke zbylým pozorováním), a sice pozorování s indexy 166 (dolní odlehly), 8, 17, 44, 154, 155, 232 (horní odlehly). V případě dolní odlehly může jít o svalovou atrofii při sarkopenii či kachexii onkologickém onemocnění, v případě horních odlehlych se může jednat typicky o ženy po ablaci mammy, kdy poté často odchází k lymfatickým otokům horní končetiny na ipsilaterální straně (tzv. *lymfedém*).



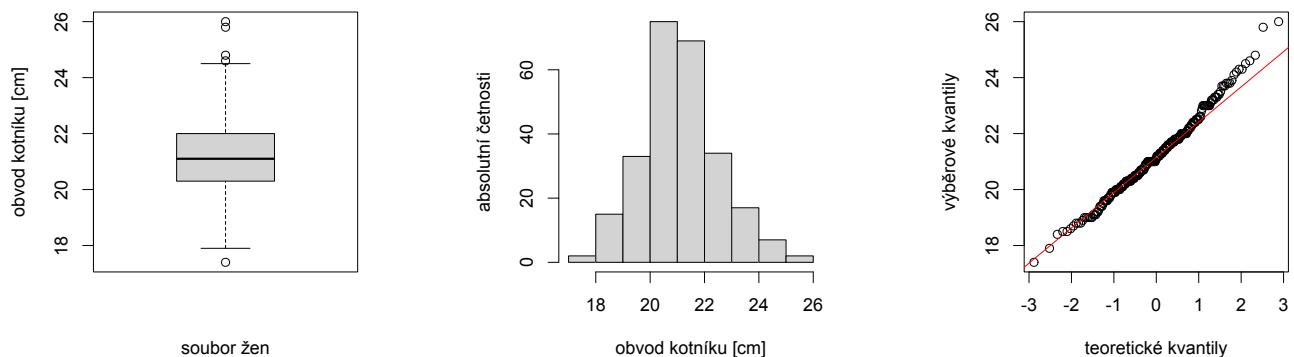
Obrázek 5: Krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod předloktí*

Na obrázku 6 je krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod kolene*. Lze nahlédnout, že proměnná obsahuje dle podmíny (1) několik suspektně odlehlych pozorování (vzhledem ke zbylým pozorováním), a sice pozorování s indexy 44 a 154 (horní odlehly). Může se jednat opět o ženy s gynoidním typem obezity (typ „hrnuška“) v rámci metabolického syndromu, kdy je somatický tuk ukládán v oblasti boků a stehen (i kolen) se vznikem tzv. tukových polštářů.

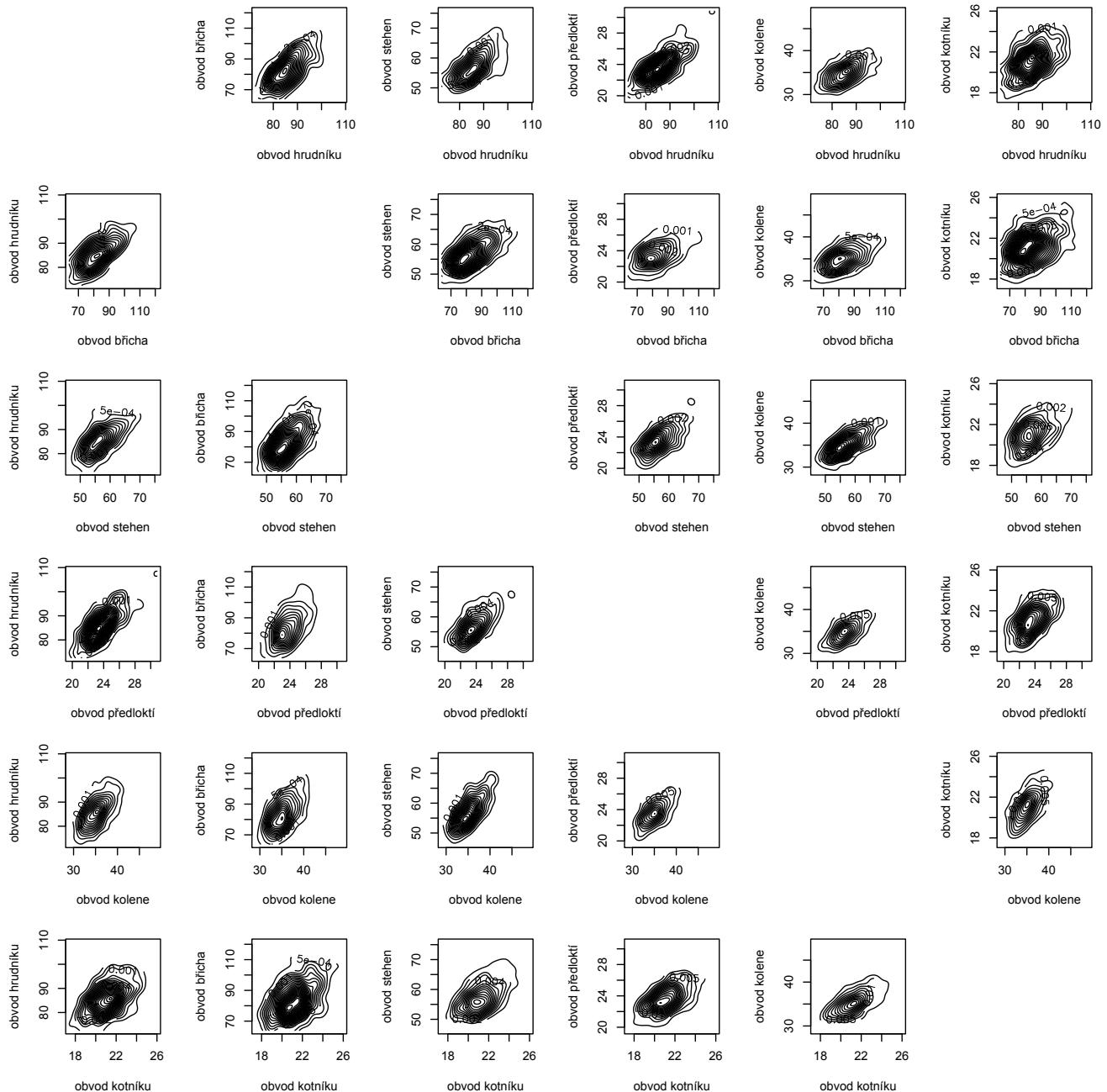


Obrázek 6: Krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod kolene*

Na obrázku 7 je krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod kotníku*. Lze nahlédnout, že proměnná obsahuje dle podmíny (1) několik suspektně odlehlych pozorování (vzhledem ke zbylým pozorováním), a sice pozorování s indexy 166 (dolní odlehly) a 44, 155, 175, 254 (horní odlehly). V případě dolní odlehly (opět žena s indexem 166) může jít o všeobecnou kachexii (pokud nejde o náhodnou chybu měření), v případě horních odlehlych se může jednat typicky o ženy s insuficiencí pravého srdce a perimaleolárními otoky, anebo o ženy s žilní insuficiencí (a perimaleolárními otoky).



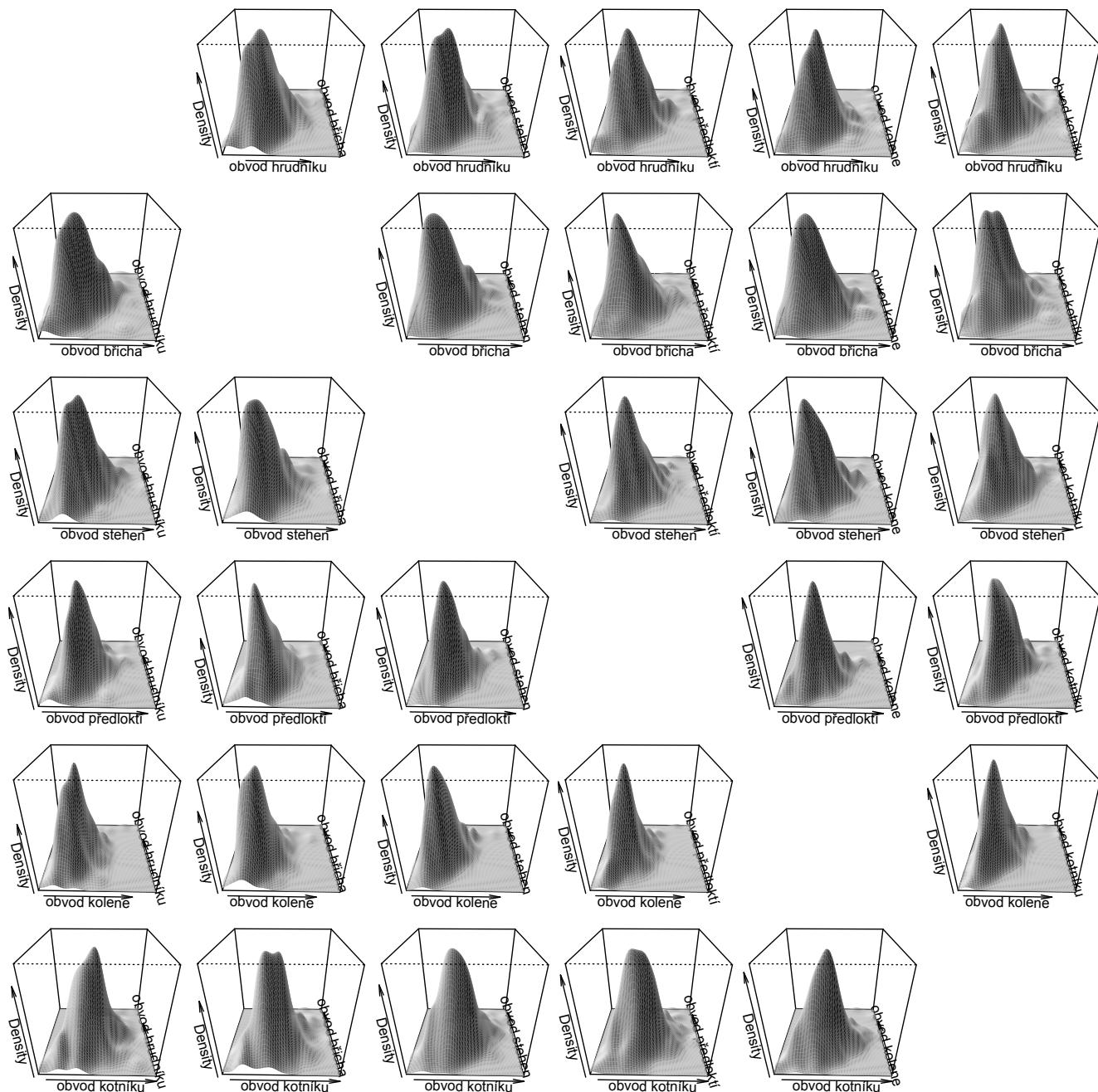
Obrázek 7: Krabicový diagram, histogram a kvantil-kvantil diagram výběrové proměnné *obvod kotníku*



Obrázek 8: Matice diagramů popisujících dvourozměrnou normalitu.

Na obrázcích 8 a 9 můžeme vidět, že každé dvě proměnné v souboru jsou relativně hodně korelované (vidíme typické vrstevnicové elipsy na obrázku 8; v případě nekorelovanosti by byly mapy spíše kruhové). V tabulce 2 nahlédneme, že předpoklad vícerozměrné (bivariantní) normality byl Henze-Zinklerovým testem ve všech dvojicích proměnných zamítnut.

Současně byla spočítána i adjustovaná Mahalanobisova distance a její χ^2 kvantil, čímž bylo identifikován čtrnáct pozorování podezřelých z vícerozměrné odlehlosti. Podle výpočtu jde o ženy s indexy 100, 101, ..., 113. Všechna tato pozorování mají adjustovanou robustní Mahalanobisovu distanci větší než 15,79, což je spočtený kritický χ^2 kvantil.



Obrázek 9: Matice plastických diagramů popisujících dvourozměrnou normalitu.

Analýza hlavních komponent. Analýza hlavních komponent byla vzhledem k relativně velkým rozdílům v rozptylech jednotlivých proměnných v souboru provedena na korelační matici, sestavené z těchto proměnných. Jak je vidět z obrázků 8 a 9 i z kovarianční matice (zde neuvádím), proměnné spolu minimálně po dvou korelují, proto má provedení analýzy hlavních komponent ve smyslu redukce dimenzionality smysl.

V tabulce 4 vidíme, že již samotná první hlavní komponenta vysvětlí 72,4 % celkové variability, zatímco druhá hlavní komponenta vysvětlí pouze dalších 10,3 % variability. To ilustruje i obrázek 10.

Lze tedy uvažovat dokonce pouze jednu (první) hlavní komponentu za účelem vysvětlení významné části celkové variability a přitom velmi dobré redukce dimenzionality. Eventuálně lze přijmout první dvě hlavní komponenty, které dohromady vysvětlí 82,7 % celkové variability. V tabulce 3 pak vidíme, že lineární koeficienty první hlavní komponenty jsou všechny kladné a relativně velké. To lze chápát tak, že první hlavní komponenta silně pozitivně souvisí se všemi šesti měřenými proměnnými; první hlavní komponentu tak lze interpretovat jako společnou vlastnost všech měřených proměnných a tou se zdá být síla vrstvy podkožního somatického tuku.

Naopak z 3 vidíme i to, že druhá hlavní komponenta lineárně pozitivně kombinuje části těla, které nejsou apriorně vybaveny silnou vrstvou podkožního tuku – předloktí, koleno a kotník. Naopak jde však o místa častých patologií a především podkožních edémů (otoků), zvláště u žen, neboť u nich jsou častější lymfatické edémy (lymfedémy) předloktí, (lýtek) a kolen typicky po některých gynekologických operacích zpravidla pro onkologickou primární příčinu. Jde o lymfedém předloktí a paže po ablaci mammy při karcinomu mammy a lymfedém dolní končetiny při gynekologických -ektomických operacích (hysterektomiích, tj. odnětích dělohy).

	ob. hrudníku	ob. břicha	ob. stehen	ob. předloktí	ob. kolene	ob. kotníku
obvod hrudníku		< 0,0001	< 0,0001	< 0,0001	0,0005	0,0215
obvod břicha	< 0,0001		0,0001	0,0001	0,0011	0,0057
obvod stehen	< 0,0001	0,0001		< 0,0001	0,0002	0,0009
obvod předloktí	< 0,0001	0,0001	< 0,0001		0,0011	0,0216
obvod kolene	0,0005	0,0011	0,0002	0,0011		0,0171
obvod kotníku	0,0215	0,0057	0,0009	0,0216	0,0171	

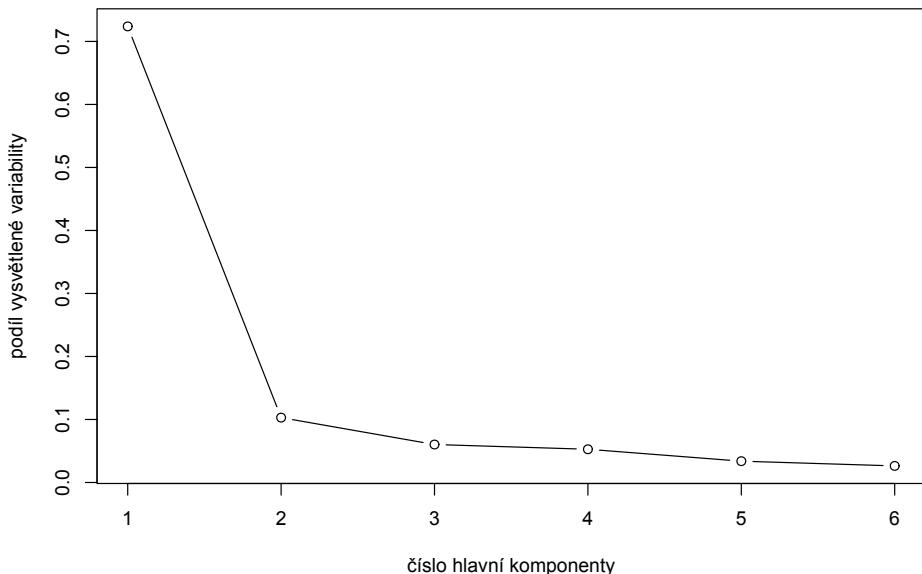
Tabulka 2: Odhadování hladin významnosti (p -hodnot) Henze-Zinklerova testu vícerozměrné normality pro jednotlivé proměnné

	PC1	PC2	PC3	PC4	PC5	PC6
obvod hrudníku	0,412	-0,435	0,420	-0,239	0,047	0,636
obvod břicha	0,398	-0,501	0,057	0,618	-0,233	-0,388
obvod stehen	0,420	-0,047	-0,647	0,047	0,625	0,095
obvod předloktí	0,429	0,033	0,176	-0,637	0,038	-0,614
obvod kolene	0,420	0,328	-0,400	-0,082	-0,702	0,238
obvod kotníku	0,368	0,670	0,458	0,382	0,242	0,043

Tabulka 3: Lineární koeficienty hlavních komponent pro jednotlivé proměnné (PC i je i -tá hlavní komponenta pro $\forall i \in \{1, 2, \dots, 6\}$)

	PC1	PC2	PC3	PC4	PC5	PC6
směrodatná odchylka	2.084	0.786	0.601	0.562	0.451	0.397
proporce vysvětlené variability	0.724	0.103	0.060	0.053	0.034	0.026
kumulativní proporce vysvětlené variability	0.724	0.827	0.887	0.940	0.974	1.000

Tabulka 4: Vlastnosti hlavních komponent (PC i je i -tá hlavní komponenta pro $\forall i \in \{1, 2, \dots, 6\}$)



Obrázek 10: Diagram popisující podíl vysvětlené variability prostřednictvím dané hlavní komponenty

2.3 Závěr

Byla ověřena jedno- i vícerozměrná normálnost vstupních dat. Předpoklad normality by byl u všech měřených proměnných obtížně udržitelný, pokud by data vstoupila do metody citlivé na normální rozložení. Pak by bylo nutné data transformovat, např. logaritmováním, eventuálně použít robustní protějšek citlivé metody.

Byla identifikována pozorování podezřelé z odlehlosti (od zbylých hodnot). Velikost těchto pozorování však nebyla nápadně jiná než ve zbytku výběru, proto se pravděpodobně u těchto pozorování jedná jen o širší variantu normy.

Analýza hlavních komponent ukázala možnost redukce systému proměnných na jednu až dvě lineární kombinace (komponenty) původních proměnných, aniž by došlo k významné ztrátě informace.

3 Appendix

Zde je uveden kód v jazyce R, ve kterém byly zpracovávány veškeré výpočty a rovněž generovány diagramy.

```
"MVN"
),
function(package){

  if(!(package %in% rownames(installed.packages()))){

    install.packages(
      package,
      dependencies = TRUE,
      repos = "http://cran.us.r-project.org"
    )

  }

  library(package, character.only = TRUE)

}

)
)

## -----
#####
## nastavuji handling se zipováním v R -----
Sys.setenv(R_ZIPCMD = "C:/Rtools/bin/zip")

## -----
#####
## nastavuji pracovní složku -----
while(!"_domaci_ukol_1_.R" %in% dir()){
  setwd(choose.dir())
}

mother_working_directory <- getwd()

## -----
#####
## vytvářím posložky pracovní složky -----
setwd(mother_working_directory)

for(my_subdirectory in c("vstupy", "vystupy")){
  if(!file.exists(my_subdirectory)){

    dir.create(file.path(
      mother_working_directory, my_subdirectory
    ))

  }
}

## -----
#####
```

Appendix

```
## loaduji data -----
setwd(
  paste(mother_working_directory, "vstupy", sep = "/")
)

my_data <- data.frame(
  setNames(
    object = read.spss(
      file = "du1_30.sav",
      to.data.frame = TRUE
    ),
    nm = c(
      "id",
      "obvod hrudníku",
      "obvod břicha",
      "obvod stehen",
      "obvod předloktí",
      "obvod kolene",
      "obvod kotníku"
    )
  ),
  check.names = FALSE
)

setwd(mother_working_directory)

## -----
#####
## Exploratory Data Analysis -----
#####

#### nejdříve vytvářím diagram závislostí jednotlivých proměnných mezi
#### sebou ----

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
  file = "scatterplot_matrix.eps",
  width = 8,
  height = 8,
  pointsize = 14
)

par(mar = c(1.1, 1.1, 0.1, 0.1))

pairs(
  my_data[, grep("obvod", colnames(my_data))],
  panel = "panel.smooth"
)

dev.off()

setwd(mother_working_directory)

#####
#### nyní vytvářím pro každou spojitu proměnnou boxplot, histogram
#### a QQ-plot ----

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

for(my_variable in colnames(my_data)[grep("obvod", colnames(my_data))]){


```

```

##### boxploty -----
cairo_ps(
  file = paste(
    gsub(
      " ",
      "_",
      iconv(my_variable, to = "ASCII//TRANSLIT")
    ),
    "_boxplot.eps",
    sep = ""
  ),
  width = 5,
  height = 5,
  pointsize = 18
)

par(mar = c(4.1, 4.1, 0.5, 0.3))

boxplot(
  x = my_data[, my_variable],
  col = "lightgrey",
  xlab = "soubor žen",
  ylab = paste(
    my_variable,
    " [cm]",
    sep = ""
  )
)

dev.off()

##### QQ-ploty -----
cairo_ps(
  file = paste(
    gsub(
      " ",
      "_",
      iconv(my_variable, to = "ASCII//TRANSLIT")
    ),
    "_qqplot.eps",
    sep = ""
  ),
  width = 5,
  height = 5,
  pointsize = 18
)

par(mar = c(4.1, 4.1, 0.5, 0.3))

qqnorm(
  y = my_data[, my_variable],
  xlab = "teoretické kvantily",
  ylab = "výběrové kvantily",
  main = ""
)

qqline(
  y = my_data[, my_variable],
  col = "red"
)

dev.off()

##### histogramy -----
cairo_ps(

```

```

file = paste(
  gsub(
    " ",
    "_",
    iconv(my_variable, to = "ASCII//TRANSLIT")
  ),
  "_histogram.eps",
  sep = ""
),
width = 5,
height = 5,
pointsize = 18
)

par(mar = c(4.1, 4.1, 0.5, 0.3))

hist(
  x = my_data[, my_variable],
  col = "lightgrey",
  xlab = paste(
    my_variable,
    " [cm]",
    sep = ""
  ),
  ylab = "absolutní četnosti",
  main = ""
)

dev.off()

}

setwd(mother_working_directory)

##### do konzole tisknu, která pozorování dané proměnné jsou suspektně
##### odlehlá ----

for(my_variable in colnames(my_data)[grepl("obvod", colnames(my_data))]){

  print("#####")

  print(
    paste(
      "Suspektně odlehlé proměnné ",
      my_variable,
      ":",
      sep = ""
    )
  )

  print("-- dolní odlehlé:")
  print(
    which(
      my_data[, my_variable] < quantile(
        my_data[, my_variable],
        probs = 1/4,
        names = FALSE
      ) - 1.5 * (
        quantile(
          my_data[, my_variable],
          probs = 3/4,
          names = FALSE
        ) - quantile(
          my_data[, my_variable],
          probs = 1/4,
          names = FALSE
        )
      )
    )
  )
}

```

```

)
print("-- horní odlehle:")
print(
  which(
    my_data[, my_variable] > quantile(
      my_data[, my_variable],
      probs = 3/4,
      names = FALSE
    ) + 1.5 * (
      quantile(
        my_data[, my_variable],
        probs = 3/4,
        names = FALSE
      ) - quantile(
        my_data[, my_variable],
        probs = 1/4,
        names = FALSE
      )
    )
  )
)
}

#### počítám p-hodnoty Kolmogorova-Smirnova testu a Shapiro-Wilkova testu ----

my_p_values <- NULL

for(my_variable in colnames(my_data)[grep("obvod", colnames(my_data))]){

  my_p_values <- rbind(
    my_p_values,
    c(
      suppressWarnings(
        ks.test(my_data[, my_variable], y = "pnorm")$p.value
      ),
      shapiro.test(my_data[, my_variable])$p.value
    )
  )

  rownames(my_p_values)[
    dim(my_p_values)[1]
  ] <- my_variable

  colnames(my_p_values) <- c(
    "p_level_kolmogorov",
    "p_level_shapiro"
  )
}

print(
  xtable(
    my_p_values,
    align = rep("", ncol(my_p_values) + 1),
    digits = 4
  ),
  floating = FALSE,
  tabular.environment = "tabular",
  hline.after = NULL,
  include.rownames = TRUE,
  include.colnames = TRUE
)
)

#### vykresluji contour diagramy bivariantní normality -----

```

```

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
  file = "bivariate_normality_matrix.eps",
  width = 10,
  height = 10,
  pointsize = 12
)

par(mar = c(4.1, 4.1, 2.1, 2.1))
par(mfrow = c(6, 6))

for(i in 2:7){

  for(j in 2:7){

    if(i == j){

      plot(
        0,
        type = "n",
        axes = FALSE,
        ann = FALSE
      )

    }else{

      mvnPlot(
        hzTest(my_data[, c(i, j)]),
        type = "contour",
        default = TRUE
      )

    }

  }

}

dev.off()

cairo_ps(
  file = "bivariate_normality_perspective_matrix.eps",
  width = 10,
  height = 10,
  pointsize = 14
)

par(mar = c(0.1, 0.1, 0.1, 0.1))
par(mfrow = c(6, 6))

for(i in 2:7){

  for(j in 2:7){

    if(i == j){

      plot(
        0,
        type = "n",
        axes = FALSE,
        ann = FALSE
      )

    }else{

      mvnPlot(
        hzTest(my_data[, c(i, j)]),
        type = "persp",

```

```

        default = TRUE,
        ylab = "",
        yaxt='n'
    )
}

}

}

dev.off()

setwd(mother_working_directory)

##### zkoumám outliery ----

my_outliers <- mvOutlier(
  my_data[, grep("obvod", colnames(my_data))],
  qqplot = TRUE,
  method = "adj.quan",
  label = TRUE
)

rownames(my_outliers$newData)[as.logical(my_outliers$outlier[, "Outlier"])]
# které indexy jsou vícerozměrně odlehle?

my_outliers$outlier[, "Mahalanobis Distance"][
  as.logical(my_outliers$outlier[, "Outlier"])
] # jak velké mají Mahalanobisovy distance

##### vytvářím tabulku p-hodnot Henze-Zirklerova testu -----
henze_zinkler_p_values <- matrix(rep(0, 6 * 6), nrow = 6)

for(i in 2:7){
  for(j in 2:7){

    if(i == j){

      henze_zinkler_p_values[i - 1, j - 1] <- 1.0

    }else{

      henze_zinkler_p_values[i - 1, j - 1] <- attr(
        hzTest(my_data[, c(i, j)]), "p.value"
      )

    }

  }

}

colnames(henze_zinkler_p_values) <- colnames(my_data)[
  grep("obvod", colnames(my_data))
]

rownames(henze_zinkler_p_values) <- colnames(my_data)[
  grep("obvod", colnames(my_data))
]

print(
  xtable(
    henze_zinkler_p_values,
    align = rep("", ncol(henze_zinkler_p_values) + 1),
    digits = 4
)
)

```

```
  ),
  floating = FALSE,
  tabular.environment = "tabular",
  hline.after = NULL,
  include.rownames = TRUE,
  include.colnames = TRUE
)

## -----
#####
## zkouším PCA -----
##### kovarianční matice -----
cov(my_data[, grep("obvod", colnames(my_data))])

##### modeluji PCA -----
my_pca <- prcomp(
  my_data[, grep("obvod", colnames(my_data))],
  center = TRUE,
  scale. = TRUE
)

##### tisknu sumář PCA -----
summary(my_pca)

my_pca[["rotation"]]

print(
  xtable(
    my_pca[["rotation"]],
    align = rep("", ncol(my_pca[["rotation"]]) + 1),
    digits = 3
  ),
  floating = FALSE,
  tabular.environment = "tabular",
  hline.after = NULL,
  include.rownames = TRUE,
  include.colnames = TRUE
)

print(
  xtable(
    summary(my_pca)[["importance"]],
    align = rep("", ncol(summary(my_pca)[["importance"]]) + 1),
    digits = 3
  ),
  floating = FALSE,
  tabular.environment = "tabular",
  hline.after = NULL,
  include.rownames = TRUE,
  include.colnames = TRUE
)

##### tisknu scree-plot -----
#screeplot(my_pca)

setwd(paste(mother_working_directory, "vystupy", sep = "/"))

cairo_ps(
  file = "scree_plot.eps",
```

```
width = 8,
height = 5,
pointsize = 12
)

par(mar = c(4.1, 4.1, 0.5, 0.3))

plot(
  summary(my_pca)[["importance"]][["Proportion of Variance", ],
  type = "b",
  xlab = "číslo hlavní komponenty",
  ylab = "podíl vysvětlené variability"
)

dev.off()

setwd(mother_working_directory)

## -----  
#####
#####
```

4 Reference

- [1] R CORE TEAM. *R: A Language and Environment for Statistical Computing* [online]. Vienna, Austria: R Foundation for Statistical Computing, 2016. Dostupné z: <https://www.R-project.org/>
- [2] CHAMBERS, J. M., W. S. CLEVELAND, B. KLEINER a P. A. TUKEY. *Graphical Methods for Data Analysis*. B.m.: Wadsworth & Brooks/Cole, 1983.
- [3] BIRNBAUM, Z. W. a Fred H. TINGEY. One-sided confidence contours for probability distribution functions. *The Annals of Mathematical Statistics*. 1951, **22**(4), 592–596.
- [4] ROYSTON, Patrick. An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*. 1982, **4**, 115–124.
- [5] HENZE, N. a B. ZIRKLER. A Class of Invariant Consistent Tests for Multivariate Normality. *Commun. Statist.-Theor. Meth.* 1990, **19**(10).
- [6] GNANADESIKAN, R. a J. R. KETTENRING. Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. *Biometrics* [online]. 1972, **28**(1), 81. Dostupné z: doi:10.2307/2528963
- [7] PETR, Hebák. *Vícerozměrné statistické metody*. Praha: Informatorium, 2004. ISBN 80-7333-036-9.
- [8] ZVÁRA, Karel. *Základy statistiky v prostředí R*. Praha, Česká republika: Karolinum, 2013. ISBN 978-80-246-2245-3.