Chapter 5 Workshop

Table of contents

Dataset `Toxaemia`

This dataset is from the `vcdExtra` package. Two signs of *toxaemia*, an abnormal condition during pregnancy characterized by high blood pressure (hypertension) and high levels of protein in the urine. If untreated, both the mother and baby are at risk of complications or death. The dataset `Toxaemia` represents 13384 expectant mothers in Bradford, England in their first pregnancy, who were also classified according to social class and the number of cigarettes smoked per day.

The dataset is a a 5 x 3 x 2 x 2 contingency table, with 60 observations on the following 5 variables:

`class` - Social class of mother, a factor with levels: 1, 2, 3, 4, 5

`smoke` - Cigarettes smoked per day during pregnancy, a factor with levels: 0, 1-19, 20+

`hyper` - Hypertension level, a factor with levels: Low, High

`urea` - Protein urea level, a factor with levels: Low, High

`Freq` - frequency in each cell, a numeric vector

Exercise 5.1

Obtain relevant graphical displays for this dataset.

Bar charts-

```
library(tidyverse)

library(vcdExtra)
data(Toxaemia)

Toxaemia |>
  ggplot() +
  aes(x=smoke, y=Freq, fill=hyper) +
  geom_bar(stat='identity')
```

```
Toxaemia |>
  ggplot() +
  aes(x=smoke, y=Freq, fill=hyper) +
  geom_bar(stat='identity',
           position = "dodge"
           )
```

```
Toxaemia |>
  ggplot() +
  aes(x=smoke, y=Freq, fill=hyper) +
  geom_bar(stat ='identity',
           position = "dodge") +
  facet_grid(urea ~ ., scales = "free")
```

Mosaic type charts

```
tab.data <- xtabs(Freq ~ smoke + hyper + urea, data=Toxaemia)

plot(tab.data)
```

```r
mosaic(tab.data, shade=TRUE, legend=TRUE)
```

```r
assoc(tab.data, shade=TRUE)
```

```r
strucplot(tab.data)
```

```r
sieve(tab.data)
```

Exercise 5.2

The genetic information of an organism is stored in its Deoxyribonucleic acid (DNA). DNA is a double stranded helix made up of four different nucleotides. These nucleotides differ in which of the four bases Adenine (A), Guanine (G), Cytosine (C), or Thymine (T) they contain. A simple pattern that we may want to detect in a DNA sequence is that of the nucleotide at position `i+1` based on the nucleotide at position `i`. The nucleotide positional data collected by a researcher in a particular case is given in the following table:

`i\(i+1)`	A	C	G	T
A	622	316	328	536
C	428	262	204	306
G	354	294	174	266
T	396	330	382	648

Perform a test of association and then obtain the symmetric plot.

```r
tabledata <- data.frame(
  A = c(622, 428, 354, 396),
  C = c(316, 262, 294, 330),
  G = c(328, 204, 174, 382),
  T = c(536, 306, 266, 648),
  row.names = c("A", "C", "G", "T")
  )
```

```r
chisq.test(tabledata)$exp
chisq.test(tabledata)
chisq.test(tabledata, simulate.p.value = T)

library(MASS)
corresp(tabledata)
plot(corresp(tabledata, nf=2))
abline(v=0)
abline(h=0)
```

```r
#or
library(FactoMineR)
CA(tabledata)
```

- More R code examples are here