Multivariate Statistical Analysis - Exercise Session 5

10.02.2022

Problem 1: Frequency analysis

First we introduce some notation. Consider a sample of size n described by two variables, x with categories $A_1, \dots, A_J$ and y with categories $B_1, \dots, B_K$. Let n_{jk} denote the number of observations having categories A_j and B_j. Let

$$f_{jk} = \frac{n_{jk}}{n}$$

denote the *relative frequencies* corresponding to n_{jk}. Additionally, we denote

$$n_{j.} = \sum_{k=1}^{K} n_{jk}, \quad n_{.k} = \sum_{j=1}^{J} n_{jk},$$

and similarly,

$$f_{j.} = \sum_{k=1}^{K} f_{jk}, \quad f_{.k} = \sum_{j=1}^{J} f_{jk}.$$

First we read the data and import required packages.

```
library(ggplot2)
library(reshape2)

data <- read.table("SCIENCEDOCTORATES.txt", header = TRUE, row.names = 1)
```

Below we show the *contigency table* of the data. For further analysis let us remove `Total` row and `Total` column. With command `proportions` we can calculate relative frequencies, row profiles and column profiles. Also, function `proportions` requires matrix as an input.

```
data
```

```
##                Y1960 Y1965 Y1970 Y1971 Y1972 Y1973 Y1974 Y1975  Total
## Engineering      794  2073  3432  3495  3475  3338  3144  2959  22710
## Mathematics      291   685  1222  1236  1281  1222  1196  1149   8282
## Physics          530  1046  1655  1740  1635  1590  1334  1293  10823
## Chemistry       1078  1444  2234  2204  2011  1849  1792  1762  14374
## EarthSciences    253   375   511   550   580   577   570   556   3972
## Biology         1245  1963  3360  3633  3580  3636  3473  3498  24388
## Agriculture      414   576   803   900   855   853   830   904   6135
## Psychology       772   954  1888  2116  2262  2444  2587  2749  15772
## Sociology        162   239   504   583   638   599   645   680   4050
## Economics        341   538   826   791   863   907   833   867   5966
## Anthropology      69    82   217   240   260   324   381   385   1958
## Others           314   502  1079  1392  1500  1609  1531  1550   9477
## Total           6263 10477 17731 18880 18940 18948 18316 18352 127907
```

```
data <- as.matrix(data[-13, -9]) # Remove total row and total column
```

Below we calculate the *relative frequencies*.

```
f <- proportions(data)
all(data / sum(data) == f)
```

```
## [1] TRUE
```

```
round(f, 2)
```

```
##                Y1960 Y1965 Y1970 Y1971 Y1972 Y1973 Y1974 Y1975
## Engineering     0.01  0.02  0.03  0.03  0.03  0.03  0.02  0.02
## Mathematics     0.00  0.01  0.01  0.01  0.01  0.01  0.01  0.01
## Physics         0.00  0.01  0.01  0.01  0.01  0.01  0.01  0.01
## Chemistry       0.01  0.01  0.02  0.02  0.02  0.01  0.01  0.01
## EarthSciences   0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
## Biology         0.01  0.02  0.03  0.03  0.03  0.03  0.03  0.03
## Agriculture     0.00  0.00  0.01  0.01  0.01  0.01  0.01  0.01
## Psychology      0.01  0.01  0.01  0.02  0.02  0.02  0.02  0.02
## Sociology       0.00  0.00  0.00  0.00  0.00  0.00  0.01  0.01
## Economics       0.00  0.00  0.01  0.01  0.01  0.01  0.01  0.01
## Anthropology    0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00
## Others          0.00  0.00  0.01  0.01  0.01  0.01  0.01  0.01
```

Next we calculate the *row profiles*. They are given by

$$f_{k|j} = \frac{n_{jk}}{n_{j.}} = \frac{f_{jk}}{f_{j.}}.$$

Frequency $f_{k|j}$ estimates the probability

$$P(y \in B_k | x \in A_j).$$

```
rowp <- proportions(data, 1)
all(rowp == sweep(data, 1, rowSums(data), "/"))
```

```
## [1] TRUE
```

```
round(rowp, 2)
```

```
##                Y1960 Y1965 Y1970 Y1971 Y1972 Y1973 Y1974 Y1975
## Engineering     0.03  0.09  0.15  0.15  0.15  0.15  0.14  0.13
## Mathematics     0.04  0.08  0.15  0.15  0.15  0.15  0.14  0.14
## Physics         0.05  0.10  0.15  0.16  0.15  0.15  0.12  0.12
## Chemistry       0.07  0.10  0.16  0.15  0.14  0.13  0.12  0.12
## EarthSciences   0.06  0.09  0.13  0.14  0.15  0.15  0.14  0.14
## Biology         0.05  0.08  0.14  0.15  0.15  0.15  0.14  0.14
## Agriculture     0.07  0.09  0.13  0.15  0.14  0.14  0.14  0.15
## Psychology      0.05  0.06  0.12  0.13  0.14  0.15  0.16  0.17
## Sociology       0.04  0.06  0.12  0.14  0.16  0.15  0.16  0.17
## Economics       0.06  0.09  0.14  0.13  0.14  0.15  0.14  0.15
## Anthropology    0.04  0.04  0.11  0.12  0.13  0.17  0.19  0.20
## Others          0.03  0.05  0.11  0.15  0.16  0.17  0.16  0.16
```

Column profiles can be calculated similarly to row profiles. They are given by

$$f_{j|k} = \frac{n_{jk}}{n_{.k}} = \frac{f_{jk}}{f_{.k}}.$$

Frequency $f_{j|k}$ estimates the probability

$$P(x \in A_j | y \in B_k).$$

```
colp <- proportions(data, 2)
all(colp == sweep(data, 2, colSums(data), "/"))
```

```
## [1] TRUE
```

```
round(colp, 2)
```

```
##                Y1960 Y1965 Y1970 Y1971 Y1972 Y1973 Y1974 Y1975
## Engineering     0.13  0.20  0.19  0.19  0.18  0.18  0.17  0.16
## Mathematics    0.05  0.07  0.07  0.07  0.07  0.06  0.07  0.06
## Physics        0.08  0.10  0.09  0.09  0.09  0.08  0.07  0.07
## Chemistry      0.17  0.14  0.13  0.12  0.11  0.10  0.10  0.10
## EarthSciences  0.04  0.04  0.03  0.03  0.03  0.03  0.03  0.03
## Biology        0.20  0.19  0.19  0.19  0.19  0.19  0.19  0.19
## Agriculture    0.07  0.05  0.05  0.05  0.05  0.05  0.05  0.05
## Psychology     0.12  0.09  0.11  0.11  0.12  0.13  0.14  0.15
## Sociology      0.03  0.02  0.03  0.03  0.03  0.03  0.04  0.04
## Economics      0.05  0.05  0.05  0.04  0.05  0.05  0.05  0.05
## Anthropology   0.01  0.01  0.01  0.01  0.01  0.02  0.02  0.02
## Others         0.05  0.05  0.06  0.07  0.08  0.08  0.08  0.08
```

Now let us calculate the *attraction repulsion matrix* D. Elements of the matrix are given by

$$d_{jk} = \frac{n_{jk}}{n^*_{jk}} = \frac{f_{jk}}{f^*_{jk}},$$

where $n^*_{jk} = \frac{n_{j.} n_{.k}}{n}$ and $f^*_{jk} = f_j f_k$.

```
# v1 and v2 are same as total row and total column
v1 <- matrix(rowSums(data), ncol = 1)
v2 <- matrix(colSums(data), nrow = 1)

# Expected number of observations under independence
e <- (v1 %*% v2) / sum(data) # n_{jk}^{*} from above

# We obtain attraction repulsion matrix D
# simply dividing each n_{jk} by n_{jk}^{*}
d <- data / e

# Values near 1: The year and science are independent
# Values < 1: The science is less frequent in that specific year
# Values > 1: The science is more frequent in that specific year
round(d, 2)
```

```
##                Y1960 Y1965 Y1970 Y1971 Y1972 Y1973 Y1974 Y1975
## Engineering     0.71  1.11  1.09  1.04  1.03  0.99  0.97  0.91
## Mathematics    0.72  1.01  1.06  1.01  1.04  1.00  1.01  0.97
## Physics        1.00  1.18  1.10  1.09  1.02  0.99  0.86  0.83
## Chemistry      1.53  1.23  1.12  1.04  0.94  0.87  0.87  0.85
## EarthSciences  1.30  1.15  0.93  0.94  0.99  0.98  1.00  0.98
## Biology        1.04  0.98  0.99  1.01  0.99  1.01  0.99  1.00
## Agriculture    1.38  1.15  0.94  0.99  0.94  0.94  0.94  1.03
## Psychology     1.00  0.74  0.86  0.91  0.97  1.05  1.15  1.21
```

```
## Sociology      0.82  0.72  0.90  0.98  1.06  1.00  1.11  1.17
## Economics      1.17  1.10  1.00  0.90  0.98  1.03  0.98  1.01
## Anthropology   0.72  0.51  0.80  0.83  0.90  1.12  1.36  1.37
## Others         0.68  0.65  0.82  1.00  1.07  1.15  1.13  1.14
```

Next we visualize attraction repulsion matrix as a heatmap. First we transform the data in appropriate form for the `ggplot2` with the `melt` function from the package `reshape2`.

```
melted <- melt(d, varnames = c("science", "year"), value.name = "ar")
head(melted)
```

```
##         science  year        ar
## 1   Engineering Y1960 0.7140280
## 2   Mathematics Y1960 0.7175789
## 3       Physics Y1960 1.0000924
## 4     Chemistry Y1960 1.5316270
## 5 EarthSciences Y1960 1.3008379
## 6       Biology Y1960 1.0425696
```

Now we are ready to plot the heatmap of attraction repulsion matrix with `ggplot2`.

```
ggplot(melted, aes(x = science, y = year, fill = ar)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = median(melted$ar), limits = range(melted$ar),
                       name = "AR value") +
  coord_fixed(ratio = 1) +
  labs(x = "Science", y = "Year") +
  theme(axis.text.x = element_text(angle = 45, size = 9, vjust = 1, hjust = 1),
        panel.background = element_blank())
```

Problem 2: Covariance matrix

Let x be a p-variate continuous random variable. Show that $\mathrm{Cov}[x]$ is positive semi-definite.

Recall the definition of positive semi-definitess:

Definition 1. A symmetric and real-valued $p \times p$ matrix A is said to be positive semi-definite if the scalar $a^T A a$ is nonnegative for every real-valued column vector $a \in \mathbb{R}^p$.

Now,

$$\begin{aligned}
&a^T \mathrm{Cov}[x] a \\
&= a^T \mathbb{E}\left[(x - \mathbb{E}[x])(x - \mathbb{E}[x])^T\right] a \\
&= \mathbb{E}\left[\underbrace{a^T (x - \mathbb{E}[x])}_{=y \in \mathbb{R}} (x - \mathbb{E}[x])^T a\right] \\
&= \mathbb{E}\left[yy^T\right] = \mathbb{E}[y^2] \geq 0.
\end{aligned}$$

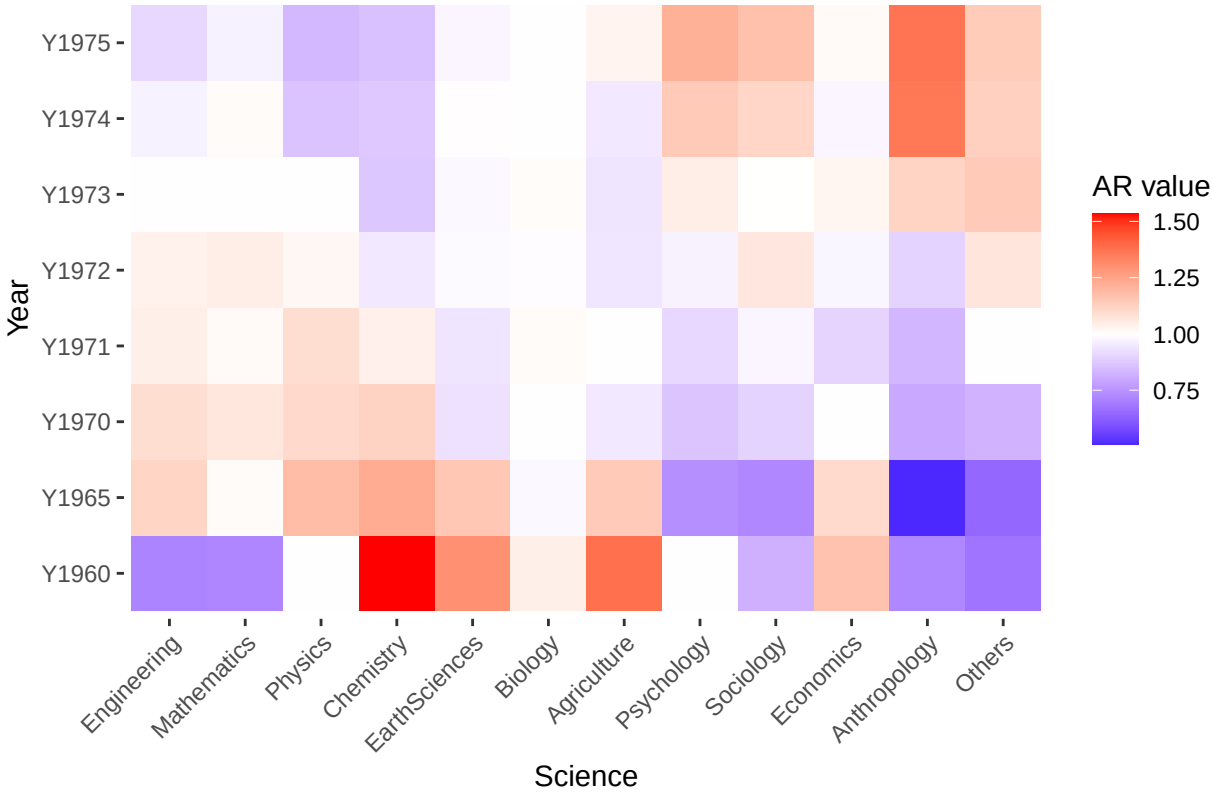

Figure 1: Heatmap of the attraction repulsion matrix.