Data Mining Summer Term 2020

Institute of Neural Information Processing

PD Dr F. Schwenker

Assignment 1 (Submission until May 12, 2020)

Exercise 1 (10 points): Statistical Measures [Pen and Paper]

Whenever we deal with a dataset to solve a particular problem, it is helpful to first analyse the data and extract some information. In this exercise, we are going to take a look at statistical measures which are very basic tools and describe the data by a single number only. However, this does not mean that they are useless or less important than other (more complex) analysis. Every measure conveys a special meaning and must be applied with care since they are also easily misinterpreted.

Here, we are using a small one-dimensional dataset

$$X = \{5, 4, 10, 1, 5, 25\} \tag{1}$$

where we can easily calculate the results by hand. Besides the measures, we are going to work with a box-and-whisker plot which is a very important visualization technique summarizing a lot of information.

Table 1: Overview of statistical measures. Please refer to the script for the definition of the measures.

Measure	Required scaling	Value of measure for X
Mode		
Arithmetic mean $\bar{x}$		
Quantile $\tilde{x}_{0.25}$		
Median $\tilde{x}_{0.5}$		
Range R		
Interquartile range Q		
Variance s^2		
Skewness g		
Quartile skewness g_Q		

1. Let's start with an overview of the measures. For this, calculate the missing values in Table 1.

a) Decide for each statistical measure about the minimally required scaling (level of measurement) which the data must have at least so that it is meaningful to calculate the measure (second column of Table 1). Hint: the list of allowed operations per scaling level on the Wikipedia page[1] may proves to be useful.

b) Apply each statistical measure to the dataset X (third column of Table 1).

2. Figure 1 shows a box-and-whisker plot of X. We now want to analyse which information we can retrieve from this plot. The figure is annotated with labels corresponding to a value of a statistical measure.

   a) Name what the labels ①, ②, ③ and ⑥ shows us.

   b) The other labels may deserve a bit of explanation. Unfortunately, it is not strictly defined how the whiskers (④ and ⑤) are set and what is treated as an outlier (⑦). The general idea is to show the relevant range of the data via the whiskers and all other data points, where the evidence is strong that they don't really belong to the dataset, are considered as outliers. Toolkits and libraries usually provide a way to configure this behaviour. However, there is a common convention for the whiskers which says that they are set to at most 1.5 times of the interquartile range. "most" because we adjust the whiskers to represent a value which is actually present in the dataset. Every data value which still lies outside this range is considered an outlier.

      i. Calculate the lower whisker: $\min\left(\{x \mid x \in X \land x \geq \tilde{x}_{0.25} - 1.5 \cdot Q\}\right) = \ ?$

      ii. Calculate the upper whisker: $\max\left(\{x \mid x \in X \land x \leq \tilde{x}_{0.75} + 1.5 \cdot Q\}\right) = \ ?$

   c) Can you infer the sign of the skewness g from the box-and-whisker plot?

3. Compared to the skewness g, the quartile skewness g_Q has the advantage of being more robust against outliers and the result is normed. Hint: use the figure on slide 54 as help for the following questions.

   a) What is the interval $[a; b]$ to which the values g_Q are bounded to?

   b) For which values do we say that the distribution is left, right or not skewed at all?

   c) Which conditions must hold in the extreme cases, i.e. when do we get the values a or b for g_Q.

   d) Give an example of a new dataset X_1 which results in minimal quartile skewness, i.e. $g_Q(X_1) = a$.

   e) Give an example of a new dataset X_2 which results in maximal quartile skewness, i.e. $g_Q(X_2) = b$.

[1] https://en.wikipedia.org/wiki/Level_of_measurement#Comparison

4. Look at the cartoons on this blog[2] and explain the basic flaw of each mentioned statistical measure, i.e. what are the main problems the author wants to draw your attention to? Or, to put it differently, what do you need to keep in mind when using these measures?

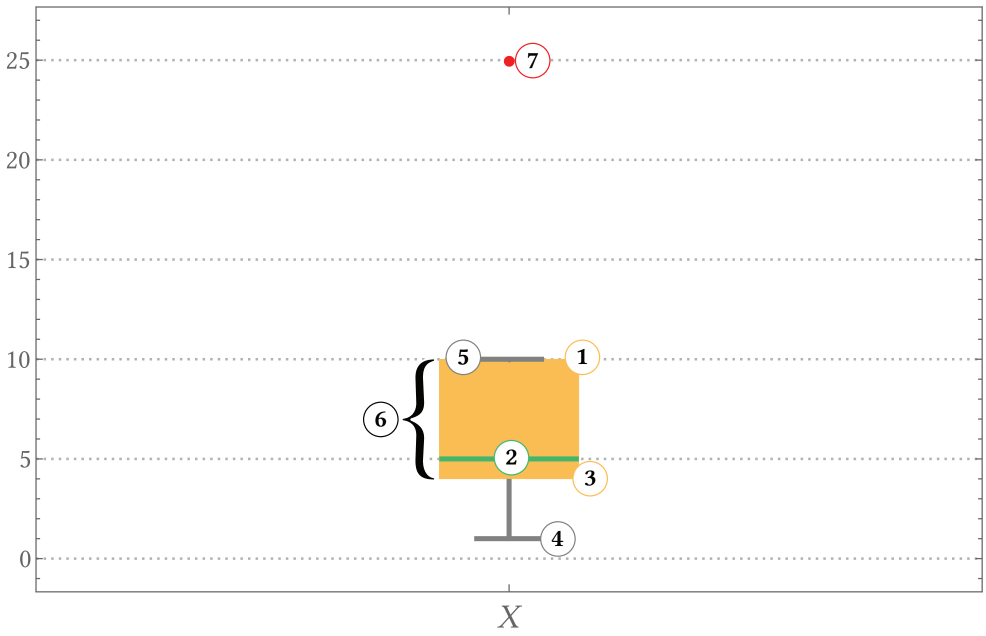

Figure 1: Annotated box-and-whisker plot of the dataset X (Equation 1).

[2] https://mathwithbaddrawings.com/2016/07/13/why-not-to-trust-statistics/