Úkol 1 - Průzkumová analýza dat a metoda hlavních komponent

4ST512 Vícerozměrná statistika

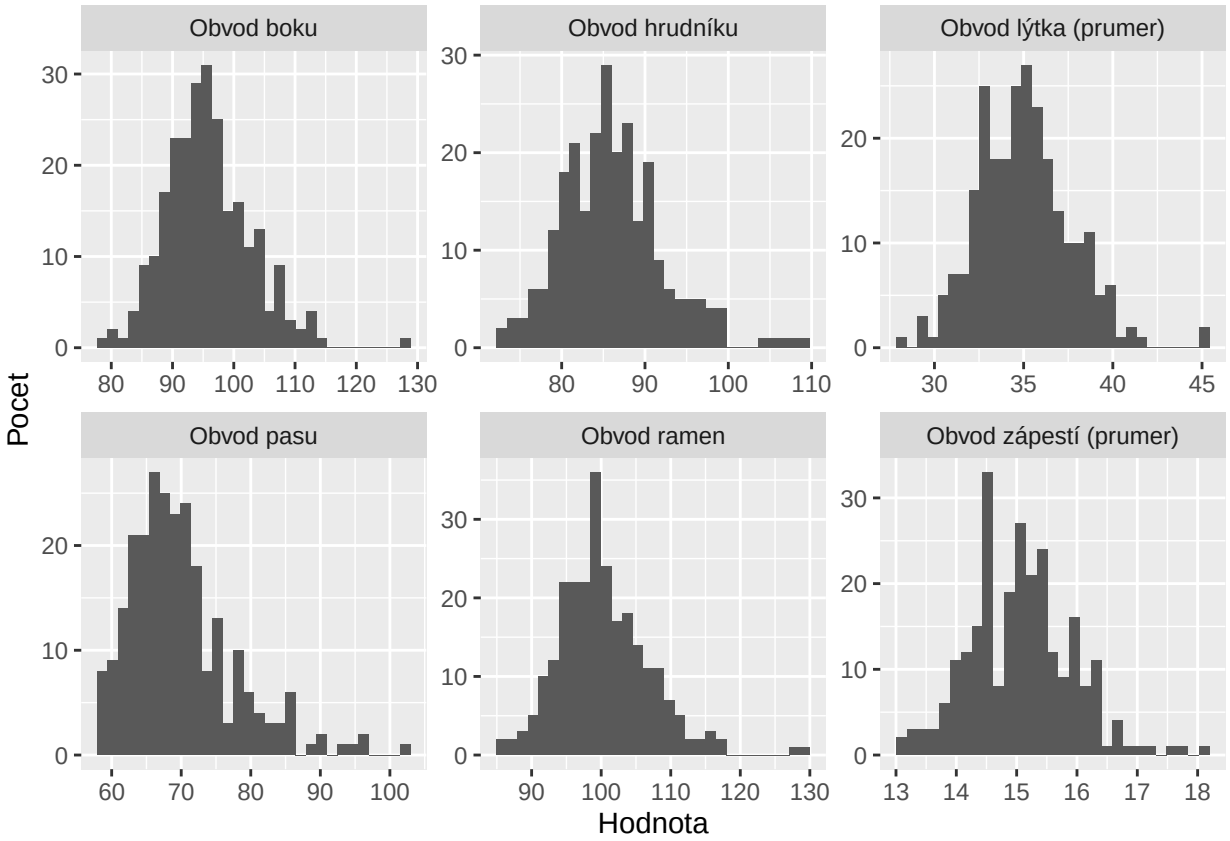
Jiří Filip

Úvod

V této práci budeme analyzovat 6 údajů (obvod ramen, hrudníku, pasu, boků, lýtek a zápěstí) 254 žen. Nejprve tato data prozkoumáme s ohledem na marginální a sdruženou pravděpodobnost, identifikujeme případné odchylky od normality a odlehlá pozorování. Následně na datech spustíme metodu hlavních komponent a pokusíme se interpretovat její výsledky.

Posouzení normality jednotlivých proměnných

Podíváme se na marginální rozdělení všech proměnných. Začneme nejprve histogramy.



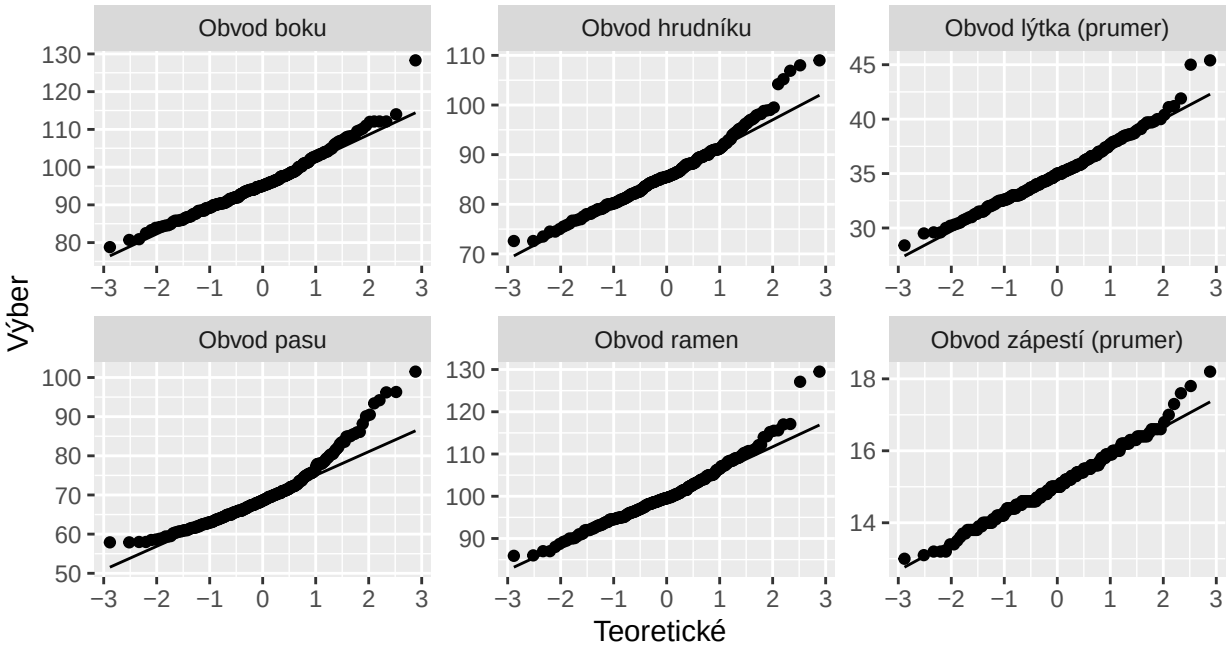
Můžeme vidět, že jednotlivé veličiny se zhruba řídí normálním rozdělením, u obvodu pasu je toto rozdělení však značně zešikmeté. Lze se proto debatovat, zda by se toto rozdělení dalo vůbec nazvat normálním.

U všech veličin je ale jasně patrné, že se zde vyskytuje hned několik odlehlých pozorování, většinou na pravé straně rozdělení. To by znamenalo, že extrémy zde příliš nejsou, když se jedná o malé obvody, ale vyskytují se spíše u obvodů velkých. Mohlo by to znamenat, že zde máme spíše obezní než podvyživené jedince.

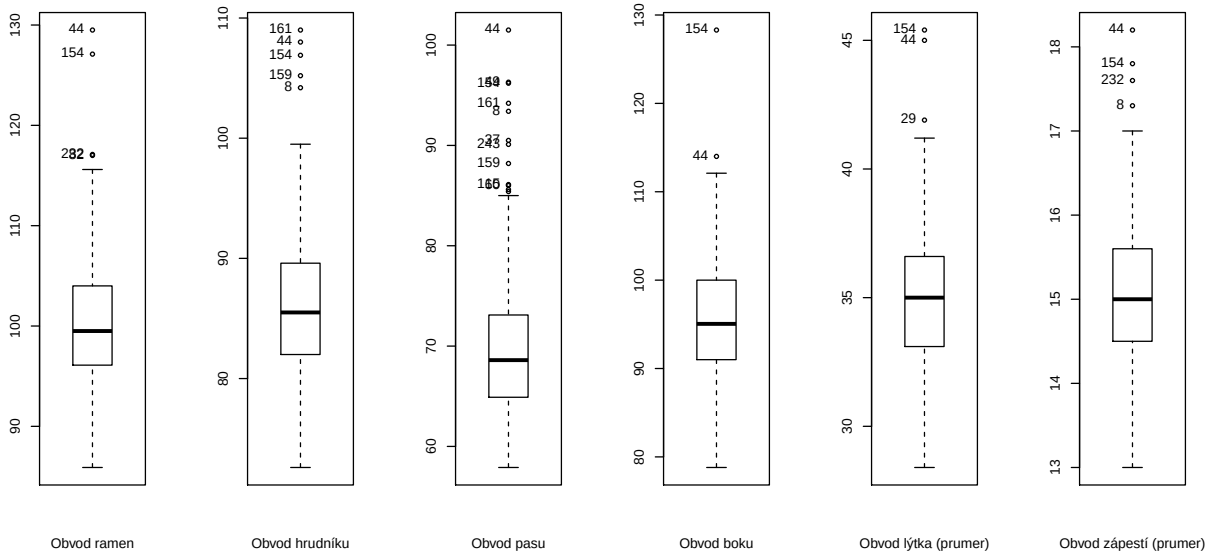
Nyní, když víme, že pozorování se zhruba řídí normálním rozdělením, zbývá identifikovat odlehlá pozorování a zjistit, zda se jedná o chybná měření, nebo zda se takoví jedinci opravdu mohou v populaci vyskytovat.

Pokud se na histogramy podíváme, vidíme, že odlehlé hodnoty nejsou tak extrémní, že by je nebylo možné vysvětlit např. obezitou (např. obvod pasu kolem 100 cm nebo obvod ramen kolem 130 cm). To by naznačovalo, že se o chybná měření nejedná.

Podívejme se ještě na Q-Q graf jednotlivých veličin, z něhož jsou odlehlá pozorování rovněž patrná. U obvodu pasu je znovu vidět, že zešikmení narušuje normalitu.



Pro samotnou identifikaci odlehlých pozorování použijme boxplot. Označme si jednotlivá pozorování jejich identifikačním číslem.

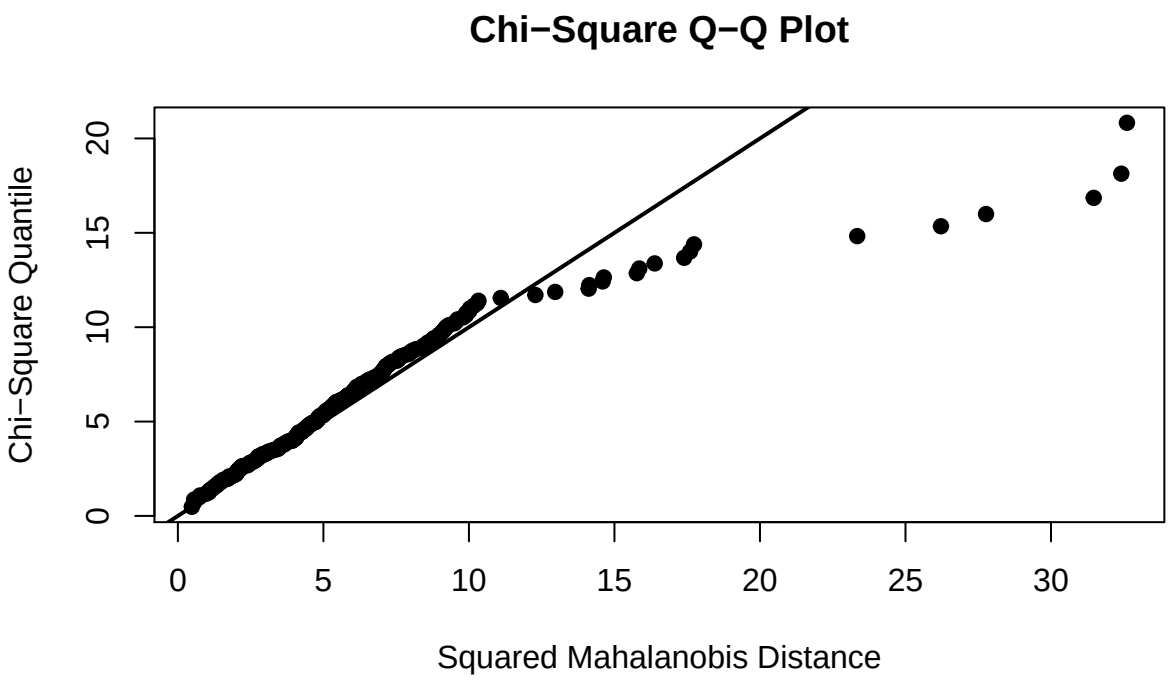


Z boxplotu snadněji vyčteme, která pozorování jsou přesně odlehlá. Např. pozorování č. 44 je odlehlé ve všech boxplotech, patrně se jedná o obézní ženu. Stejně tak žena č. 8 (která není uvedená jako odlehlá ve všech, ale ve většině ano). Pozorování č. 161 má velký obvod hrudníku a pasu, mohlo by se jednat o těhotnou ženu.

Lze vidět, že obvod pasu má zdaleka nejvíce odlehlých hodnot. Toto by se dalo vysvětlit právě těhotenstvím. Např. žena č. 37 se nevyskytuje jako odlehlá hodnota u ostatních proměnných. Posouzení vícerozměrné normality a odlehlých hodnot však provedeme pomocí chí-kvadrát diagramu.

Posouzení vícerozměrné normality

Podívejme se na chi-kvadrát diagram k posouzení vícerozměrné normality.



Odhlehlá pozorování jsou následující: Z004, Z008, Z037, Z044, Z049, Z095, Z154, Z160, Z183, Z222, Z232, Z237, Z243, Z248, Z253

Mohlo by se jednat jak o těhotné (např. žena č. 37), tak o obézní ženy (č. 44).

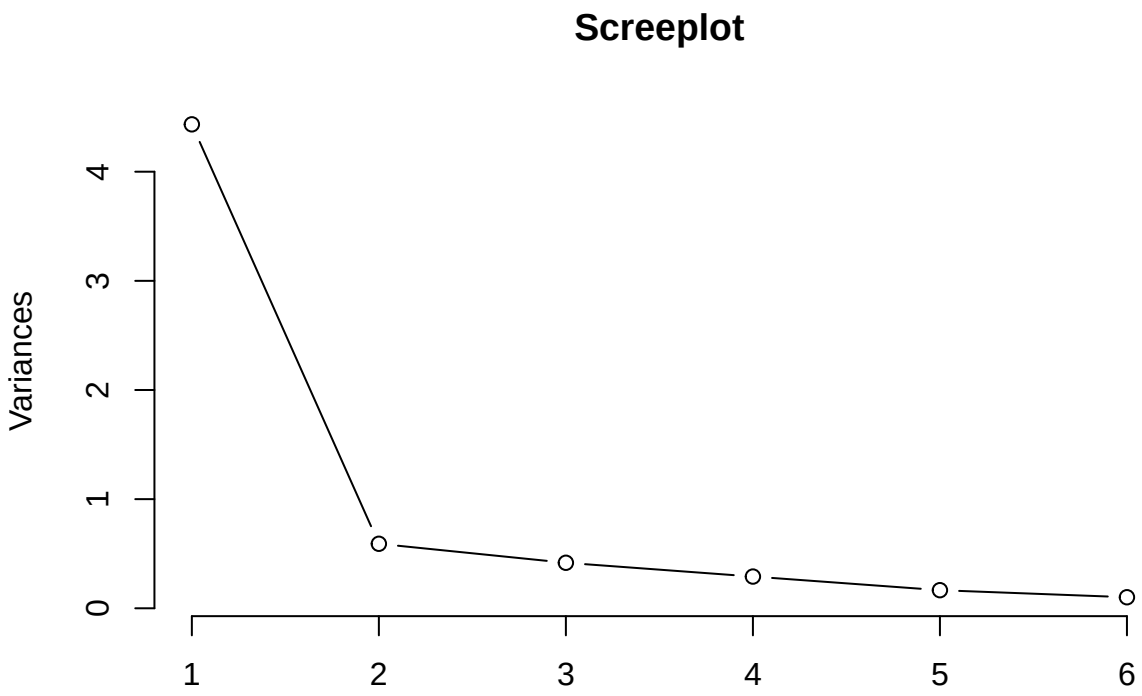
Kromě odlehlých pozorování lze na grafu vidět, že obvody se řídí vícerozměrným normálním rozdělením (až na odlehlé hodnoty, které značně vybočují).

Metoda hlavních komponent

Pomocí metody hlavních komponent budeme analyzovat korelační matici (proměnné nemají stejnou škálu, průměry a rozptyly se liší). Lze vidět, že již první komponenta vysvětluje 73,9 % variability původních dat. První tři komponenty pak vysvětlují přes 90 % celkové variability. Řekněme tedy, že data mají dimenzionalitu 3.

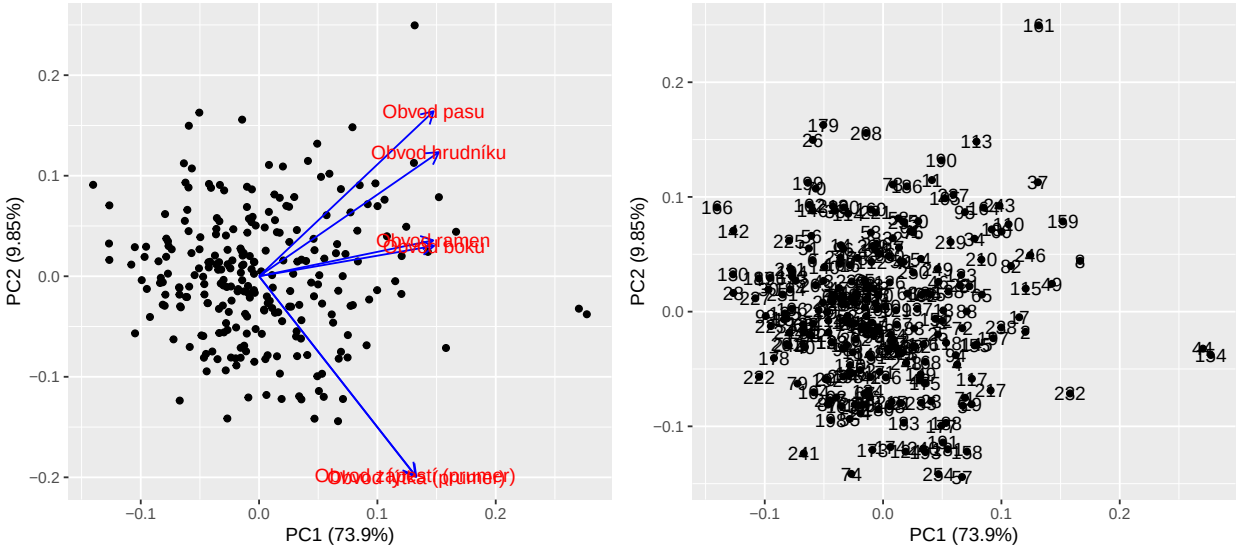
```
## Importance of components:
##                           PC1     PC2    PC3     PC4     PC5     PC6
## Standard deviation     2.106 0.76881 0.6458 0.53920 0.40749 0.31824
## Proportion of Variance 0.739 0.09851 0.0695 0.04846 0.02768 0.01688
## Cumulative Proportion  0.739 0.83749 0.9070 0.95544 0.98312 1.00000
```

Na screeplotu si prohlédněme variabilitu v závislosti na hlavní komponentě. Můžeme vidět, že první 3 komponenty vysvětlují většinu původních proměnných.



Na dvou obrázcích dole si můžeme prohlédnout projekci do prostoru založeného na prvních dvou hlavních komponentách. Kromě toho lze vidět jednotlivé zátěže, které nám ukazují promítnutí původních proměnných do podprostoru hlavních komponent 1 a 2. Ze zátěží rovněž vyčteme, že rozměry jsou mezi sebou závislé (obvod ramen a obvod boků, pasu a hrudníku, zápěstí a lýtek). Dává smysl, že obvod ramen nám dává informaci o obvodu boků, stejně tak u dalších dvou závislých dvojic.

Druhý z obrázků obsahuje i identifikační číslo ženy. Můžeme vidět odlehlá pozorování (např. žena č. 44, 161 nebo 232).



Nedá se jednoznačně říci, že by hlavní komponenty šly nějak přesně interpretovat. Přesto se o to pokusme. Všechny proměnné mají velmi podobnou korelaci s první hlavní komponentou (okolo 0,4). Můžeme tak první hlavní komponentu interpretovat jako "mohutnost" subjektu.

Druhá komponenta je silně (circa -0,56) negativně korelována s obvodem lýtek a zápěstí, zatímco je zde nezanedbatelná korelace s obvodem pasu a hrudníku.

Když tedy PC2 roste, klesá obvod lýtek a zápěstí, zatímco se zvyšuje obvod pasu a hrudníku. Nabízí se možnost uvažovat například nad těhotenstvím, i když to zní poněkud absurdně. PC2 by mohlo vyjadřovat "stádium těhotenství" (např. počet měsíců), kdy by při rostoucím počtu měsíce ženě rostl pas a hrudník (prsa). Nesmíme ale zapomenout na negativní korelaci s obvodem lýtek a zápěstí. S rostoucí dobou těhotenství by se potom tyto dva obvody musely snižovat, což zní poněkud nepravděpodobně.

Třetí komponenta je ještě rozporuplnější. Silně koreluje s obvodem boků a lýtek, naopak záporně s obvodem ramen, hrudníku a zápěstí. S velkou nadsázkou bychom mohli interpretovat třetí komponentu jako "mohutnost dolní poloviny těla", ale i tak by to nevysvětlilo, proč by tato "mohutnost" negativně korelovala s obvodem ramen, hrudníku a zápěstí.

```
##                              PC1         PC2        PC3         PC4
## Obvod ramen            0.4188555  0.10165836 -0.4111865  0.58626428
## Obvod hrudníku         0.4323497  0.35124480 -0.2499071  0.06084086
## Obvod pasu             0.4197767  0.46706811  0.1864739 -0.18872693
## Obvod boků             0.4209387  0.08610293  0.4820640 -0.39223125
## Obvod lýtka (průměr)   0.3781602 -0.56818032  0.4978600  0.44168464
## Obvod zápěstí (průměr) 0.3758016 -0.56382395 -0.5034376 -0.51772972
##                                PC5         PC6
## Obvod ramen            -0.48209190  0.26274118
## Obvod hrudníku          0.35146044 -0.70712180
## Obvod pasu              0.40625228  0.60844359
## Obvod boků             -0.61983811 -0.21205334
## Obvod lýtka (průměr)    0.29933491 -0.04018374
## Obvod zápěstí (průměr)  0.07225905  0.11899952
```