Statistics

L1: Descriptive Statistics

Katarzyna Filipiak

Institute of Mathematics
Poznań University of Technology

2025/2026

Organization

Lectures – 30h: dr hab inż. Katarzyna Filipiak, prof. PP

Labs – 30h: dr hab inż. Katarzyna Filipiak, prof. PP
dr inż. Monika Mokrzycka

R: https://cloud.r-project.org/ and RStudio Desktop

Assessment methods:

Lectures - written test concerning mainly the theoretical part of the subject

Labs - evaluation of two tests with the use of R software

Program

Statistics – the art of learning from data; it is concerned with the collection of data, their subsequent description, and their analysis, which often leads to the drawing of conclusions.

- Descriptive statistics
- Random variables (revisiting)
- Mathematical statistics
  - Estimation
  - Hypothesis testing
  - Correlation and regression analysis
  - Goodness-of-fit tests
  - Contingency tables and independence
  - Nonparametric tests

Types of the data

Qualitative data – characteristics, that are not numerical, e.g., the color of eyes, pain level, income level, the place of leaving

Quantitative data – numerical characteristics:

- discrete – when the set of possible values is finite or countable (such a characteristic is measured by 'counting'), e.g., number of failures of some equipment during a week, number of errors on one page of a given book
- continuous – when the set of possible values is uncountable, e.g., time, hight, length

Presentation of the data – discrete case

- raw or ordered data:
  - data input: `c(`*comma separated data*`)`
  - reading raw data from csv file:

  standard: `read.csv(`*name*`, sep = ";")`
  polish coding style (with comma): `read.csv(`*name*`, sep = ";", dec = ",")`

- (point) frequency table: `table(`*data*`)`

- probability / frequency line-graph (line-histogram):
  `discrete.histogram(`*data*`)` `discrete.histogram(`*data*`,freq = T)`
  Caution! "arm" package required
  `plot(table(`*data*`)/length(`*data*`))` `plot(table(`*data*`))`

- pie chart: `pie(table(`*data*`))`

Example

Perform an experiment asking your colleagues for the number of bicycles in the household (student+parents+sisters+brothers or student+partner+children).

(a) Construct frequency table.

(b) Draw frequency and probability line graphs.

data input:	`bikes = c(1,4,0,...,3)`
frequency table:	`table(bikes)`
package installation:	`install.packages("arm")`
package reading:	`library(arm)`
line-histogram:	`discrete.histogram(bikes)`

Presentation of the data – continuous case

- raw data
- interval frequency table: `table(cut(`*data*`,`k`))`
  (k – number of intervals)
- frequency histogram:
  `hist(`*data*`, main=`*title*`, xlab=`*label of OX*`)`
- probability histogram:
  `hist(`*data*`, main=`*title*`, xlab=`*label of OX*`, freq = F)`
- pie chart: `pie(table(cut(`*data*`,`k`)))`

Construction of frequency table

General rules:

- each observation has to be set to one class
- non-empty classes
- mutually exclusive classes

Number of classes, k (n - number of observations):

$$k \approx \sqrt{n}, \quad \boxed{\frac{\sqrt{n}}{2} \leq k \leq \sqrt{n}}, \quad k \leq 5 \log n, \quad k \approx 1 + 3,322 \log n$$

Length of the interval, h:

$$h \approx \frac{x_{max} - x_{min}}{k} \quad \text{(rounded up)}$$

Example

The following sample data represents the ozone concentration (measured in parts per 100 million) of air in the downtown of the city during 78 consecutive summer days. Construct a frequency histogram for this data.

3.5	1.7	3.1	4.5	3.0	3.7	4.1	9.4	2.5	6.1
6.8	1.1	5.8	4.2	6.0	7.6	3.5	5.3	3.0	8.1
2.4	7.5	4.7	5.4	1.4	6.6	5.9	4.7	5.1	2.0
6.8	5.8	5.7	6.5	2.8	4.1	6.0	6.7	6.2	6.2
5.5	3.4	6.0	7.4	2.5	3.7	5.6	1.4	7.6	5.6
6.2	3.1	4.4	5.5	3.7	5.8	6.6	6.6	3.8	4.0
5.7	4.4	4.7	5.8	3.3	5.3	1.4	3.9	4.4	3.4
9.4	6.6	4.7	1.6	6.8	5.4	5.6	11.7		

maksimum: `max(data)` minimum: `min(data)`

rounding up: `ceiling(data)` △

Numerical description of the data

$x_1, x_2, \ldots, x_n$ – data

$x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ – data arranged in increasing order

Numerical quantities:

- central tendency measures (e.g., mean, median, mode, quartiles)
- variability measures (e.g. variance, standard deviation, range)
- asymmetry measures
- concentration measures
- ...

Central tendency measures – mean

Mean (average)

- for raw data: $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ or `mean(`*data*`)`

- for grouped data: $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$ (n_i – frequency of the ith class)

- for interval data: $\overline{x} = \frac{1}{n}\sum_{i=1}^{k} m_i n_i$ (m_i – middle of the ith interval)

Central tendency measures – mode

Mode (dominant value)

For raw data or grouped data

observation that occurs the most frequently in the data set

Central tendency measures – quantiles

$100p$ quantile (percentile) – the data value $x_{[p]}$ having the property that at least $100p\%$ of the data are less than or equal to it and at least $100(1-p)\%$ of the data values are greater than or equal to it.

$$x_{[p]} = \begin{cases} x_{(\lfloor pn \rfloor + 1)} & \text{if } \ pn \notin \mathbb{N} \\ \frac{1}{2}\left(x_{(pn)} + x_{(pn+1)}\right) & \text{if } \ pn \in \mathbb{N} \end{cases}$$

For grouped data:

- indicate the class in which cumulative frequency is greater or equal to pn
- $x_{[p]}$ is equal to the value of observation in this class

`quantile(`*data*`, probs=`*vector of probabilities*`)`

Central tendency measures – quartiles

$Q_1 = x_{[0.25]}$ – first quartile (25th quantile)

$Q_2 = x_{[0.5]} = x_{me}$ – second quartile or median (50th quantile)

$Q_3 = x_{[0.75]}$ – third quartile (75th quantile)

`quantile(`*data*`)`

All central tendency measures together:

`summary(`*data*`)`

Variability measures

Variance:

- for raw data: $s^2 = \frac{1}{\ell}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \frac{1}{\ell}\left(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\right)$

- for grouped data:

$$s^2 = \frac{1}{\ell}\sum_{i=1}^{k}(m_i - \overline{x})^2 n_i = \frac{1}{\ell}\left(\sum_{i=1}^{k} m_i^2 n_i - n\overline{x}^2\right)$$

$\ell = n$ – for population variance

$\ell = n - 1$ – for sample variance: `var(data)`

Standard deviation: $s = \sqrt{s^2}$ `sd(data)`

Variability measures

Range: $R = x_{(n)} - x_{(1)}$

Interquartile range: $R_Q = Q_3 - Q_1$

Variability index: $v = \frac{s}{\overline{x}} \cdot 100\%$

0 – 20%	– **weak** variability of the data
20 – 40%	– **medium** variability of the data
40 – 60%	– **strong** variability of the data
over 60%	– **very strong** variability of the data

Example

The following table represents the notes of two groups of students at the end of some course. Compare these groups by central tendency and variability measures.

group A	3.0	3.0	4.0	4.5	4.5
group B	2.0	3.5	4.0	4.5	5.0

Graphical interpretation of data

Box-plot (Stem-and-Leaf plot) – graphical summarizing of central tendency measures and variability measures

On a horizontal axis we draw:

$$x_{min},\ Q_1,\ x_{me},\ Q_3,\ x_{max}.$$

Then, we impose on the line a 'box', which starts at the first and continues to the third quartile, with the value of the second quartile indicated by a vertical line. The length of the 'box' is equal to the interquartile range R_Q. The length between x_{min} and x_{max} is equal to the range.

`boxplot(data)`

Example - cont.

The following table represents the notes of two groups of students at the end of some course. Compare these groups using box-plots.

group A	3.0	3.0	4.0	4.5	4.5
group B	2.0	3.5	4.0	4.5	5.0

`boxplot(`*groupA*`,` *groupB*`)`

Chebyshev's Theorem

For an arbitrary data set

- the interval $(\overline{x} - 2s; \overline{x} + 2s)$ contains at least 75% $\left(1 - \frac{1}{2^2} = \frac{3}{4}\right)$ of the data
- the interval $(\overline{x} - 3s; \overline{x} + 3s)$ contains at least 89% $\left(1 - \frac{1}{3^2} = \frac{8}{9}\right)$ of the data
- the interval $(\overline{x} - ks; \overline{x} + ks)$ contains at least $1 - \frac{1}{k^2}$ of the data $(k > 1)$