# Descriptive statistics: a review of some basic concepts

MDT - Master Digital Transformation

May 29, 2020

# Descriptive statistics versus statistical inference

**Descriptive statistics** mainly relies on the process of summarizing and organizing the data so they can be easily understood.

**Inferential statistics** relies on the process of analysing a sample of data and using it to draw inferences about the population from which it was drawn.

# Some data classifications

Data can be classified in several ways. For example, based on whether the data are measured on a numerical scale or not we have

- **Quantitative data** which are observations measured on a numerical scale.
  - **discrete data**(e.g. number of car accidents in different Italian cities)
  - **continuous data** (e.g. blood pressure of patients of a hospital)
- **Qualitative or categorical data** which are data which can only be classified into one of the groups of categories.
  - **nominal data** (e.g. marital status of people in a community)
  - **ordinal data** which have category that should be listed in a specific order (e.g. levels of satisfaction of customers of a given product)

# Frequencies

One of the common methods for organizing data is to construct frequency tables.
A **(absolute) frequency** is the number of *units* in each category on the scale of measurement, i.e. it is the number of units taking that value (corresponding to that category, or to that interval of values).

Example. The numbers of accidents experienced by 80 machinists in a certain industry over a period of one year were found to be as shown below.

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2 | 0 | 0 | 1 | 0 | 3 | 0 | 6 | 0 | 0 | 8 | 0 | 2 | 0 | 1 |
| 5 | 1 | 0 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 |
| 0 | 0 | 0 | 5 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| 0 | 3 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 2 | 0 | 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | | | | | | | | | | |

How we can organize these data in a *frequency table*?

# Frequency table I

In a frequency table, value (or category of interval of values) are reported along with the corresponding frequency.

Example.The frequency table for the number of accidents is

| Number of accidents | Frequency |
|---|---|
| 0 | 55 |
| 1 | 14 |
| 2 | 5 |
| 3 | 2 |
| 5 | 2 |
| 6 | 1 |
| 8 | 1 |
| | 80 |

# Frequency table II

In general, for data $x_1, \ldots, x_m$ with corresponding frequencies $n_1, \ldots, n_m$, and $\sum_{i=1}^{m} n_i = N$, a frequencies table is

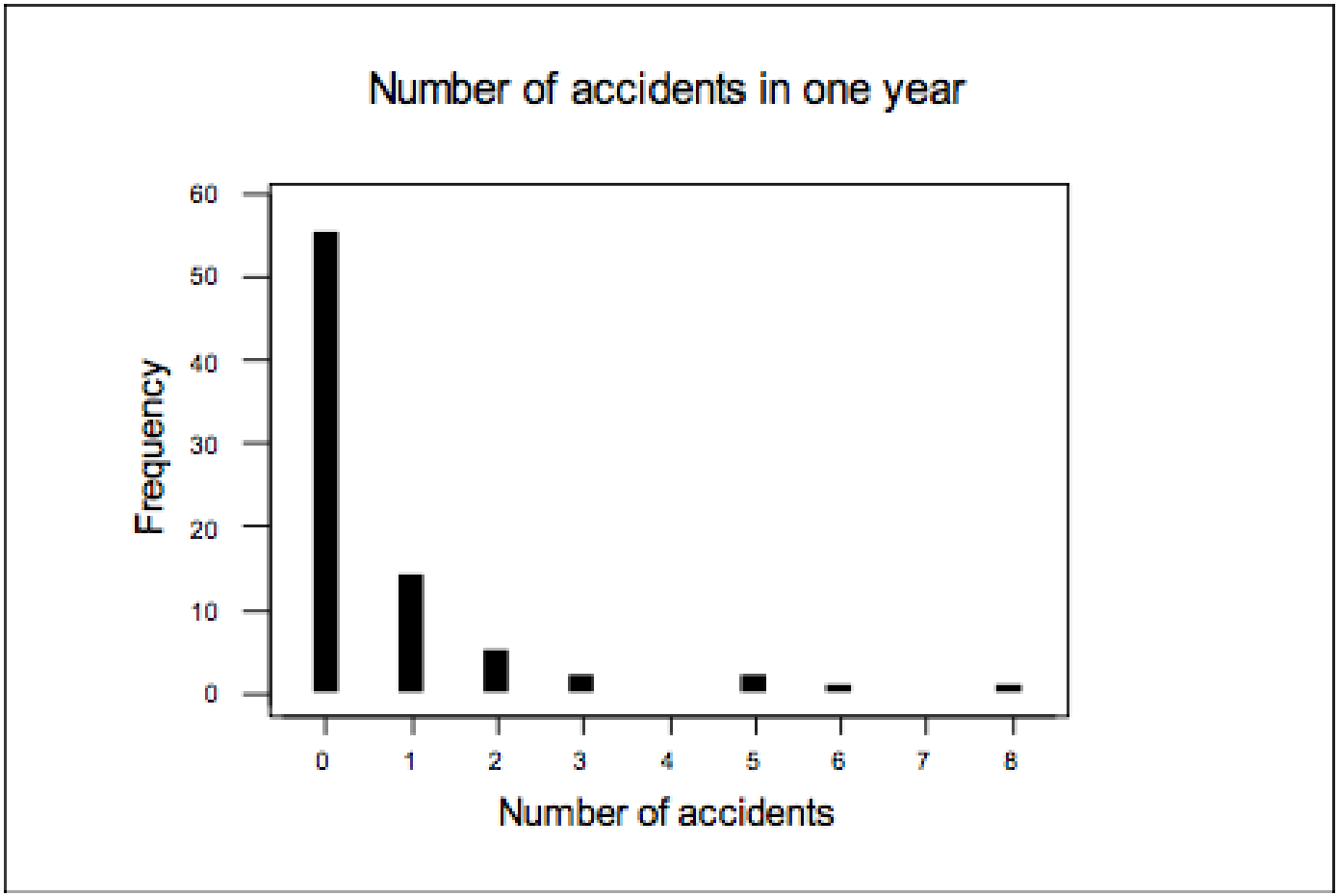
| $x_i$ | $n_i$ |
|---|---|
| $x_1$ | $n_1$ |
| $x_2$ | $n_2$ |
| $\vdots$ | $\vdots$ |
| $x_m$ | $n_m$ |
| | $N$ |

Relative frequency for $x_i$ is given by $\frac{n_i}{N}$, percentage frequency is $\frac{n_i}{N}100\%$, while for a given $k \leq m$, $n_1 + \ldots n_k$ is the cumulative frequency corresponding to $x_k$.

Frequency table can be constructed also for qualitative data.

# Bar chart

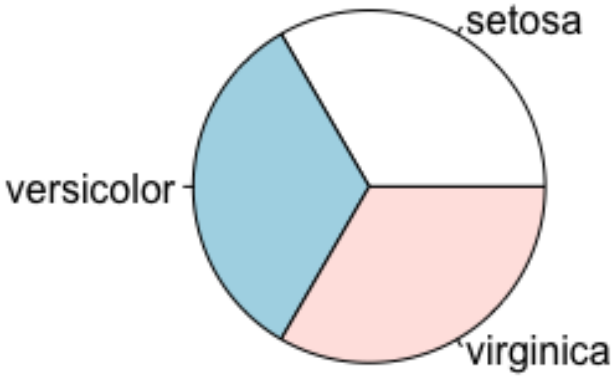
A **bar chart** consists of bars corresponding to each of the possible values (or categories), whose heights are equal to the frequencies.



# Pie chart

**Pie charts** are especially useful for presenting categorical data. The pie "slices" are drawn such that they have an area proportional to the frequency.

Example. The (Fisher's or Anderson's) iris data set gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris: setosa, versicolor, and virginica.



# Steam-and-leaf diagram I

In a steam-and-leaf diagram, each observation is separated into a *stem* consisting of all but the final digit and a *leaf*, the final digit. The stems are placed in a vertical column with the smallest at the top, and each leaf is in the row to the right of its stem, in increasing order out from the stem.

Example. Consider the following set of 80 data points which are compressive strengths in pounds per square inch of 80 specimens of a new aluminum-lithium alloy undergoing evaluation.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 105 | 97 | 245 | 163 | 207 | 134 | 218 | 199 | 160 | 196 |
| 221 | 154 | 228 | 131 | 180 | 178 | 157 | 151 | 175 | 201 |
| 183 | 153 | 174 | 154 | 190 | 76 | 101 | 142 | 149 | 200 |
| 186 | 174 | 199 | 115 | 193 | 167 | 171 | 163 | 87 | 176 |
| 121 | 120 | 181 | 160 | 194 | 184 | 165 | 145 | 160 | 150 |
| 181 | 168 | 158 | 208 | 133 | 135 | 172 | 171 | 237 | 170 |
| 180 | 167 | 176 | 158 | 156 | 229 | 158 | 148 | 150 | 118 |
| 143 | 141 | 110 | 133 | 123 | 146 | 169 | 158 | 135 | 149 |

# Steam-and-leaf diagram II

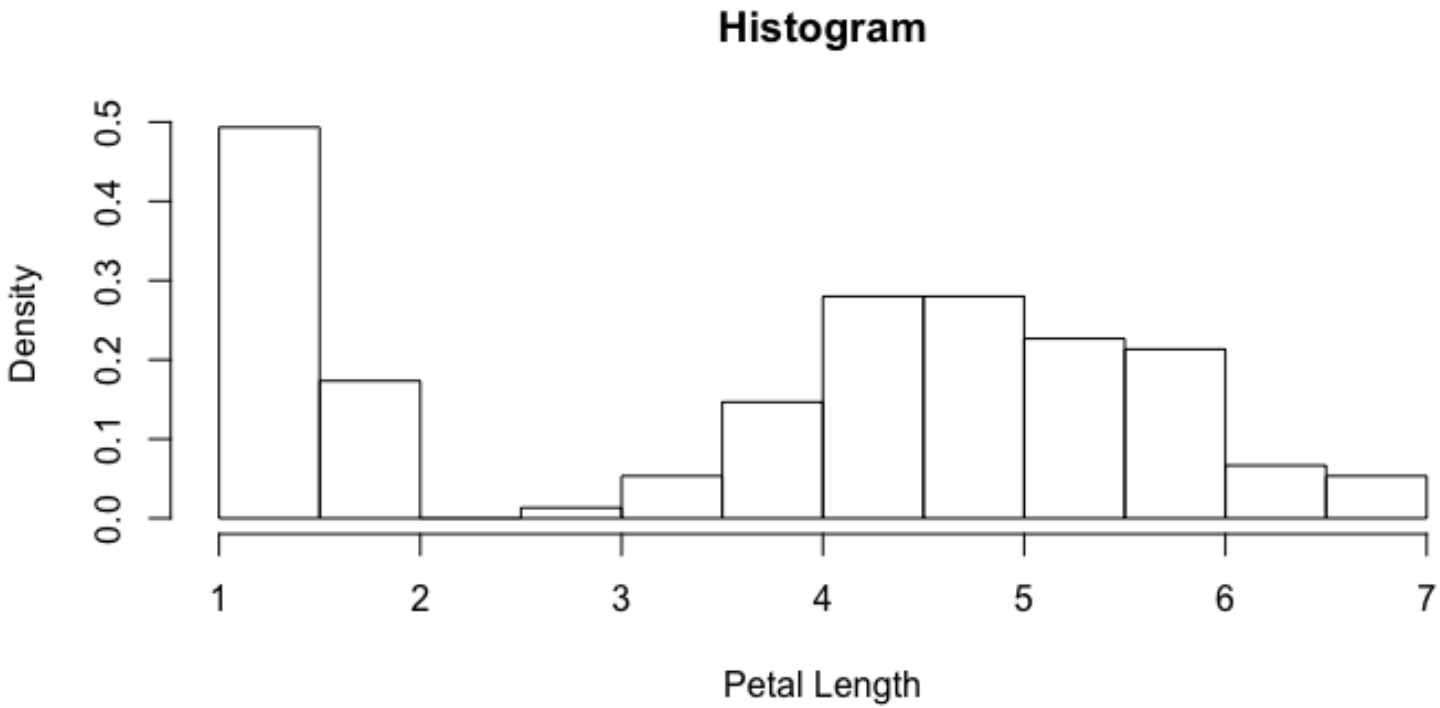
The steam-and-leaf plot is

```
 7 | 6
 8 | 7
 9 | 7
10 | 15
11 | 058
12 | 013
13 | 133455
14 | 12356899
15 | 001344678888
16 | 0003357789
17 | 0112445668
18 | 0011346
19 | 034699
20 | 0178
21 | 8
22 | 189
23 | 7
24 | 5
```

# Histogram

When the data are continuous (or discrete with many values), the histogram is a useful representation of the frequency distribution. To construct a histogram, the first step is to divide the range of values into a series of intervals. The bins are specified as consecutive, non-overlapping intervals, with non necessarily equal sizes. The area of each bin is equal to the frequency of the intervals.

Example. Example of histogram using iris dataset.



# Some measures of central tendency I

Given a set of data $x_1, x_2, \ldots, x_m$, the (arithmetic) **mean** is defined as

$$\frac{x_1 + \ldots + x_m}{m} = \frac{\sum_{i=1}^{m} x_i}{m}.$$

Letting $n_i$ be the frequency corresponding to $x_i$, and $N = \sum_{i=1}^{m} n_i$

$$\frac{\sum_{i=1}^{m} x_i n_i}{N}.$$

Example. For the number of accidents the mean is equal to 0.657

# Some measures of central tendency I

Given a set of data $x_1, x_2, \ldots, x_m$, the **median** is the middle number of the ordered dataset. If the dataset has an even number of elements, then the median is the average of the middle two numbers.

Example. Consider the weights in Kg of a group of 7 people

$$75.3, \quad 82.1 \quad 64.8, \quad 76.3, \quad 81.8, \quad 90.1, \quad 74.2.$$

To find the median, data are placed in order as

$$64.8, \quad 74.2, \quad 75.3, \quad 76.3, \quad 81.8, \quad 82.1, \quad 90.1,$$

and the median is 76.3

Example. For the accident data the median is 0.

The median is the *second quartile* of the distribution.

The **first quartile** $q_1$ is the middle number of the half of the data below the median, and the **third quartile** $q_3$ is the middle number of the half of the data above the median.

# Some measures of central tendency II

The **mode** is the most frequently value or category of the data set.

A dataset can have more than one mode or no mode.

Example. For the accident data the mode is 0.

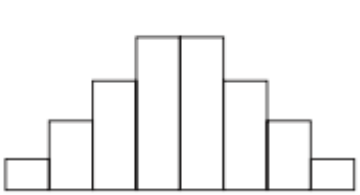
Mean can be defined just for quantitative data. Mode can be defined also for categorical (both ordinal and nominal) data, while median can be defined (with due modifications) for categorical ordinal data.

Mean is sensitive to the presence of outliers, while median is more *robust*. Finally, mode ignores outliers.

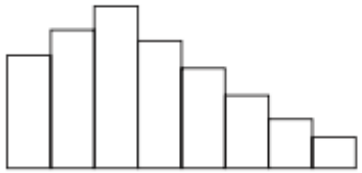
# Some measures of central tendency III

If the distribution is symmetric then the mean is equal to the median.

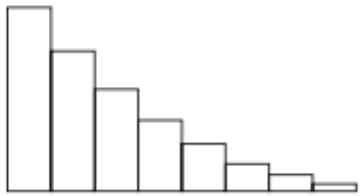
If, in addition, the distribution is unimodal, then mean = median = mode.



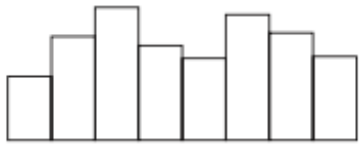
*Symmetrical* or *Bell Shaped*
e.g. exam results



*Positively* (or right) *Skewed*
e.g. earnings of people in the UK



*Reverse J Shaped*
e.g. lifetimes of light bulbs



*Bimodal* (i.e. twin peaks)
e.g. heights of 14 yr old boys and girls

# Some measures of spread

- Given $m$ observations $x_1, \ldots, x_m$, the **range** is

$$x_{(m)} - x_{(1)},$$

where $x_{(1)} = \min\{x_1, \ldots, x_m\}$, and $x_{(m)} = \max\{x_1, \ldots, x_m\}$.

- Given $m$ observations $x_1, \ldots, x_m$, the **variance** is

$$\sigma_X^2 = \frac{1}{m} \sum_i (x_i - \mu_X)^2,$$

where $\mu_X$ stands for the mean of $x_1, \ldots, x_m$ or, letting $n_i$ being the frequency corresponding to $x_i$ and setting $N = \sum_{i=1}^{m} n_i$.

$$\sigma_X^2 = \frac{1}{N} \sum_{i=1}^{n} (x_i - \mu_X)^2 n_i.$$

The **standard deviation** is defined as $\sqrt{\sigma^2}$.

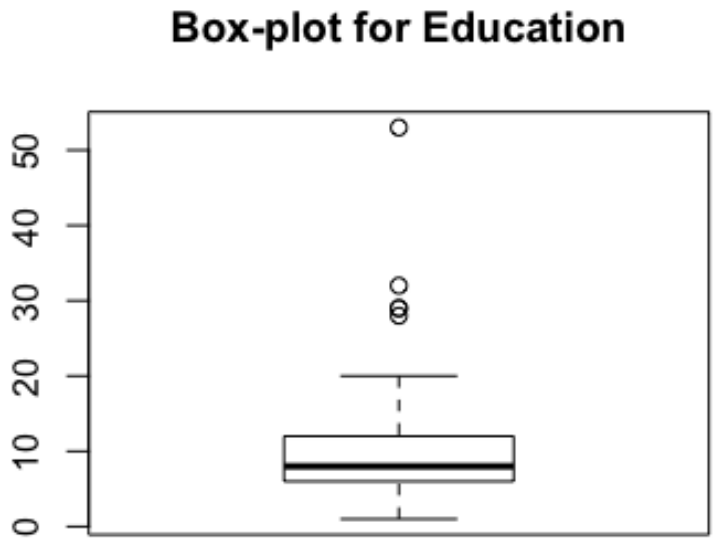
- Letting $q_1$ and $q_3$ respectively denoting the first and third quartiles, the **interquartile range** is

$$IQ = q_3 - q_1.$$

# Box-plot

**Box-plots** represent the five number summary: minimum, lower quartile, median, upper quartile, maximum. Larger possible whiskers are placed at $q_1 - 1.5 \times IQ$ and $q_3 + 1.5 \times IQ$, while values out past $q_1 - 1.5 \times IQ$ or $q_3 + 1.5 \times IQ$ are *outliers*

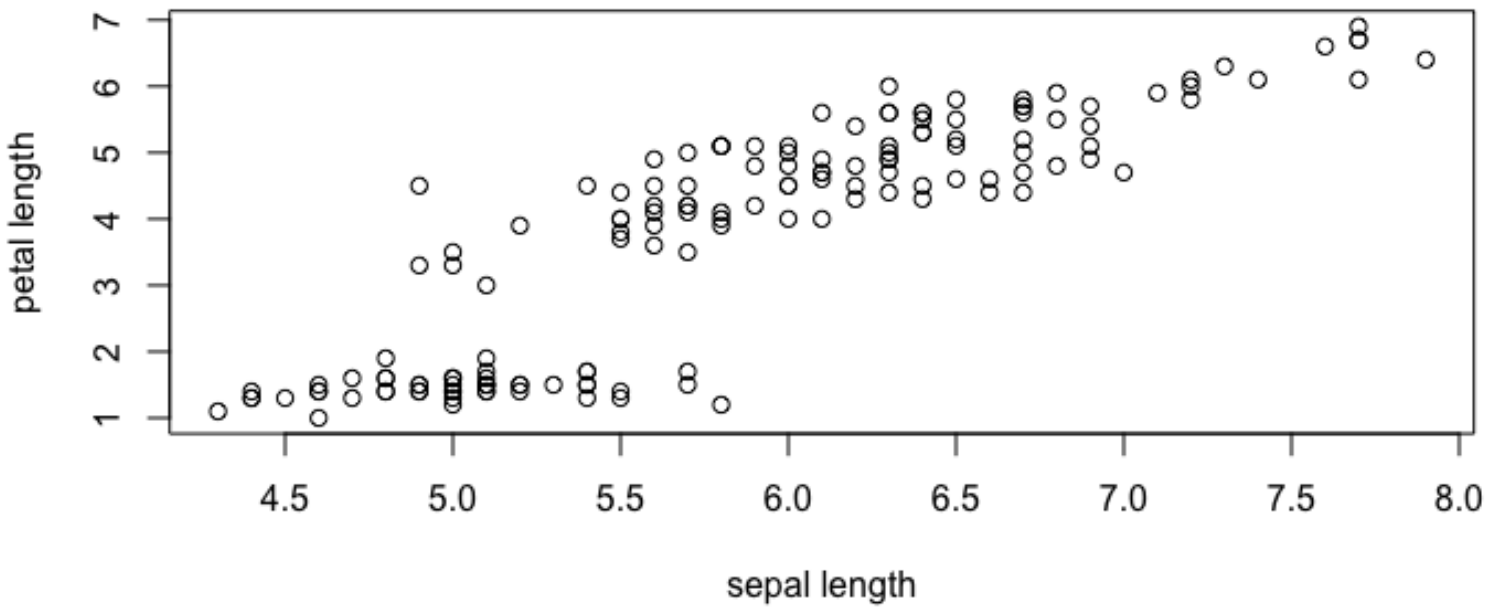
Example. The datset swiss contains measure of fertility and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888. The variable *Education* refers to the % of education beyond primary school.



# Scatterplot

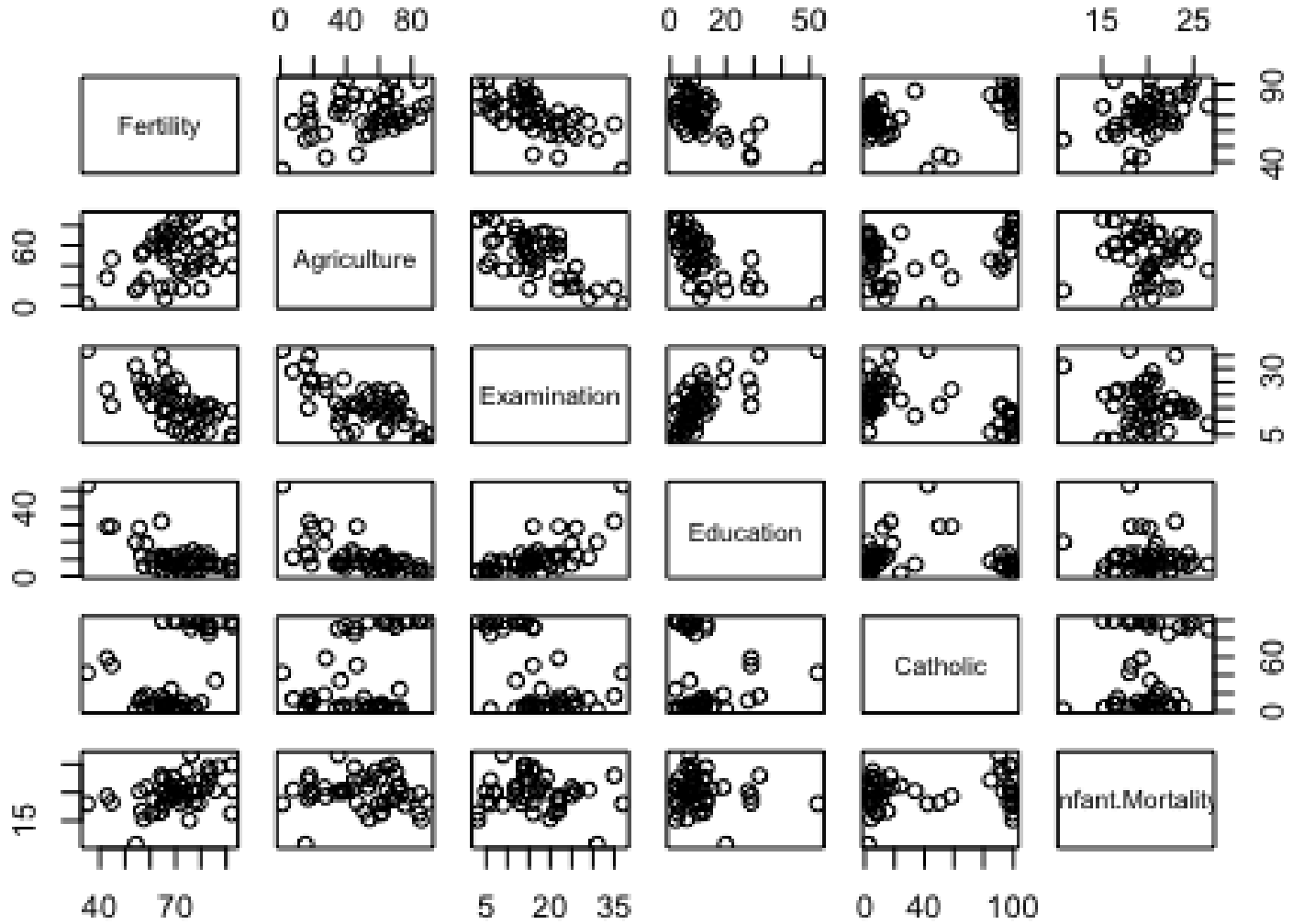
Scatterplots are bivariate (or trivariate) plots of variables against each other. They help us understand relationships among the variables of a data set.

Example. Example of scatterplot using the iris dataset.



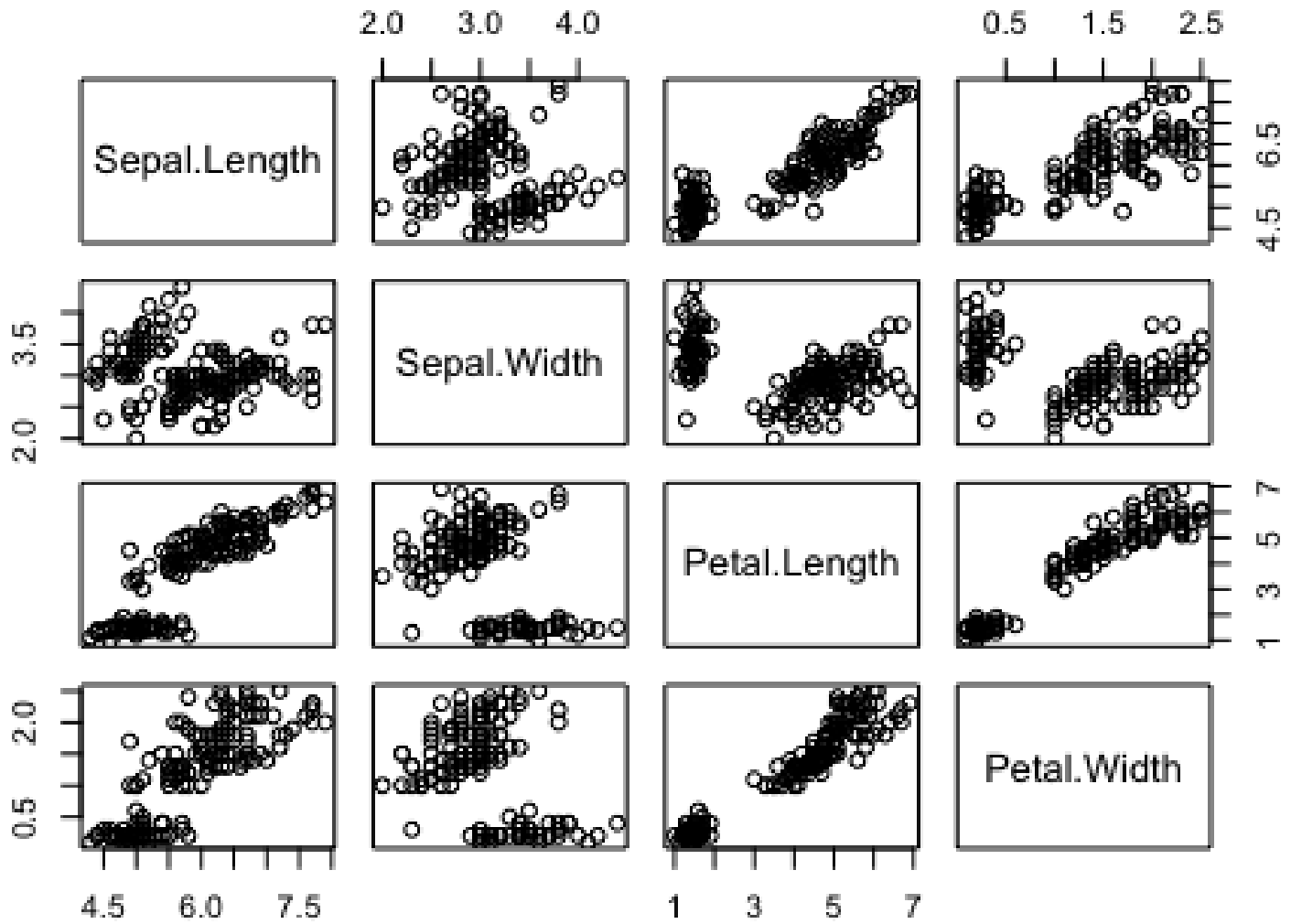
# Scatterplot matrix I

Example. Scatterplot matrix for swiss dataset.



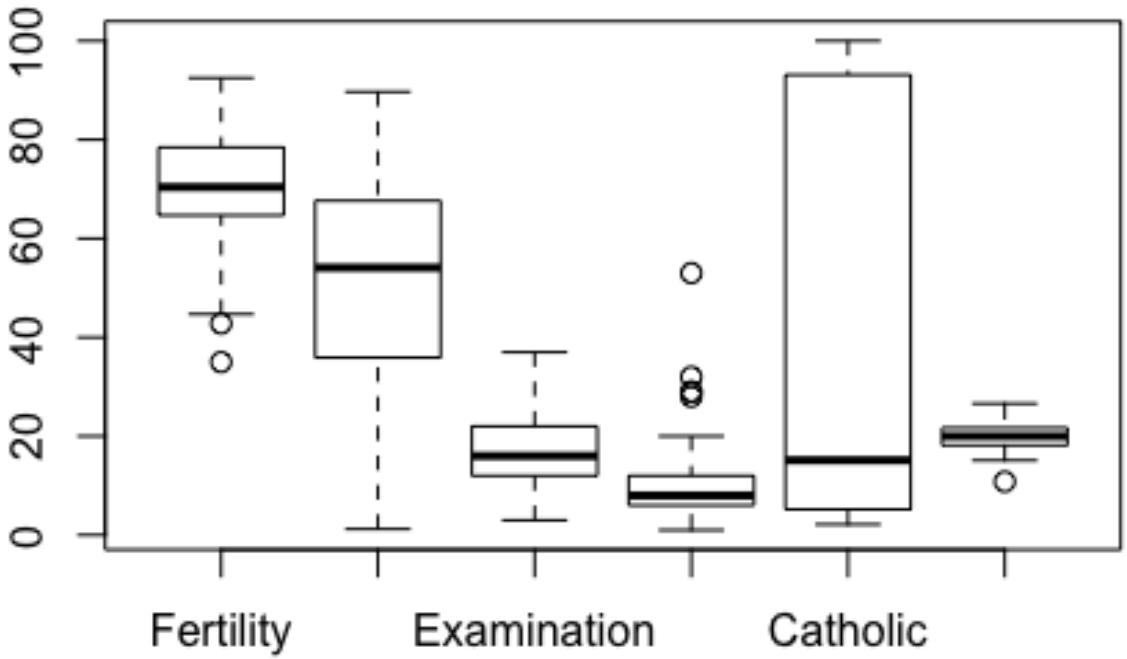
# Scatterplot matrix II

Example. Scatterplot matrix for iris dataset.



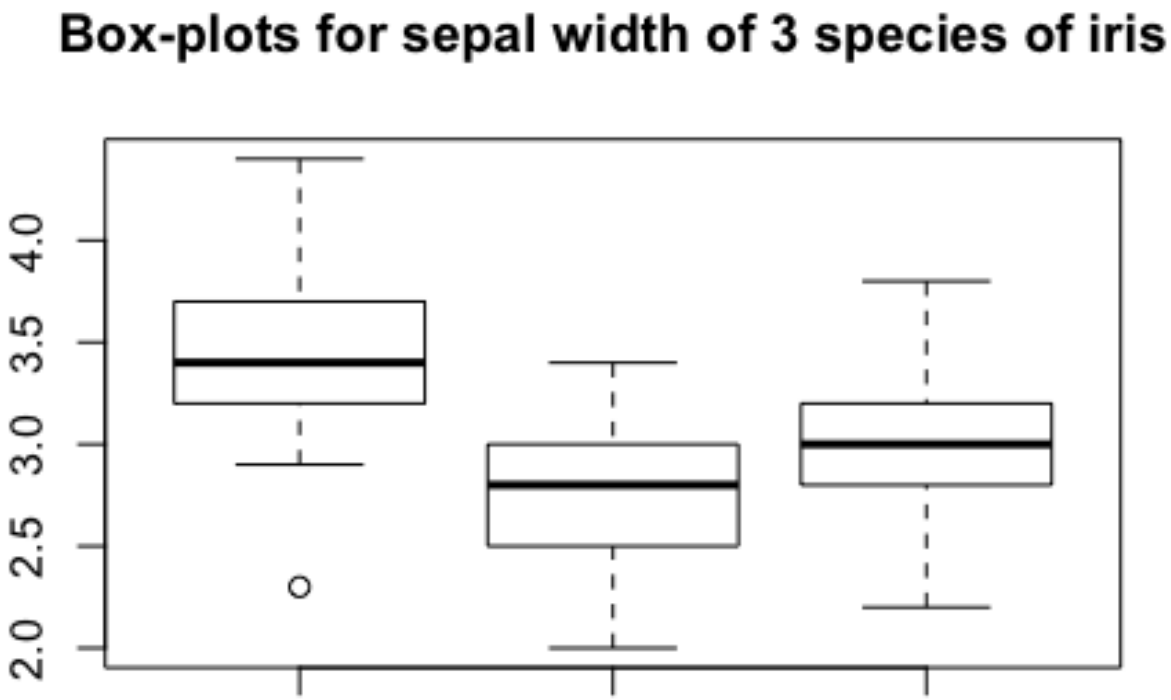
# Boxplots I

Example. Boxplots for swiss dataset.



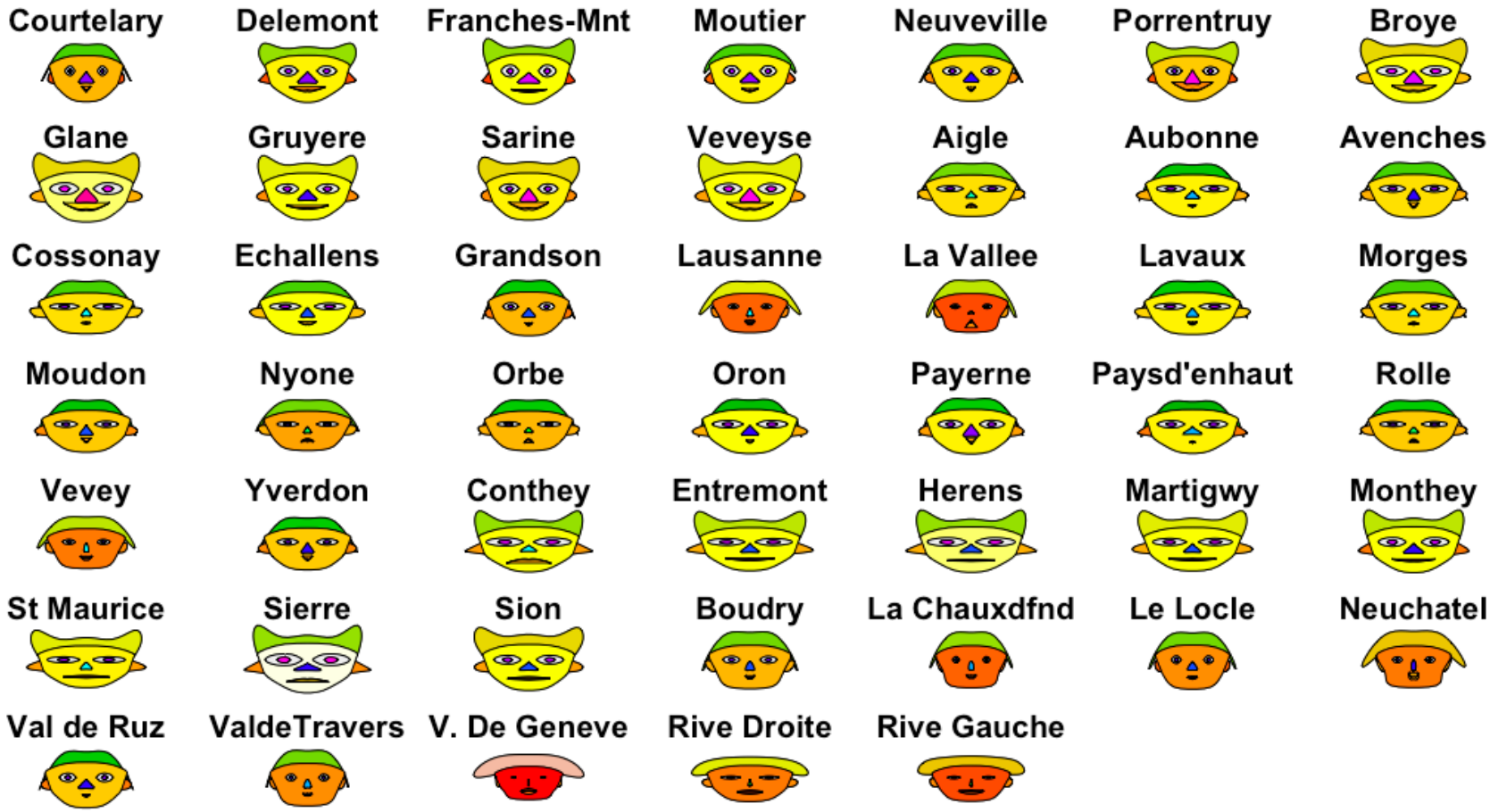
# Boxplots II

Example. Boxplots for iris dataset.



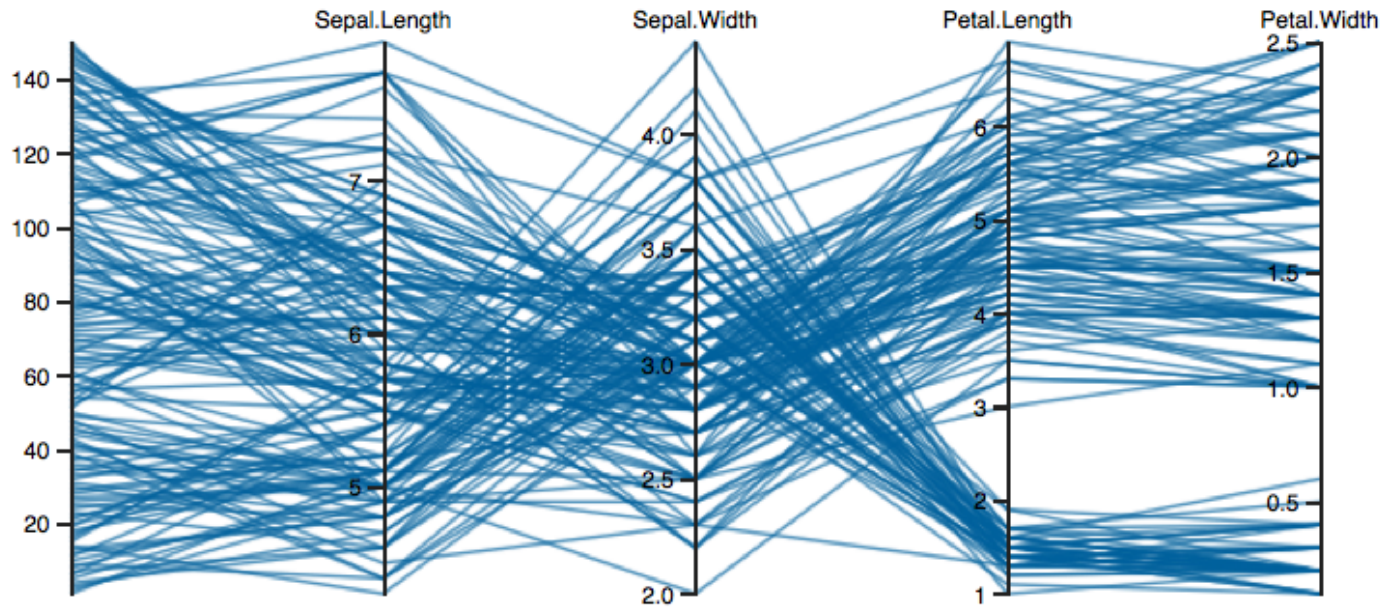
# Chernoff faces

Example. Chernoff faces for swiss dataset.



# Parallel coordinates plot

Example. Parallel coordinates plot for iris dataset.



# Andrews curves

Example. Andrews curves for swiss dataset.