Course of Study Bachelor Computer Science	Exercises Statistics WS 2020/21

Sheet IV

Descriptive Statistics - Frequency Tables and Distributions

1. Consider the results of the national elections in Germany in 2013 and 2017:

Party	Results 2013 (%)	Results 2017 (%)
CDU	26,8%	34,1%
SPD	20,5%	25,7%
AfD	12,6%	4,7%
FDP	10,7%	4,8%
DIE LINKE	9,2%	8,6%
GRUENE	8,9%	8,4%
CSU	6,2%	7,4%
Others	5,0%	6,2%

   Summarize the results of 2017 in a pie and abar chart. Compare the results in 2013 and 2017 with an appropriate bar chart.

2. The data shown in the list are the times in milliseconds it took one of us to move the mouse over a small target in a series of 20 trials. The times are sorted from shortest to longest.

   568, 577, 581, 640, 641, 645, 657, 673, 696, 703, 720, 728, 729, 777, 808, 824, 825, 865, 875, 1007

   (a) Compute and draw the cumulative frequency distribution.

   (b) Compute using the cumulative frequency distribution the proportion of response times

      i. less equal 800
      ii. greater than 725
      iii. greater than 642 and less equal 777
      iv. equal 696

in the sample.

(c) Consider the following classes $(500, 600]$, $(600, 700]$, $(700, 800]$, $(800, 900]$, $(900, 1000]$, $(1000, 1100]$.

- Compute the grouped frequency distribution and draw the histogram.
- Assume that the values within each interval are distributed uniformly. Determine the proportion of response times from above and draw the corresponding cumulative distribution function.

(d) The classes are now $(500, 600]$, $(600, 900]$,$(1000, 1200]$. Mention that the classes have different width. Compute the grouped frequency distribution and draw the histogram. Can you interpret the y-values in the diagram?

Descriptive Statistics - Measures

1. Make up data sets with 5 numbers each that have:

   (a) the same mean but different standard deviations.

   (b) the same mean but different medians.

   (c) the same median but different means.

2. Consider a stock portfolio that began with a value of 1000 $ and had annual returns of 13%, 22%, 12%, -5%, and -13%.

   (a) Compute the value after each of the five years.

   (b) Compute the annual rate of return.

   Use the **geometric mean**: $\sqrt[n]{\prod_{i=1}^{n} x_i}$

   (c) Based on the result of b), which annual returns do you expect in the next two years? Would it make sense to prdeict the annual return 20 years later?

3. A sample of 30 distance scores measured in yards has a mean of 7, a variance of 16, and a standard deviation of 4.

   (a) You want to convert all your distances from yards to feet, so you multiply each score in the sample by 3. What are the new mean, variance, and standard deviation?

(b) You then decide that you only want to look at the distance past a certain point. Thus, after multiplying the original scores by 3, you decide to subtract 4 feet from each of the scores. Now what are the new mean, variance, and standard deviation?

4. Which of the following measures of location can be used for a qualitative variable, a quantitative continuous variable resp. an ordinal variable?

   - Mode
   - Median
   - Mean

5. You have the following 25 observations of the variable Number. Calculate the arithmetic mean, the geometric mean, the harmonic mean and the trimmed 20% mean.

Number	Absolute frequency
1	5
2	4
3	1
4	7
5	2
6	3
7	1
8	2
Sum	25

6. Which of the following measures of dispersion can be used for a qualitative variable resp. a quantitative continuous variable?

   - Variance
   - Standard deviation
   - Interquartile range
   - Range

7. Define for each of the measures mean, quantile, variance, geometric mean, harmonic mean and trimmed mean based on their definitions given in the lecture a R function.

   (a) Use the sample x_i : $3, 7, 2, 5, 6, 10, 6, 3, 6, 5$ to test the functions. Calculate the 3 quartile and the 10% trimmed mean for the given sample.

(b) In R there are several methods offered to compute quantiles. These methods are speciefied by the argument type $\in 1, 2, ..., 9$. Identify the differences between a computation of type 1 and type 7 which is default computation. Use the help function to get more informations about the computation of the qunatiles

8. An experiment compared the ability of three groups of participants to remember briefly-presented chess positions. The data are shown below. The numbers represent the number of pieces correctly remembered from three chess positions.

   Compare the performance for each group by computing mean, median, min, max, quartiles, interquartile range, variance. Create side-by-side box plots for these three groups. What can you say about the differences between these groups from the box plots?

Non-players	Beginners	Tournament Players
22.1	32.5	40.1
22.3	37.1	45.6
26.2	39.1	51.2
29.6	40.5	56.4
31.7	45.5	58.1
33.5	51.3	71.1
38.9	52.6	74.9
39.7	55.7	75.9
43.2	55.9	80.3
43.2	57.7	85.3

9. Exercise 3.1 from Heumann, Schomaker: Introduction to Statistics and Data Analysis, page 63
   A hiking entusiast has a app for his smartphone which summarizes his hikes by using a GPS device. The distance hiked (in km) and maximum altitude (in m) for the last 10 hikes:

Distance	12.5	29.9	14.8	18.7	7.6	16.2	16.5	27.4	12.1	17.5
Altitude	342	1245	502	555	398	670	796	912	238	466

   (a) Calculate the arithmetic mean and median for both distance and altitude.

   (b) Determine the first and third quartile for both distance and altitude. Discuss the shape of the distribution given the values in a) and b).

(c) Calculate the interquartile range and standard deviation for both variables. Compare the variability of both variables.

(d) Draw the box plot for both distance and altitude.

(e) Assume distance is measured as only short (5-15 km), moderate (15-20 km) and long (20-30 km). Summarize the grouped data in a frequency table. Calculate the weighted arithmetic mean under the assumption that the raw data is not known.

10. The data set mpg of the ggplot package contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008 - this was used as a proxy for the popularity of the car.

(a) Inspect the description of the data set using the ?mpg() command.

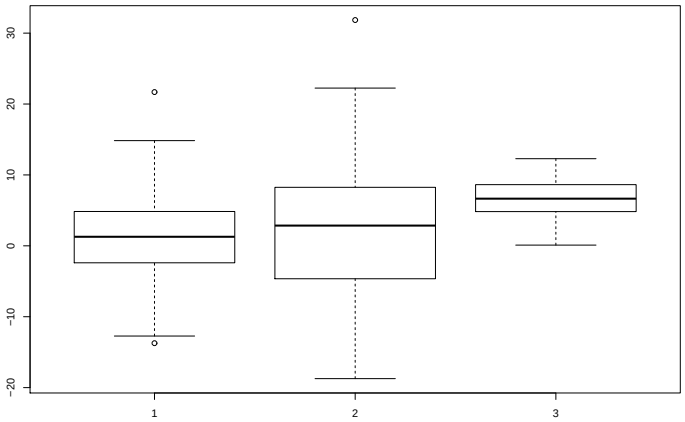
(b) Select only the variables displ (engine displacement) and hwy (highway miles per gallon) from the data set. Group the values of the variable displ into the the groups "low" ($1 \leq \text{displ} < 3$), "medium" ($3 \leq \text{displ} < 5$) and "big" ($5 \leq \text{displ} < 8$). Use the cut() command to do this. Add a column displ_class which denotes the belonging to one of the groups.

(c) Calculate the mean, minimum, maximum and the three quartile of the variable hwy depending on the values of displ and depending on displ_class.

(d) Draw boxplots of the variable hwy grouped by displ resp. displ_class and interpret the results.

Descriptive Statistics - Shape

1. Use the following boxplots to answer the questions below:



(a) Which of the three distributions has the highest measure of location?

(b) Which of the three distributions has the largest range?

(c) Which of the three distributions has the largest interquartile range?

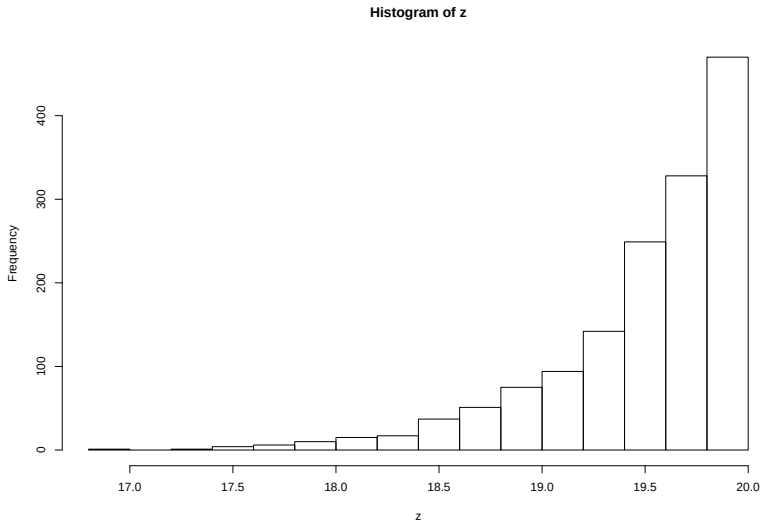
(d) Which of the three distributions has the highest maximum value?

(e) Which of the three distributions has the smallest maximum value?

(f) Discuss skewness/symmetry of the three distributions.

Motivate your answers!

2. Use the following histogram to answer the questions below:



Is the distribution left-skewed, right-skewed or symmetric? Motivate your answers!