Statistics for Geosciences – Winter term 2021/22 **Exercise 1**

Robert Czudaj, Julian Rodemann 27.10.2021

Exercise 1 - Data collection: fundamental terms

Specify population, sample (size), statistical unit, statistical variable (feature), the variable's type (discrete, dichotomous, ...) and, if provided, the variable's realizations in the following situations.

a) Locations of lightning strikes in Germany for 2020 are collected in order to visualize them.

b) A Student asks his two seatmates whether they like their new teacher or not in order to get an impression of the new teacher's popularity in the class. One likes the new teacher, the other does not.

c) Hourly wind direction data $w_i \in [0, 360), i \in \{1, ..., 24\}$, is collected at Zugspitze at a given day in order to get an impression of the air movement on that day.

Solution:

a) **Population:** All lightning strikes in Germany in the year of 2020, **sample:** corresponds to population (complete survey), **statistical unit:** lightning strike, **statistical variable:** location of lightning strike. **type:** continuous

a) **Population:** All students in the class, **sample:** Two seatmates, **statistical unit:** Student in the class, **statistical variable:** Student's opinion on the new teacher, **type:** dichotomous (binary) **variable's realizations:** $\{\text{like}, \text{do not like}\}$

a) **Population:** Wind movement on a given day at Zugspitze, **sample:** 24 hourly measurements, **statistical unit:** hourly wind measurement, **statistical variable:** wind direction $w_i \in [0, 360)$, type: angular (cyclic)

Exercise 2 - Levels of measurement

Describe the following data by specifying which level of measurement they have (nominal, ordinal, dichotomous etc.):

a) Party preference at an election

b) Level of difficulty in a video game

c) Age of zoo animals

d) Calendar time with the birth of Christ at point zero

e) Enrollment number of a student

f) Grades in school

Solution:

a) Nominal scale
No natural ordering

b) Ordinal scale
Levels are ordered, but differences normally not equal between difficulty levels 1, 2 and 3

c) Proportional scale
Natural zero point (birth date)

d) Interval scale
No natural zero point (birth of Christ)

e) Nominal scale
Even if numeric, number 112233 isn't "half as good" as number 224466

f) Ordinal scale
Natural ordering, but no equal differences

Exercise 3 - Measures of central tendency and variation

Measuring the age of $n = 10$ randomly selected people on the street one obtains the following sample:

Observational unit i	1	2	3	4	5	6	7	8	9	10
x_i	39	44	29	51	58	54	54	49	48	43

Calculate the following measures based on the sample and judge whether the measures are meaningful for the given situations:

a) Arithmetic mean $\bar{x}$

b) Median $\tilde{x}_{0.5}$
Additional question: Why do the median and the arithmetic mean differ for this sample?

c) Mode x_{mod}

d) Sample standard deviation s

Solution:

a) Arithmetic mean:
$\bar{x} = \frac{1}{10}(39 + 44 + \ldots + 48 + 44) = 46.9$

Judgement: Measure is meaningful for metric variables, as long as no extreme outliers are present.

b) Median:
Reorder the data: $29, 39, 43, 44, 48, 49, 51, 54, 54, 58$
n is even $\rightarrow \frac{n}{2} = 5$
$\Rightarrow \tilde{x}_{0.5} = \frac{1}{2}[x_{(5)} + x_{(6)}] = 48.5$

Judgement: Measure is meaningful for metric variables.

Additional question: They differ as the distribution is negatively skewed and the arithmetic mean is not robust against extreme values.

c) Mode:
Only age 54 has a frequency of 2, all other ages occur only once!
$\Rightarrow x_{mod} = 54$

Judgement: Measure not very meaningful for metric variables!

d) Sample standard deviation:
Recap that the arithmetic mean is $\bar{x} = 46.9$
Standard definition: $s = \sqrt{\frac{1}{n-1} \sum\limits_{i=1}^{n} (x_i - \bar{x})^2}$
Often easier to use the "Verschiebungssatz":

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$\Rightarrow \sum\limits_{i=1}^{n} x_i^2 = (39^2 + 44^2 + \ldots + 43^2) = 22649$
$\Rightarrow \sum\limits_{i=1}^{n} x_i^2 - n\bar{x}^2 = 22649 - 10 \cdot 46.9^2 = 652.9$
$\Rightarrow s = \sqrt{\frac{1}{9} \cdot 652.9} \approx \sqrt{72.54} \approx 8.52$

Judgement: Measure makes sense for the metric variable age.

Exercise 4 - Descriptive Statistics

The following table gives the amount of rain (in litres per square metre), measured at the volcano Merapi (Indonesia) between January 1st and January 20th, 1995.

Rain	Date	Rain	Date
2	01.01.1995	50	11.01.1995
9	02.01.1995	12	12.01.1995
18	03.01.1995	0	13.01.1995
2	04.01.1995	0	14.01.1995
23	05.01.1995	0	15.01.1995
42	06.01.1995	0	16.01.1995
11	07.01.1995	3	17.01.1995
13	08.01.1995	3	18.01.1995
40	09.01.1995	40	19.01.1995
12	10.01.1995	48	20.01.1995

a) What is the level of measurement of the variable `rain`?

b) Draw a histogram using the intervals [0,10),[10,20),[20,30),[30,40),[40,50].

c) Calculate the the following measures of location and dispersion based on the sample: mode, median, arithmetic mean, lower quartile, upper quartile, sample variance, sample standard deviation, and coefficient of variation.

d) Use the results obtained in (c) to draw a boxplot of the empirical distribution of `rain`, and interpret it.

Solution:

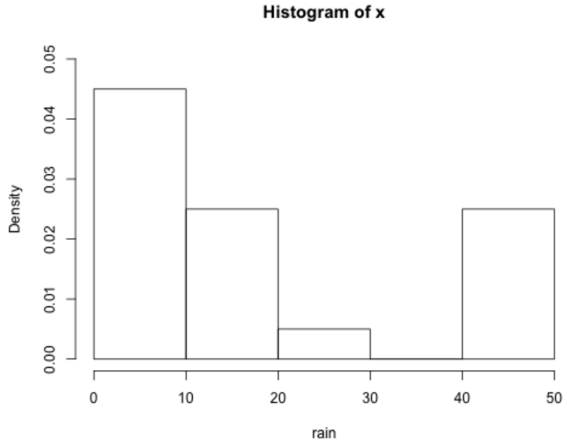
a) Proportional scale, as rain has a naturally defined zero point.

b) Drawing a histogram:

(a) Calculate relative frequencies f_j per category j

(b) Calculate the height of each bar: $\text{height}_j = \frac{f_j}{\text{width}_j}$, s.t. the area of each bar is proportional to its relative frequency

Category	h_j	f_j	height_j
$[0, 10)$	9	$\frac{9}{20} = 0.45$	$\frac{0.45}{10} = 0.045$
$[10, 20)$	5	$\frac{5}{20} = 0.25$	$\frac{0.25}{10} = 0.025$
$[20, 30)$	1	0.05	0.005
$[30, 40)$	0	0	0
$[40, 50]$	5	0.25	0.025



c) **Mode:**

$x_{mod} = 0$, as the most frequent value

Median and quartiles:

First we order our dataset: 0 0 0 0 2 2 3 3 9 11 12 12 13 18 23 40 40 42 48 50

$n = 20$ is even $\rightarrow x_{med} = \frac{1}{2}(x_{(n/2)} + x_{(n/2+1)}) = \frac{1}{2}(x_{(10)} + x_{(11)}) = \frac{1}{2}(11 + 12) = 11.5$

$n \cdot 0.25 = 5$ is an integer $\rightarrow x_{0.25} = \frac{1}{2}(x_{(np)} + x_{(np+1)}) = \frac{1}{2}(x_{(5)} + x_{(6)}) = \frac{1}{2}(2 + 2) = 2$

$n \cdot 0.75 = 15$ is an integer $\rightarrow x_{0.75} = \frac{1}{2}(x_{(15)} + x_{(16)}) = \frac{1}{2}(23 + 40) = 31.5$

Arithmetic mean:

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20} \cdot 328 = 16.4$

Sample variance:

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{19} \cdot 5926.8 \approx 311.94$

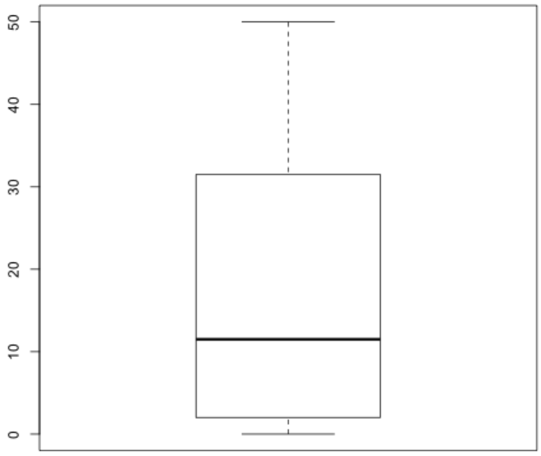
Sample standard deviation:

$s = \sqrt{s^2} \approx 17.66$

Coefficient of variation:

$v = \frac{s}{\bar{x}} = \frac{17.66}{16.4} \approx 1.08$

d)



Interpretation:

- The distribution is positively skewed, i.e. mostly small values
- The middle 50% of the data lie in the interval $[2, 31.5]$
- 25% of values are ≤ 2, 25% are ≥ 31.5

Exercise 5 - Association of categorical variables

In the following, we will analyze data about all Titanic passengers that were on board during its sinking. Use the information given in the contingency tables to calculate and interpret Odds Ratios to quantify the difference in the chance to survive (Yes: Person survived, No: Person did not survive) between the following groups:

a) Passengers from first and third class

	1st	2nd	3rd	Crew
No	122	167	528	673
Yes	203	118	178	212

b) Male and female passengers

	Male	Female
No	1364	126
Yes	367	344

c) Male crew members and male passengers from third class

Contingency table only based on males:

	1st	2nd	3rd	Crew
No	118	154	422	670
Yes	62	25	88	192

Solution:

a)

$$\text{OR} = \frac{\#(\text{Yes}, \text{1st})/\#(\text{No}, \text{1st})}{\#(\text{Yes}, \text{3rd})/\#(\text{No}, \text{3rd})} = \frac{203/122}{178/528} \approx 4.94$$

⇒ The chance to survive was 4.9 times higher for first class passengers than for third class passengers

b)

$$\text{OR} = \frac{\#(\text{Yes}, \text{Male})/\#(\text{No}, \text{Male})}{\#(\text{Yes}, \text{Female})/\#(\text{No}, \text{Female})} = \frac{367/1364}{344/126} \approx 0.10$$

⇒ The chance for male passengers to survive was only 10% of the chance for female passengers to survive

c)

$$\text{OR} = \frac{\#(\text{Yes}, \text{Crew})/\#(\text{No}, \text{Crew})}{\#(\text{Yes}, \text{3rd})/\#(\text{No}, \text{3rd})} = \frac{192/670}{88/422} \approx 1.37$$

⇒ The chance for male crew members to survive was 37% higher than the chance for male third-class passengers to survive