UNICORN UNIVERSITY

Datová matice a transformace dat

2. seminář k předmětu ***Statistické metody v analýze dat***

5.10.2022

Martina Šimková

ZS 2022/2023

Úvodní informace

- Ing. Martina Šimková, Ph.D.
  - simkova.martinka@gmail.com
  - martina.simkova.1@unicorncollege.cz

- Bodové hodnocení:
  - 20 bodů – aktivita na cvičení
  - 30 bodů – 2 testy (á 15 bodů)
  - 50 bodů – závěrečná zkouška

Shrnutí poznatků z přednášky

- Datová matice
  - dvourozměrná tabulka popisující jednotlivá pozorování (objekty) v řádcích a ve sloupcích jsou naměřené hodnoty proměnných (ukazatelů)
- Výběrová rozdělení
  - je třeba rozlišit, kdy pracujeme s celou populací (základní soubor) a kdy pracujeme s výběrem
  - pouze **náhodný výběr** je základem pro realizaci statistických úsudků
  - základní výběrové charakteristiky:
    - Hustota pravděpodobnosti
    - Výběrové průměry
    - Kovarianční matice
    - Korelační matice
  - Obvykle předpokládáme **normální rozdělení** proměnných
- Lineární transformace dat
- Vzdálenosti objektů
- Testování normality – jednorozměrné, vícerozměrné

Instalace balíčků do R

```
install.packages("pastecs")
install.packages("ggpubr")
install.packages("ggplot2")
install.packages("robustHD")
install.packages("MVN")
install.packages("mvnormtest"
```

Datová matice

UNICORN UNIVERSITY

Datová matice – teoretický příklad

STATISTICKÉ PROMĚNNÉ (ZNAKY)

STATISTICKÉ OBJEKTY

	x1	x2	x3	x4	x5	x6	...
1							...
2							...
3							...
4			VÝBĚROVÝ				...
5			SOUBOR				...
6							...
7							...
8							...
9							...
...	...	...	...	...	...	...	...

Datová matice – praktický příklad

- Náhodný výběr 15 pozorování ze 100 šetřených osob

ID osoby	Pohlaví	Věk	Vzdělání	Rodinný stav	Ekonomická aktivita	Roční hrubý příjem	Subj. hodn. zdrav.stavu
021	1	32	3	1	1	294 882	1
003	2	29	4	1	1	460 705	1
019	1	61	1	4	3	164 532	4
078	1	52	2	2	1	399 375	2
100	2	48	2	3	1	1 040 412	1
091	1	36	3	2	1	384 000	1
040	2	31	3	2	1	300 000	1
056	2	54	4	4	1	264 000	2
061	2	40	1	3	2	86 515	2
032	1	34	4	2	1	377 597	1
011	2	37	4	1	1	372 965	3
084	2	58	3	2	1	189 913	4
075	2	29	4	2	1	159 018	1
085	1	35	4	2	1	225 693	1
093	2	46	2	3	2	57 547	2

- Je nutné si vždy ujasnit, o jaký typ proměnné se jedná.

PŘÍKLAD 1

V datovém souboru **MMDA_02_data.xlsx** jsou některé zajímavé ukazatele za několik amerických měst.

Vypočítejte pro uvedené ukazatele základní výběrové charakteristiky:

- Výběrový průměr
- Výběrovou směrodatnou odchylku
- Výběrový rozptyl
- Výběrovou kovarianční matici
- Výběrovou korelační matici

Výpočty proveďte v MS Excel a následně ověřte v R (soubor **MMDA_02_data.csv**).

PŘÍKLAD 1

- Výběrové průměry $\bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$
- Výběrové směrodatné odchylky s_j $s_j = \sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x_j})^2}{n-1}}$
- Výběrové rozptyly s_j^2 $s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x_j})^2}{n-1}$
- Wishartova matice Q $\boldsymbol{Q} = \sum_{i=1}^{n}(\boldsymbol{x_i}-\bar{\boldsymbol{x}})\,(\boldsymbol{x_i}-\bar{\boldsymbol{x}})^T$
- Výběrová kovarianční matice S $\boldsymbol{S} = \frac{1}{n-1}\boldsymbol{Q}$
  - na diagonále výběrové rozptyly, mimo diagonálu výběrové kovariance
- Výběrová korelační matice R
  - na diagonále jedničky, mimo diagonálu korelační koeficienty: $r(x_j, x_{j'}) = \frac{Q(x_j, x_{j'})}{\sqrt{Q(x_j)Q(x_{j'})}}$

PŘÍKLAD 1

V datovém souboru **MMDA_02_data.xlsx** jsou některé zajímavé ukazatele za několik amerických měst.

Import dat do R: File → Import Datasets → From text (base) → MMDA_02_data.csv

(!!! nutno mít krátkou cestu k datům)

```
# Výběrové charakteristiky
install.packages("pastecs")
library(pastecs)
stat.desc(MMDA_02_data)

x <- as.matrix(MMDA_02_data[,2:8])
one <- as.matrix(rep(1, dim(x)[1]))
n <- dim(x)[1]
xbar <- 1/n*t(x)%*%one

mean(MMDA_02_data$Rainfall)

cov_x <- cov(x)
cor_x <- cor(x)
```

Exploratorní analýza dat

- **IDENTIFIKACE ODLEHLÝCH A EXTRÉMNÍCH POZOROVÁNÍ**
- Ověřování předpokladů, především tzv. testy normality
- Náhrada chybějících hodnot
- Třídění dat do intervalů

Identifikace odlehlých a extrémních pozorování

- Grafická analýza
- Míry polohy – kvantily
- Lineární transformace proměnných ve výběru
- Vzdálenosti objektů

Lineární transformace proměnných ve výběru

- **CENTROVANÉ PROMĚNNÉ** – od všech pozorování odečteme výběrový průměr

$$x_{ij,C} = x_{ij} - \bar{x}_j, \qquad i = 1,2,\dots,n, \qquad j = 1,2,\dots,p.$$

- **NORMOVANÉ PROMĚNNÉ** – centrované proměnné vydělíme směrodatnou odchylkou

$$x_{ij,N} = \frac{x_{ij} - \bar{x}_j}{s(x_j)} = z_j, \qquad i = 1,2,\dots,n, \qquad j = 1,2,\dots,p.$$

- **LIBOVOLNÁ LINEÁRNÍ KOMBINACE PROMĚNNÝCH** – např. umělé proměnné

$$u_i = \boldsymbol{c}^T \boldsymbol{x}_i = \sum_{j=1}^{p} c_j x_{ij}\,, \qquad i = 1,2,\dots,n.$$

PŘÍKLAD 2

V datovém souboru **MMDA_02_data.xlsx** jsou některé zajímavé ukazatele za několik amerických měst.

Transformujte tyto ukazatele centrováním a normováním.

Vzdálenosti objektů

- Míra vzdálenosti mezi i-tým a i‘-tým objektem v datové matici: $d_j(i;i') = x_{ij} - x_{i'j}$
- **EUKLIDOVSKÁ VZDÁLENOST:**

$$D_E(i;i') = \sqrt{\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2}$$

- **NORMOVANÁ VZDÁLENOST:**

$$D_N(i;i') = \sqrt{\sum_{j=1}^{p}(z_{ij}-z_{i'j})^2}$$

- **MAHALANOBISOVA VZDÁLENOST:**

$$D_M(i;i') = \boldsymbol{d}^T\boldsymbol{S}^{-1}\boldsymbol{d} = (\boldsymbol{x}_i - \boldsymbol{x}_{i'})^T\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \boldsymbol{x}_{i'})$$

PŘÍKLAD 3

V datovém souboru **MMDA_02_data.xlsx** jsou některé zajímavé ukazatele za několik amerických měst.

Vypočítejte Mahalanobisovy vzdálenosti.

PŘÍKLAD 3

V datovém souboru **MMDA_02_data.xlsx** jsou některé zajímavé ukazatele za několik amerických měst.

Vypočítejte Mahalanobisovy vzdálenosti.

```
# Mahalanobisova vzdálenost
require(graphics)
vzdalenost_mh <- mahalanobis(x, xbar, cov_x, inverted = FALSE)
plot(density(vzdalenost_mh, bw = 0.3), main="Squared Mahalanobis distances, n=428, p=9"); rug(vzdalenost_mh)

# ulozeni do csv
write.table(vzdalenost_mh, „c:/R/Data_UCL_MGR_MMDA/vzdalenost_mh.txt", sep="\t",dec = ",")
```

PŘÍKLAD 4

V datovém souboru **MMDA_02_data.xlsx** jsou některé zajímavé ukazatele za několik amerických měst. Najděte města, která můžeme z hlediska jedné či více proměnných považovat za odlehlá pozorování. Využijte k tomu všechna kritéria, která znáte, včetně grafické analýzy.

- Normované proměnné (větší než 2)
- Mahalanobisova vzdálenost (>12 nebo TK: F-rozdělení → p-value < α)

$$F_2 = \frac{(n-p)n}{(n^2-1)p} D^2 \qquad F_2 \sim F(p, n-p)$$

Exploratorní analýza dat

- Identifikace odlehlých a extrémních pozorování ✓
- **OVĚŘOVÁNÍ PŘEDPOKLADŮ, PŘEDEVŠÍM TZV. TESTY NORMALITY**
- Náhrada chybějících hodnot
- Třídění dat do intervalů

Testy normality

- Některé statistické metody předpokládají výběr z normálního rozdělení
- Testování **jednorozměrné** normality:

  H_0: *normalita*

  H_1: *non* H_0

  - Chí-kvadrát test dobré shody
  - Kolmogorovův test
  - Shapiro-Wilk test
  - Testy založené na šikmosti a špičatosti
  - Grafické posouzení jednorozměrné normality pomocí grafu výběrové distribuční funkce, porovnání výběrových a teoretických kvantilů (Q-Q diagram) apod.

- Testování **vícerozměrné** normality:
  - Vícerozměrný test Shapiro-Wilk
  - Chí-kvadrát diagram – grafické ověření dvourozměrné normality pomocí srovnání Mahalanobisových vzdáleností s kvantily rozdělení Chí-kvadrát

PŘÍKLAD 5

V datovém souboru **MMDA_02_data.xlsx** jsou některé zajímavé ukazatele za několik amerických měst.

1. Otestujte normalitu jednotlivých proměnných pomocí těchto testů:
   - Graf hustoty pravděpodobnosti
   - Q-Q graf
   - Shapiro-Wilkův test
   - Kolmogorovův test

2. Posuďte, jestli ukazatele mají vícerozměrné normální rozdělení, pomocí:
   - Vícerozměrného Shapiro-Wilkova testu
   - Chí-kvadrát diagramu

PŘÍKLAD 5

```
# Testy jednorozměrné normality
    # 1.grafické nástroje
        # Q-Q graf pro normalitu
        qqnorm(MMDA_02_data$Mortality, pch = 1, frame = TRUE)
        qqline(MMDA_02_data$Mortality, col = "steelblue", lwd = 2)

        # Hustota pravdepodobnosti
        hist(MMDA_02_data$Mortality,probability=TRUE)
        lines(density(MMDA_02_data$Mortality),col="red")

    # 2.statistické nástroje
        # Shapiro-Wilkův test
        shapiro.test(MMDA_02_data$Mortality)

        # Kolmogorov-Smirnov test
        ks.test(MMDA_02_data$Mortality, "pnorm", mean=mean(MMDA_02_data$Mortality), sd=sd(MMDA_02_data$Mortality))

# Testy vícerozměrné normality
    # Shapiro-Wilkův test
    mshapiro.test(t(x))

    # Q-Q graf chisq-mahalanobis
    qqplot(qchisq(ppoints(60), df = 6), vzdalenost_mh)
    abline(0, 1, col = 'gray')
```

PŘÍKLAD 6 – AKTIVITA ZA 2 BODY

V datovém souboru **MMDA_02_data.xlsx** jsou některé zajímavé ukazatele za několik amerických měst.

1. Identifikujte odlehlá pozorování u proměnných **x4** a **x5**. Jak se změní výběrový průměr a výběrová směrodatná odchylka, pokud u těchto proměnných odlehlé pozorování vyloučíte?

2. Otestujte normalitu proměnných **x2** a **x3** pomocí známých grafů a testů.

Exploratorní analýza dat

- Identifikace odlehlých a extrémních pozorování ✓
- Ověřování předpokladů, především tzv. testy normality ✓
- **NÁHRADA CHYBĚJÍCÍCH HODNOT**
- Třídění dat do intervalů

Náhrada chybějících hodnot

- Když datová matice není kompletní
- Pokud u objektu chybí větší počet údajů, lze tento objekt vypustit
- Hodnoty, které jsou náhodně nevyplněné lze doplnit uměle (zmírnění ztráty informace)
- Umělé náhrady (odhad):
  - **Průměrem** příslušné proměnné nebo příslušného objektu (nejjednodušší, nerespektuje však variabilitu ani korelační strukturu dat)
  - **Náhodným číslem** z rozdělení příslušné proměnné s parametry odhadnutými z výběru
  - **Regresí** – odhad založený na regresní rovnici
- Software: umožňuje při hodnocení dvojice proměnných vyloučit jen ty řádky, které se přímo týkají aspoň jedné z proměnných (bez ohledu na to, že v jiných sloupcích použitých řádků některé údaje chybí)
- Obvyklé kódování: 9; 99; 999; -99 apod.

Exploratorní analýza dat

- Identifikace odlehlých a extrémních pozorování ✓
- Ověřování předpokladů, především tzv. testy normality ✓
- Náhrada chybějících hodnot ✓
- **TŘÍDĚNÍ DAT DO INTERVALŮ**

Třídění dat do intervalů - opakování

- Přechod od původních hodnot kvantitativní proměnné k intervalovému rozdělení četností → přechod k ordinální proměnné = sloupcový vektor v datové matici obsahuje pořadová čísla intervalů
- Neexistuje obecný předpis pro stanovení počtu a délky intervalů
  - Ztráta informace (málo intervalů) x ztráta přehlednosti (mnoho intervalů)
- např. Sturgessovo pravidlo pro počet intervalů K:

$$K = log_2(2n) \cong 1 + 3{,}3 \log n$$

  - Délka intervalů (H) nejčastěji stejná, např. podle variačního rozpětí (R):

$$H \approx \frac{R}{K}$$

Dotazy?

UNICORN UNIVERSITY

UNICORN
UNIVERSITY