Some Notes about Outliers

Dipl.- Ing. (FH) Klaus Hoffmann, Derikumer Weg 26, Neuss, 13.01.2024

Introduction

The intended audience of this paper are engineers, practitioners, who are not primarily statisticians and naturally interested readers. For this reason, this paper is based on a rather heuristic approach to make the explanations easy to understand.

The science of statistics deals with the collection, analysis, interpretation and presentation of data. Effective interpretation of data based on good procedures and thorough examination. The goal of statistics is to gain an understanding of data. The interpretation must come from the analyst and not only from the applied statistic software. If the analyst can thoroughly grasp the basics of statistics, the analysing person can be more confident in the decisions. This essay deals with the occurrence of outliers, a statistical problem in which statistical knowledge alone is insufficient to understand. [1]

> *One well- known historical example for an outlier is related to the ozone hole over the Antarctica. Although conspicuous values were detected for years, the measured values were evaluated as obviously incorrectly measured viz. the values were interpreted as outliers and ignored.* [2]
>
> *One noted example for the possible effects of one single outlier is located in the field of economy, which is represented through the Per Capita Income (PCI) in the city of Heilbronn. This statistic is distorted by one single billionaire who lives there. The existence of this one individual causes an increase of the PCI to be the highest in Germany.* [3, 4, 5]

This paper is focused on numerical, continuous- valued data interval or ratio scales and univariate data sets, this means the evaluation of only one characteristic. The article provides a conceptual overview of outliers with special focus on common techniques used to detect them. [29, 35] Whereat the detecting for outliers is also defined as a part of data cleaning. [33] The treatise does not discuss the outlier management techniques of deletion, substitution and transformation. [29] The focus is restricted to elementary univariate methods to give the reader a cardinal insight of the fundamentally different difficulties, solution statements and to do calculations without the application of specialised software. This restriction represents also a line to the area of e.g., “machine learning” and many “data mining applications” which is also based on statistical methods. [31, 35]

This paper is oriented on giving an overview of the applications of methods, mainly categorized into two types: informal methods and formal methods.

Informal Methods

Informal methods include several outlier labelling methods on the basis of the Gaussian- distribution. In addition, robust statistics for distributions that are not normal are also briefly depicted. The most frequent applied techniques are the Z_{score}, modified Z_{score}, MAD_E, Tukey’s method (Box- Plot) and other graphical methods to illustrate the outlying observation. [35] In addition the Q-Q- Plot and the histogram can support the analyst by visualising whether the distribution is normal and to detect potential outliers.

Formal Methods

Formal methods are test- based methods. [35] This paper deals with tests on the basis of the Gaussian- distribution. Therefore, the mentioned formal methods require a test based on an informal method or a formal method to examine whether the distribution is normal. The Grubb’s and the Dixon- tests to detect potential outliers, are described in a rather brief form. The generalized extreme studentized deviate test (ESD- Test) will be explained in more detail.

Outliers

In statistics, an outlier is a data point which does not fit to the rest of the data, it differs significantly from the mainstream of the other observations. [1, 6, 9, 35] It is a case - or a very few cases - that seems to be unattached to the rest of the distribution. [32] An outlier may be defined as an observation in a set of data that appears to be inconsistent with the remainder of that data set. [13] It has a low probability that it originates from the same statistical distribution as the other observations in the data set. [30] An outlier, which is also called an extreme value or an unusual value, may be due to a variability in the measurement, an indication of novel data, or it may be the result of experimental error; the latter are sometimes excluded from the data set. [1, 6, 9] The suspicious value can be an indication of exciting possibility, but can also cause serious problems in statistical analyses. [6, 9, 33] An analyst who will be confronted with outliers will be forced to decide how to handle them. Outliers can distort statistical analyses and violate their assumptions. [9, 29] To maximize generalizability, outliers must be properly handled prior to data analysis regardless of the cause. Several statistical techniques can be used to detect, classify, and manage outliers. The presence of an excessive number of outliers should raise an alarm for researchers, as it may indicate serious problems with the sampling procedures or the conceptual definition of the population of interest. [29] Removing outliers is legitimate only for specific reasons. Outliers can occur by chance in any distribution, but they can indicate novel behaviour or structures in the data set, measurement error, or that the population has a heavy- tailed distribution. [9] They may also represent legitimate extreme cases of the target population. [29] While in the case of heavy- tailed distributions, the data indicate that the distribution has a high skewness and that one should be very cautious in using tools or intuitions that assume a normal distribution. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub- populations, [6] nonnormal distributions and unequal variances. [6, 9] It has to be noted, that it cannot be statistically shown that an outlier originates from a different distribution than the rest of the data. [30] Such an extreme value can be generated due to incidental systematic error but also a flaw in the theory. [6] The reasoning being, a statistical outlier is unlikely to arise by chance. Similarly, an outlying observation in a process control environment is an important signal of a process problem; if all the outlying values were rejected, process control would be rendered ineffective. [13]

> Note: Extreme cases that are legitimate outliers can have a strong impact and therefore need to be diagnosed and addressed. [33] Outlier detection is a principal step in statistical applications. [35]

Types of Outliers

An important aspect of an outlier detection technique is the nature of the expected outlier. Outliers can be classified into the following three categories [23, 24]:

- point outliers [23] or global outliers [24]
- contextual outliers [23] or conditional outliers [24]
- collective outliers [23]

Point Outliers

A data point is considered a global outlier if its value is far outside the entirety of the data set in which it is found. [24] This is the simplest type of outlier and is the focus of the majority of research on outlier detection. [23]

Contextual Outliers

A data point is considered a contextual outlier if its value significantly deviates from the rest of the data points in the same context. This means that the same value may not be considered to be an outlier if it occurred in a different context. [24]

Collective Outliers

A subset of data points within a data set is considered abnormal if those values as a collection deviate significantly from the entire data set, but the values of the individual data points are not themselves anomalous in either a contextual or global sense. [24]

Applications of Methods for the Detecting of Outliers

Although only a basic understanding of the various methods of outlier detection should be conveyed, the subsequent exemplary mentioned applications should show the far- reaching importance.

- Mobile Phone Fraud Detection.

In this activity monitoring problem, the calling behaviour of each account is scanned to issue an alarm when an account appears to have been misused. [23]

- Insider Trading Detection

Insider trading is a phenomenon found in stock markets, where people make illegal profits by acting on, or leaking, inside information before the information is made public. It could be knowledge about a pending merger or acquisition, a terrorist attack affecting a particular industry, a pending legislation affecting a particular industry. Fraud has to be detected in an online manner and as early as possible, to prevent people or organizations from making illegal profits or criminal activities. [23]

- Medical and Public Health Detection

The data typically consists of patient records which may have several different types of features such as patient age, blood group, weight. The data might also have temporal as well as spatial aspect to it. The data can have outliers due to several reasons such as abnormal patient condition, instrumentation errors or recording errors. [23]

- Industrial Damage Detection

Industrial units suffer damage due to continuous usage and the normal wear, corrosion etc. Such damages need to be detected early to prevent further escalation and losses. The data in this domain is usually sensor data recorded using different sensors and collected for analysis. [23]

- Sensor Networks

Sensor networks have lately become an important topic of research from data analysis perspective, since the data collected from various wireless sensors has several unique characteristics. Outliers in such data collected can either imply one or more faulty sensors, or the sensors are detecting events that are interesting for analysts. [23]

Effects and Causes of Outliers

Outliers can be very informative about the subject area and data collection process. It is essential to understand how outliers occur and whether they might happen again as a normal part of the process or study area. It is important to resist the temptation to remove outliers inappropriately. Outlying values generally have an appreciable influence on calculated mean values and even more influence on calculated standard deviations, because of its possible inflation. [13, 35] Outliers increase the variability of the data, which decreases statistical power and may adversely lead to model misspecification and biased estimates etc. [13, 9, 35] Consequently, excluding outliers can cause results to become statistically significant. [13 ,9] Outliers lead to both Type I and Type II errors, frequently with no clue as to which effect, they have in a particular analysis. And they can lead to results that do not generalize except to another sample with the same kind of outlier. [31] Deciding how to handle these values depends on investigating their underlying cause. The appropriate efforts depend on what causes the outliers. In broad strokes, there are five main causes for outlier data entry or measurement errors, sampling problems, unusual conditions and natural variation. [13, 9]

Data Entry and Measurement Errors

Errors can occur during measurement [30, 6] e.g., instrument error, physical apparatus for taking measurements may have suffered a transient malfunction. [6] The measurement systems should be shown to be capable for the process they measure. Outliers also come from incorrect specifications that are based on the wrong distributional assumptions at the time the specifications are generated. [30, 32] Missing- value codes in computer syntax so that missing- value indicators are read as real data. [32] During data entry, generally human errors can produce weird values. Unfortunately, a common cause of outliers - particularly very extreme values - are human errors or other aberration in the analytical process. [13, 29, 32] Outliers can also arise deliberately due to fraudulent behaviour. [6] It is essential to have an access to the original record to correct the input or even remeasure the subject to determine the correct value. These types of errors are easy cases to understand. If that value is not possible it is necessary to delete the data point because it is proven. [13]

Sampling Problems and unusual Conditions

Unfortunately, the study might accidentally obtain a subject that is not from the target population. A sample may have been contaminated with elements from outside the population being examined. [6] The subject was measured under abnormal conditions. Consequently, the data was excluded from the analyses because it was not a member of the assumed population. If the analyst can establish that a subject does not represent the population, the analysing person can remove that data. However, the analysing person must be able to attribute a specific cause or reason for why that sample item does not fit the target population. [13]

Natural Variation

Natural variation respectively and natural deviation in population [6] can produce outliers and it is not necessarily a problem. However, random chance might include extreme values in smaller data sets. Hence, the process or population which are studied might produce weird values naturally. There is nothing wrong with these data points. They are unusual, but they are a normal part of the data distribution. Therefore, there is no justifiable reason to remove that value. While it is an oddball, it accurately reflects the potential surprises and uncertainty inherent in a system. When the analyst removes them, the model makes the process seem more predictable than it actually is. Even though this unusual observation is influential, it is best to left it in the model. It is bad practice to remove data points simply to produce a better fitting model or statistically significant results. If the extreme value is a legitimate observation that is a natural part of the population the analyst is studying, the analysing person should be leave it in the data set. [13]

Advices for the Removing Outliers

Sometimes it is the best to keep outliers in the data set. They can represent valuable information as a part of the study area. [13, 30] Often, values that seem to be outliers are the right or left tail of a skewed distribution. [30] Retaining these points can be hard, particularly when it reduces statistical significance. Excluding extreme values solely due to their extremity can distort the results by removing information about the variability inherent in the study area. The rejection of an extreme value on statistical grounds alone is not generally recommend. [13, 30] The non- consideration of extreme values caused the subject area to appear less variable than it is in reality. When considering whether to remove an outlier, the analyst needs to evaluate if it appropriately reflects the target population, subject- area, research question, and research methodology. It has to be clarified whether anything unusual happen while measuring these observations, such as e.g., power failures, abnormal experimental conditions, or anything else out of the norm. [13] If no root cause can be determined, and a retest can be justified, the potential outlier should be recorded for future evaluation as more data become available.

If an outlier is in question:

- In the case of a measurement error or data entry error, the correction of the value has to be done if possible. If it is not possible to fix it, that observation has to be removed.
- If the value is not a part of the presupposed population, then it is legitimate to remove the outlier. [13, 29] (Note: It cannot be statistically shown that an outlier originates from a different distribution than the rest of the data.[30])
- In the case of a value that is a natural part of the presupposed population under study, the value should not be eliminated. [13]
- Tabachnick and Fidell postulate two scenarios in which variable deletion is appropriate: (a) the variable is highly correlated with other variables or (b) the variable is not essential for the analysis. [29]

When it is decided to remove outliers, it is needed to document the excluded data points and explain the reasoning.[13] (Whereupon from some authors, removing will be seen as the most conservative and probably the safest approach to outlier management. [29]) It is inevitable for the analyst to attribute a specific cause or causes for removing outliers. [13] Another approach is to perform the analysis with and without these observations and discuss the differences. [13, 30] Comparing the results in this manner is particularly useful when the analyst is unsure about removing an outlier and when there is substantial disagreement over this question. [13] If outliers do not change the results of the analysis, they can be retained. [29]

General Strategies for the Detecting of Outliers

As mentioned, there are two general strategies for detecting of outliers. The first are the applications of informal approaches, the visualising statistical methods. The second strategy the formal approaches, which are outlier tests. These tests are intended to identify outliers and distinguish them from chance variation, allowing the analyst to inspect suspect data and if necessary correct or remove erroneous values. [9, 34] The analyst has to decide whether the cases that are outliers are properly part of the population from which you intended to sample. Cases with extreme scores, which are, nonetheless, apparently connected to the rest of the cases, are more likely to be a legitimate part of the sample. [32] The both strategies can also be applied in the case of application of robust statistic tests. [9] Robust statistical procedures which are not greatly affected by the presence of occasional extreme values, but which still perform well when no outliers are present. [13]

> Note: There is no rigid mathematical definition of what constitutes an outlier [6, 9]; determining whether an observation is an outlier is ultimately a subjective. [6]

Finding outliers depends on subject- area knowledge and an insight of the data collection process. While there is no solid mathematical definition, there are guidelines, graphs viz. methods of descriptive statistics and statistical tests which can be used to find outlier. There are a variety of ways to find outliers, [9] whereat some of which are treated as synonymous with novelty detection.[6] All of these methods employ different strategies for finding values that are unusual compared to the rest of the data set.

Methodologies to detect Outliers

Sorting Data Sets

Sorting a data set to generate ranking lists for each variable is a simple but effective way to highlight unusual values. This methodology is useful, particularly when the number of data points is not too large. This approach does not quantify the degree of abnormality of an outlier, on the other hand it will enable the analysing person to see the unusually high or low value at a glance. [13, 29] The ranking list can be supplemented by domain specific thresholds to select the most relevant suspicious value. [23]

Graphing Data Sets

Humans simply are incapable of processing information about lengthy numerical arrays. Graphs provide an effective means of downplaying the details of the data and emphasizing the important features, the distributional shape, location, presence of unusual observations, like outliers etc. Graphical methods greatly simplify the assessment of analytical data. The use of visualising statistical methods facilitates the detection of outliers, because these values are characterised by a visible distance from the remainder of the data. [13] While illustrative measures such as dot plots, histograms, boxplots etc. provide visual indications of the presence of possible outliers, it is recommended that researchers corroborate these approaches with objective quantifiable measures to ensure accurate outlier identification. [29] On the other hand, graphs can be somewhat detrimental in some situations because it often is difficult to recover the numeric values from the visual display. In contrast, graphs usually are superior for revealing patterns, trends, and relative quantities within data sets regardless of their size. All numerical summaries of data are based on assumptions about the nature of those data. If these assumptions are met, then descriptive statistics provide an accurate representation of the data features. But in the extent the assumptions are not met, descriptive statistics can be inaccurate and misleading. Graphical presentations are not nearly as reliant on such underlying assumptions and so they can be used to summarize the data without the attendant dangers of misrepresentation. Another advantage is that graphical analysis facilitates greater interaction between the analyst and the data. Effective visual presentations highlight interesting and unusual aspects of the quantitative information under investigation. This encourages the researcher to pursue these features to identify their sources and implications for understanding the processes that are generating the data. [31]

Histograms

The histogram [29] which is also known as frequency distribution, [34] or frequency histogram [32] is by far the most commonly used procedure for displaying data. A histogram is a graphical display that is used to demonstrate central tendency distributions, the relative concentration or "density" of observations. The morphology of histograms typically mimics a normal distribution with a cluster of cases near the mean and a trail of cases heading toward both ends of the distribution. [29, 32, 34] Many naturally occurring things have this shape of distribution. [34] The data density at any specific location within the whole range is represented by the vertical height of a point. A histogram is usually presented as a vertical bar chart showing the number of observations in each of a series of intervals. The horizontal axis is divided into segments corresponding to the intervals. Whereat intervals are often called "bins”. On each segment a rectangle is constructed whose area is proportional to the frequency in the group. The areas under the histogram can be interpreted as probabilities such as the area covered by each "bar". The histogram provides a great deal of information about the distribution in a very concise manner. Histograms are all somewhat sensitive to the choice of intervals. Even relatively small changes in start point or interval width can noticeably change the appearance of the plot, especially for smaller data sets. [13] The problems stem from the arbitrary nature of the “bins” used to categorize the continuous data values. If the y- value is small, relative to the range of the data, in combination with narrow bins, then the histogram will follow the contours of the distribution closely. The empirical representation of the distribution could be quite "bumpy." Alternatively, wider bins in combination with larger y- values produce a smoother histogram. But they also increase the risk of distorting substantively important features -like outliers- in the distribution of the variable. The problem is that wide “bins” eliminate any possibility of showing local variations in the densities contained within the respective bins. [9, 29] Most frequently histograms display no more than about 20 bars. [33] The very use of the “bins” is a distortion of information because any data variability within the “bins” cannot be displayed in the histogram. At the same time, the discrete nature of the “bins” generates discontinuities that are manifested visually in the sharp corners of the histogram bars; the latter certainly, are not an intrinsic part of the data. Histograms also emphasize the existence of outliers. [9, 29] The basic idea is to visualise the data distribution for a single variable and find values that fall outside the distribution. [9, 29] It is recommended that histograms be utilized only as a preliminary assessment in the search for outliers. This is because, except in extreme outlier cases, the use of histograms may not be definitive. [29]

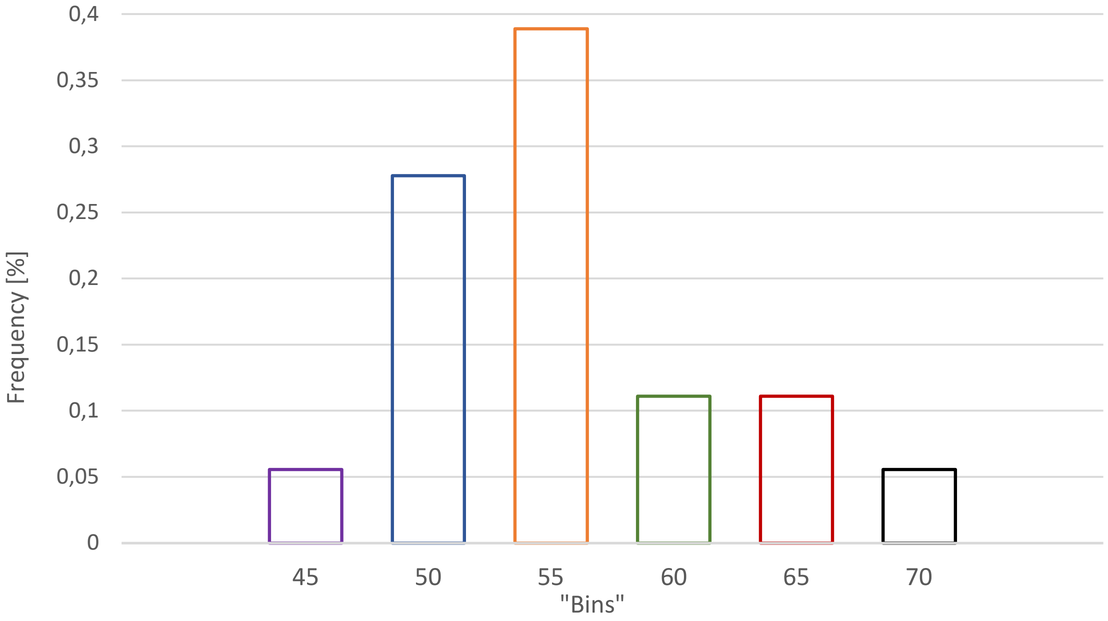

Figure 1 shows a histogram with the thiamphenicol data from [13]

Unidimensional Scatterplots

The chart below using unidimensional scatterplots (Figure 2) for each of four variables. Unidimensional scatterplots show clearly that the variable distributions differ from each other in important and easily recognizable ways. For anyone confronted with the information in this graphical form, there is no question that the variables have divergent distributions. Instead, attention would centre on the more interesting question of why the four sets of values show such differences from each other. This brief example illustrates very nicely the general advantages of graphical approaches to data analysis. The graphs provide useful summaries for large, complicated data sets.

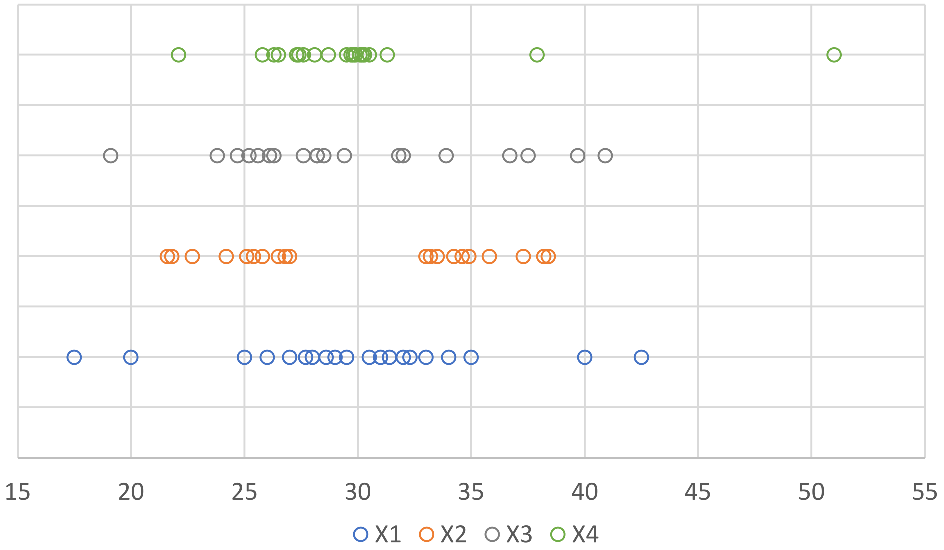

Figure 2 shows an even a cursory glance at the Figure reveals that X1 has a unimodal, symmetric distribution, whereas the distribution for X2 is symmetric but bimodal. At the same time, X3 distribution is skewed positive, whereas X4 distribution is compressed to the left, but offset by a single outlying observation (Outlier) with an extremely large value. The sample arithmetic mean and standard deviation are accurate summaries of the distribution for X1, but they seriously misrepresent the other four variables.

A unidimensional scatterplot simply shows each observation as a point plotted along a scale line that represents the range of data values. This type of graph can convey a great deal of information without the potential loss of information or distortion encountered in a histogram. The main drawback of a unidimensional scatterplot is that it is effectively limited to small data sets. With large numbers of observations, there is a drawback of overplotting. This makes it difficult to discern individual observations and concentrations of data points within the overall distribution. There are two general and mutually supportive strategies for minimizing the effects of overplotting. First, it is important to select a plotting symbol that allows readers to detect overplotted points. The relatively large open circles are effective for this purpose. Small and or solid points would coalesce into incomprehensible blobs within the display. Similarly, if the plotting symbols had straight sides (e.g. squares), then it would be more difficult to separate them visually into individual data points when they overlap within the display. The overplotting can be reduced by displacing the points somewhat in the direction perpendicular to the scale line of the variable. This Process is called “jittering”. In a “jittered” unidimensional scatterplot, it is important to keep the range of the random variation small relative to the variation in the substantive variable. [13]

Dot Plots

A dot plot (Figure 3) which also called index plot is a useful display method whenever data values are associated with identifying information such as a label or an index number. Dot plots are useful in a variety of situations and there are several different versions of the basic display. One axis of the dot plot, usually the horizontal, represents the scale for the variable under investigation. The other axis, usually the vertical, contains rows that provide a label or an index for each data value. Observations are sorted according to the values of the variable under investigation and then plotted as points at the appropriate scale location within each row. In this manner, the dot plot identifies both the specific data points and the numeric values that are associated with them. The dot plot is effectively the same as a transposed quantile plot or Q-Q- Plot and the shape of the point array can be interpreted in a manner similar to those displays. [13]

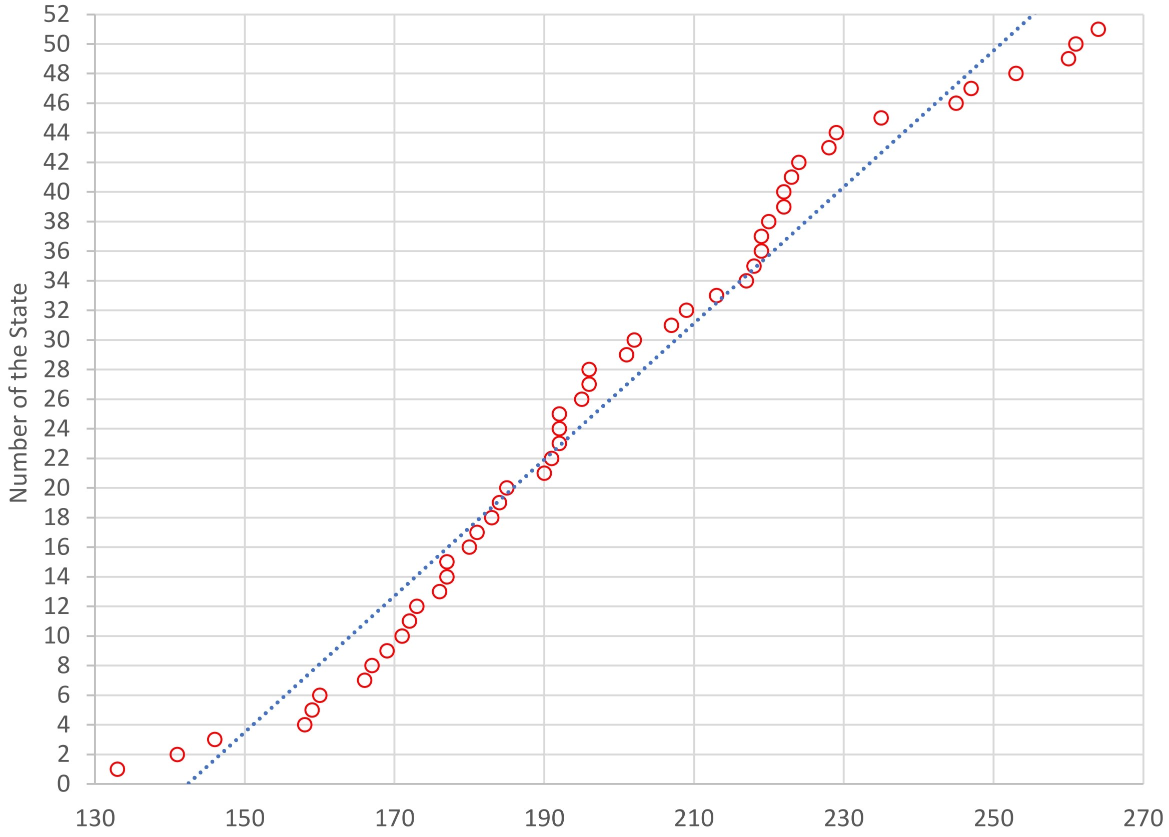

Figure 3 shows a bimodal distribution (corresponding to the nearly vertical sections in the point array) and the asymmetric tails (in the long and fairly steep point array in the lower left side of the graph and the shorter, shallower string of points near the upper right- hand corner). The dot plot succinctly provides a great deal of information about the distribution and the specific observations within the data like outlier. In the case of a unimodal distribution the graph would show a plot shaped like a "transposed S." with a more “linear part” in the middle. Note: The modality of a distribution concerns how many peaks. A distribution with a single peak - that is, one value with a high frequency - is a unimodal distribution. Multimodal distributions have two or more peaks and when there are exactly two peaks, the distribution is bimodal. [33]

The dot plot is a graphical processing task that human observers can carry out quite accurately. By contrast, pie charts require analysts to make comparisons between the angles, arcs and areas that define the sizes of the pie wedges. Like dot plots, bar charts also require judgments about locations along the scale. In summary, dot plots are excellent graphical displays for labelled data. They contain a great deal of information, concerning distribution, outliers, are easy to interpret and overcome a number of the problems associated with other kinds of displays. For these reasons, they should be used frequently in empirical research. [31]

Background of the Parametric Statistical Methods

Visual inspection alone cannot always identify an outlier and can lead to mislabelling an observation as an outlier. Because data are used in estimation with classical measures (parametric statistic) such as the arithmetic mean being highly sensitive to outliers, statistical methods were developed to accommodate outliers and to reduce their impact on the analysis. [30] Parametric statistical tests - like the later on described outlier test - assume an underlying normal distribution specified by the arithmetic mean and the standard deviation and where this assumption is violated, the results of such tests will be unreliable due to a loss of statistical power. It is therefore necessary to test if such an assumption is valid before proceeding to analyse the data.

- The normal probability plot is a graphical method on the basis of a non- parametric statistic approach for assessing whether or not a data set is approximately normally distributed.
- The probability plot correlation coefficient (PPCC) is a test statistic of the linearity of the relationship between two variables for assessing whether or not a data set is approximately normally distributed. [18]

Against the background of the importance of the parametric statistical methods, like the application of the Zscore and parametric tests, the fundamental central limit theorem has to be de described briefly.

Central Limit Theorem

In the case of an infinite large size of numbers of samples, the normal distribution respectively the Gauss- distribution represents the central limit theorem [42, 1, 43] developed by Pierre-Simon (Marquis de) Laplace (1778) (1749 -1827) [44] The law of large numbers is a natural law [45] and fundamental for the inductive statistics. [43] The law of large numbers represents formally the convergence of the means. [46] It is necessary to underscore that probability density functions in general are mathematical models which embodies a set of statistical assumptions concerning the generation of sample data and similar data from a larger population, consequential density functions are approximation functions to describe the reality. Also, valid here is the principle: A mathematical model is only as strong as its underlying assumptions. [1, 47, 43, 45, 48] The use of the central limit theorem makes it possible to describe the arithmetic mean, the standard deviation the confidence interval of the theoretical distribution and resulting from this the probability of the existence of outliers.

Continuous Probability Density Function

The law of large numbers says, if it is taken more samples from any population (formular 1), then the mean of the sampling distribution (formular 2), tends to get

$$\mu = \frac{1}{N}\sum_{k=1}^{N} x_i$$

Formular 1

$$\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_i$$

Formular 2

closer to the true population mean, the population standard deviation (s) (Formular 3) (Whereby σ represents also the inflexion point of the function of Gauss- distribution on the x- axis [42, 49]) and sample standard deviation (σ) (Formular 4), consequently closer to reality. [4]

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

Formular 3

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

Formular 4

Both results of formular 3 (σ) and formular 4 (s) represent the variability of the random variable (x) around the expected value (μ, $\bar{x}$). [49]

Confidence Interval and Probability of Outliers

The determination of the confidence interval is based on the arithmetic mean $\bar{x}$ of the sample to estimate the arithmetic mean of population (μ) and the standard variation (s) to estimate the population standard deviation (σ). The influence of the sample size and the knowledge of the distribution are also essential in these analyses. [43] The confidence interval is an interval estimation. [42, 1, 50, 47, 43, 49] The confidence interval contains the most plausible values of the unknown parameter of interest. [47] This interval estimate for the unknown population parameters depends on: the desired confidence level, information that is known about the distribution, the sample and its size. Concerning to the approximate standard normal distribution with a known population standard deviation (σ) it is mostly common to define a confidence interval of 95% of the samples which will be within a confidence coefficient of $Z\alpha = 1.960$ which represents a standard deviation 1,96 σ of the population mean μ.

This confidence interval implies two possibilities: Either the interval contains 95% of the true mean μ and samples produced an $\bar{x}$ that is within the interval of the true mean μ. The second possibility happens for 5% (α_{error}) of the samples, because they are outside of the interval. Concerning the labelling of outliers, the α_{error} represents the existence of outliers. In the context of outlier detection, the α_{error} will be labelled α_{out}. The relationship between the total probability, the confidence interval (CI) and the probability of the existence of outliers (α_{out}) can be written as follows (one – side interval estimation):

$$1 = \mathrm{CI} + \alpha_{\mathrm{out}}$$

Formular 5

In the case of a two - side interval estimation as follows:

$$1 = \mathrm{CI} + \frac{\alpha_{\mathrm{out_{left}}}}{2} + \frac{\alpha_{\mathrm{out_{right}}}}{2}$$

Formular 6

The confidence coefficient $Z\alpha$ is the number of standard deviations where the outlier lies from the mean with a certain probability. The most convention in economics, technical- and also most social sciences sets confidence levels at either 90 % ($Z\alpha_{out}= \pm 1.645$, $\alpha_{out}= 10\%$), 95% ($Z\alpha_{out}= \pm 1.960$, $\alpha_{out}= 5\%$) certainty is considered as probable respectively significant [47], 99% ($Z\alpha_{out}= \pm 2.576$, $\alpha_{out}= 1\%$) security is considered as significant respectively “very significant” [47] and 99.9 % ($Z\alpha_{out}= \pm 3.290$, $\alpha_{out}= 0.1\ \%$) security is considered as “highly significant.” [42, 1, 50, 47, 43, 49] Like the level of confidence the probability of the existence of outliers must be pre- set and not subject to revision as a result of the calculations. [51, 1] The difference concerning the probability of the existence of outliers and the most often applied significance level α or $\alpha_{error} = 0.5$ is that $\alpha_{out} \leq 0.5$ and that the value is mainly given by the confidence coefficient ($Z\alpha_{out}$) and not in percent. Next to confidence coefficient $Z\alpha$, for the labelling of outliers, some authors named also a value of a threshold of $Z\alpha_{out}$ respectively of $\sigma= \pm 2.576$ ($Z\alpha_{out}= \pm 2.576$, $\alpha_{out}= 1\%$). [34] A standard cut- off value for finding outliers is a $Z\alpha_{out}$ of ± 3, respectively $\sigma = \pm 3$, the three- sigma rule ($Z\alpha_{out}= \pm 3.000$, $\alpha_{out}= 0.27\ \%$). This rule denotes that roughly 1 in 370 observations will differ by three times the standard deviation. [10, 33, 6] Similar to the value which represents the term “very significant”, the value of $\sigma= \pm 3.290$ ($Z\alpha_{out}= \pm 3.290$, $\alpha_{out}= 0.1\ \%$) can be found in the literature to identify outliers. [29, 32, 34] The author argued that: “The extremeness of a standardized score depends on the size of the sample; with a very large number of data (n), a few standardized scores in excess of 3.29 are expected.” [32].

$$Z_{\alpha_{out1}} = \pm 1.96 = \alpha_{out} = 5.00\% \qquad [13, 40]$$

Formular 7

$$Z_{\alpha_{out2}} = \pm 2.58 = \alpha_{out} = 1.00\%$$ [13, 34, 40]

Formular 8

$$Z_{\alpha_{out3}} = \pm 3.00 = \alpha_{out} = 0.27\%$$ [6, 10,33]

Formular 9

$$Z_{\alpha_{out4}} = \pm 3.29 = \alpha_{out} = 0.10\%$$ [29, 32, 34]

Formular 10

In order to check whether data points are outliers, they have to be converted to Z_{score}, which will also called "Studentization". Z_{score} -analysis is an important and objective way to determine whether a suspected outlier is truly a concern. [29] If the population is assumed to be normal, the "Studentization" respectively conversion can be applied. [11] The Z_{score} can quantify the unusualness of an observation. Z_{score} and $Z_{\alpha out}$ are the number of standard deviations (σ) above and below the arithmetic mean $\bar{x}$ that each value falls. [10] To calculate the Z_{score} for an observation, it is necessary to take the raw measurement (x), subtract each by the arithmetic mean ($\bar{x}$) and divided by the standard deviation (S).

$$Z_{score} = \frac{x - \bar{x}}{S}$$

Formular 11

If the absolute value Z_{score} of a data is greater than the absolute value of the chosen cut- off value $Z_{\alpha out}$ the suspicious data can be labelled as a potential outlier.

The Z_{scores} can mislead with small data sets (n) because the maximum Z_{score} is limited. The hereafter quoted equation describes the maximal achievable Z_{score} as function of the number of samples

$$Z_{score(max.)} = \frac{(n-1)}{\sqrt{n}}$$

[12] Formular 12

Unfortunately, the value of $Z_{score(max.)}$ is quite limited for small data sets (n). When n ≤ 10 the $Z_{score(max.)}$ cannot exceed σ = ± 3 ($Z_{\alpha out}$ = ± 3) regardless of the combination of values. Consequently, no value can be detected as an outlier according to the three- sigma rule. [12] Therefore it is possible to name the minimum number of data n >10 to apply the three- sigma rule for the labelling of outliers. The presence of the outlier influences the Z_{score} because it inflates the arithmetic mean $\bar{x}$ and standard deviation S. When the Z_{scores} will be calculated without the outlier, the values would be different. If a data set contains outliers, Z_{scores} will be biased therefore they appear to be less extreme (i.e., closer to zero). It is not appropriate to consider a Z_{score} as being approximately normally distributed in any cases. The Z_{score} is not satisfactory for outlier labelling, especially in small data sets. Although the basic idea of using the Z_{score} is a helpful approximation, they are unsatisfactory because the summaries x and σ are not resistant in respect of outliers. [21] However, if the data do not follow the normal distribution, this approach might not be accurate. In general, it is assumed that the number of data exceed 30. Incipient with a number greater than 30 is can be assumed that the data are approximately normally distributed and the Z_{score} is a satisfactory estimation for outlier labelling. [22, 34] When the number of samples get fewer than 30, the sampling distribution has a different shape and can be considered suitable e.g. the t- distribution.

Note: Concerning the setting of limit values like the threshold of σ ($Z_{\alpha out}$) and the significance in general, it is worth remembering the words of Sir Ronald Aylmer Fisher:

> Fisher acknowledged that the dogmatic use of a fixed level of significance was silly: 'no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas' (Fisher, 1956). [34]

In other cases, robust methods should be used like the estimator MAD (the median of the absolute deviations about the median) etc. [21, 13]

Modified Zscore

This method uses two estimators for outlier labelling, the median $\tilde{x}$ and median absolute deviation MAD instead of arithmetic mean $\bar{x}$ and the classical standard deviation S to resolve the limitation of Z_{score} in which standard deviation S can be affected by extreme observation. [35]

$$\text{MAD} = \text{median } |xi - \tilde{x}|_{i = 1,2, \ldots , n}$$

Formular 13

Whereat $\tilde{x}$ is the sample median and MAD being the sample median absolute deviation. The median absolute deviation (MAD) is not directly comparable to the classical standard deviation (S). Under the condition that the underlying distribution is approximately normal, the MAD can be modified to provide approximated standard deviation S.

$$\text{MAD} \approx \frac{\delta}{1.483} \approx \frac{S}{1.483}$$

Formular 14

The approximated standard deviation S on the basis of the MAD is called MAD_E

$$MAD_E = 1.483 \text{ MAD}$$

[13, 21] Formular 15

The modified Z_{score} is denoted by Mi and is calculated as follows

$$\text{Mi} = \frac{|xi - \tilde{x}|_{i = 1,2, \ldots , n}}{\text{MAD}_\text{E}}$$

Formular 16

Iglewicz and Hoaglin proposed that the absolute values of Mi greater than 3.5 i.e. |Mi| > 3.5, the observation is considered as an outlier. [21, 35]

Notes:

1. The MAD and MAD_E share the disadvantage that they both become zero if more than half of the data set are equal, perhaps because of excessive rounding or a large number of zero observations. This would, of course, be a problematic data set in any case.

2. Because of the limitations of MAD_E, it is sometimes useful to use the arithmetic mean absolute deviation instead of the median absolute deviation. Although this is less robust than MAD_E, it does not become zero unless all the values in the data set are identical. sMAD is a

compromise; if MAD_E is non-zero, sMAD = MAD_E; if MAD_E is zero, sMAD is the arithmetic mean absolute deviation. [13]

On the basis that MAD_E approximately represents the standard deviation of the Gaussian- distribution σ a threshold of |Mi| > 3.5 represents the probability of the existence of one outlier of 0,00012 (0,012 %) for each side and 0,00022 (0,022 %) of both sides of the Gaussian- distribution.

Median Absolute Deviation (MAD_E) Method

The method MAD_E is a robust technique that uses median $\tilde{x}$ and median absolute deviation MAD instead of arithmetic mean $\bar{x}$ and standard deviation S, as they are highly unaffected by extreme observations. This technique is defined as follows:

$$2MAD_E \text{ Method: } \tilde{x} \pm 2MAD_E$$

Formular 17

$$3MAD_E \text{ Method: } \tilde{x} \pm 3MAD_E$$

Formular 18

The values that lie outside the interval of $\tilde{x} \pm 2MAD_E$ or $\tilde{x} \pm 3MAD_E$ are considered as outliers. [35]

Q-Q- Plots

A special example of a scatter plot, [13] dot- plot or index- plot [31], is a normal probability plot [13, 32], sometimes also called a quantile- quantile plot “Q- Q- Plot” [13] or “normal – plot”. Whereat the Q-Q- plot represents the most common variant. [16] The use includes the identification of skewness, kurtosis, [6, 15, 20] a need for transformations and also the detection of outliers. [6, 15, 32, 34]

> Note: Quartiles divide a distribution into fourths, percentiles divide a distribution into one hundredth and deciles divide it into tenths. [33] The quantiles of a distribution are a set of summary statistics that locate relative positions within the complete ordered array of data values. [31, 32]

A point on the plot corresponds to one of the quantiles of a distribution plotted against the same quantiles of the reference distribution, with which the first will be compared. [14, 31]

Q-Q- plot is formed by:

y- axis: z_i- scores of the observation

x- axis: z_i- score of the inverse cumulative function (Φ^{-1}) of the reference distribution [16]

This defines a parametric curve where the parameter is the index of the quantile interval. To produce a probability plot, the order statistics of the observed values or the transformed order statistics has to be generated. [18] This calculation of the quantiles (pi) of the observations has to be plotted based on the ranks. [14, 32] One way of forming approximate normal scores for n data points $x_1 \ldots x_n$, of the uniform order statistic medians p_i [16], is as follows [13]:

1. Obtain the ranks r_i for the data set. [13]

2. The normal probabilities p_i, for each data point respectively rank. [13] Different sources quote diverse formulars for the approximation, calculation of the quantiles [16]. The formula used by the basic "stats" package in R for that continuity correction [20] is as follows:

$$p_i = \frac{(r_i - \mathrm{a})}{(\mathrm{n} - 1 + 2\mathrm{a})}$$

Formular 19

where r_i is the rank of the data, a is set to 0,375 and n the number of values. If the number of value n is less than or equal to 10 then a is set to be 0,5 otherwise. [13, 15] Others sources recommend the following equation, developed by Blom (1958):

$$p_i = \frac{(r_i - 0{,}375)}{(n + 0{,}25)}$$

[14,18,19] Formular 20

Most references quote the following formular:

$$p_i = \frac{(r_i - 0{,}5)}{n}$$

Formular 21

The last of these ranks r_i, in this equation, corresponds to the 100th percentile which represents the maximum value of the theoretical distribution, which is sometimes infinite. [14, 20] In this equation, the quantity 0.5 is subtracted from each r_i value in the numerator to avoid extreme quantiles of exactly 0 or 1. The latter would cause problems if empirical quantiles were to be compared against quantiles derived from a theoretical, such as the normal. This adjustment has no effect on the shape of any graphical displays that use the quantiles. [31]

3. The uniform order statistic medians pi and the percent point function [16], which is also called the inverse of the cumulative distribution function (Φ^{-1}) and "Probit" [17], is needed to generate the x-values of the Q-Q- plot. The Φ^{-1} – function generates the probability, based on the calculation of the uniform order statistic medians, the pi - values. [16] Φ^{-1} gives the $Z_{i\text{-score}}$ associated with probability pi values between $0 \leq pi \leq 1$ onto a standard normal distribution. [17] If one or both of the axis in a Q- Q- plot is based on a theoretical distribution with a continuous cumulative distribution function (CDF), all quantiles are uniquely defined and the $Z_{i\text{- score}}$ of the normal distribution can be obtained by Φ^{-1}. [14]

$$\mathrm{z}_{\mathrm{i-score}} = \Phi^{-1}, \mathrm{f}(\mathrm{p}_\mathrm{i})$$

Formular 22

Whereat Φ^{-1} is for the Gaussian- distribution as function of the probabilities pi. [14] The Q-Q- plot based on the presentation of the $Z_{i\text{-scores}}$ grounded on Φ^{-1} as x- values and the $Z_{i\text{-scores}}$ of the measurement as y- values.

The diagram below shows the different $Z_{i\text{-scores}}$ for Φ^{-1}, for the Gaussian- distribution and the $Z_{i\text{-scores}}$ of the measurement calculated based on the procedure of "Studentization" of measurements [11]. Both scores a pictured as function of the probabilities pi based on the quantiles.

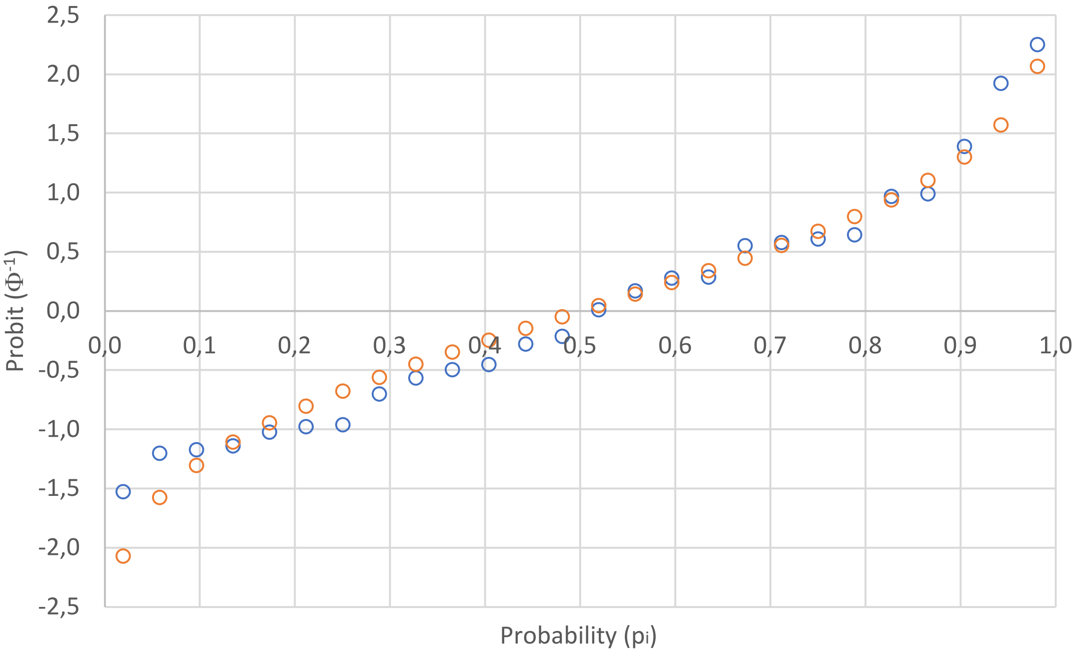

Figure 4 displays a Q-Q- Plot, whereby the orange dots represent the inverted cumulative distribution function (Φ^{-1}) for the Gaussian- distribution and the blue dots represent the $Z_{i\text{-scores}}$ of exemplary measurements, as function of the probability pi based on the quantiles. Whereat the y- axis represents the $Z_{i\text{-scores}}$ of the measurement and the $Z_{i\text{-scores}}$ of the inverted cumulative distribution function (Φ^{-1}) for the Gaussian- distribution.

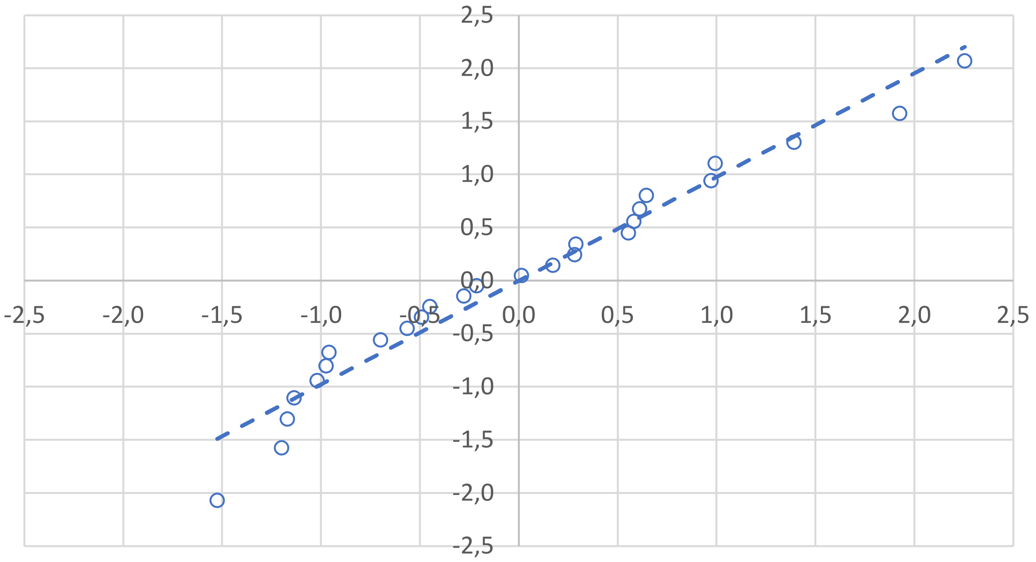

Figure 5 shows the plotting Z_{score} against xi- values gives the Q-Q- plot. [13] Whereat the values of the x- axis in contrast to first graph are not standardised.

The points plotted in a Q- Q- plot are always increasing when viewed from left to right. [20, 32] If the two distributions being compared are identical, the Q–Q plot follows the 45°- line (y = x) the angle bisector [20]. If the two distributions agree after linearly transforming the values in one of the distributions, then the Q- Q- plot follows some line, but not necessarily the line y = x. If the general trend of the Q- Q- plot is flatter than the line y = x, the distribution plotted on the horizontal axis is more dispersed than the distribution plotted on the vertical axis. Conversely, if the general trend of the

Q- Q- plot is steeper than the line y = x, the distribution plotted on the vertical axis is more dispersed than the distribution plotted on the horizontal axis. Q- Q- plots are often arced, or "S" shaped, indicating that one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other. [14] Deviations from the line (usually at either end) indicate some deviation from normality. If the data points fall on or close to a straight line, the data are close to be normally distributed. This guide line is shown as the dashed line in Figure 5. In this case, most of the points fall fairly close to the line; only the slight curvature in the data suggests any non- normality. [13] If normality is present, the residuals are normally and independently distributed around the 45° - line (y=x). [32]

Box- Plot and Interquartile Range

A Q-Q- plot shows all of the data. Sometimes, however, this degree of detail is not necessary in a graphical display. [31] A Box- Plot - often also called a 'box- and- whisker' plot or "box- and- whisker diagram" - is a useful method of summarising data sets, particularly where the data fall into different categorical grouping variable and continuous variable. [13, 28] Box- Plots are based on the quantiles of a distribution. Analysts frequently use them during data analysis because the displayed data set shows the "most important" quantiles respectively the characterising data. The Box- Plot represents, the central tendency, [31] dispersion, skewness, and spread, around the median $\tilde{x}$ as well as highlighting outliers. The interquartile range (IQR) - the hight of the box - is a measure of the spread and dispersion of the data. [28, 29, 30, 31, 32, 33, 34] The symmetry of the distribution is indicated by the relative distances from the median line to the upper and lower edges of the box and also by the relative sizes of the two whiskers. The box shows the central region of the distribution. [31] The used quartiles and interquartile range (IQR) have the advantages to be also relatively robust compared to other quantitative methods concerning the detecting of outliers. [9, 31] The Box- Plot displays outliers using asterisks that fall outside the subsequent described “whiskers”. These graphs are often precursors to hypothesis tests. [28] The different features can represent different statistics, but the most common choice is the five- number summary as follows [13, 28]:

1. The minimum value of the data set. [26, 28, 31]
2. The first or lower quartile (Q1) which represents 25th percentile of the data set. [26,28, 29, 30, 31, 34]
3. The central solid line inside of the box is the median $\tilde{x}$ [13, 21] which represents 50th percentile of the data set. Whereat the median is also called the second quartile. [34] The median is a measure of central tendency in statistics. [26, 28, 29, 30, 31, 33, 34].
4. The third or upper quartile (Q3) which represents 75th percentile of the data set. [26, 28, 29, 30, 31]
5. The maximum value of the data set. [26, 28, 31]

These five values highlight the data distribution shape, spread, and central tendency. All these measures are nonparametric and do not make assumptions about the data distribution. This aspect makes a box and whisker plots especially suitable for the early stages of analysis. This graph works by breaking the data set down into predefined quartiles. When the sample size is too small, the quartile estimates might not be meaningful. Consequently, these plots work best when at least 20 data points per group are available. [28] The bottom and top of the rectangular “box” show the lower and upper quartiles, respectively. The box shows the range of the central 50% of the data set. The length of the box is the interquartile range (IQR), the indicator of the dispersion of the data. [13] It represents the range of values between the third quartile (75%) and the first quartile (25%) (IQR= Q3 – Q1= $Q_{0.75}$ – $Q_{0.25}$), that equates 50%. [13, 9, 21, 26, 28, 29, 31, 33] Percentiles respectively quartiles indicate the percentage of data that fall below a particular value and it describes the relative standing of a value. [28] One formula for finding first quartile (Q1) and third (Q3) is quoted below. Whereat i is the index of a data value and can be calculated on the bases of number of date n.

$$i = \frac{(\lfloor (n+1)/2 \rfloor) + 1}{2}$$

Formular 23

To get the lower quartile, Q1, with an ascending counting of i from the top of the list. The position i can be calculated as follows [21]

$$n\ odd\ :\ \frac{x_{(\lfloor i \rfloor)} + x_{(\lfloor i+1 \rfloor)}}{2}$$

Formular 24

$$n\ even\text{:}\ x_{(\lfloor i \rfloor)}$$

Formular 25

To get the upper quartile, Q3, with an ascending counting of i from the bottom of the list. The position i can be calculated as follows [21].

$$i = \frac{(\lfloor (n+1)/2 \rfloor) + 1}{2}$$

Formular 23

When data are arranged in ascending order, the median $\tilde{x}$ is a number that measures the "center" of the data set it represents the 50th percentile and the central solid line inside of the box. For symmetrical distribution, the arithmetic mean $\bar{x}$ and median $\tilde{x}$ have the same expected value. [13] The median $\tilde{x}$ as the "middle value," but it does not actually have to be one of the observed values. It is a number that separates ordered data into halves. If n is an odd number - similar to the calculation of Q1 and Q3 - the median is the middle value of the ordered data. If n is an even number, the median is equal to the two middle values added together and divided by two after the data has been ordered. [1,13]

$$n\ odd\ :\ x_{(n+1)/2}$$

Formular 26

$$n\ even\text{:}\ \frac{x_{n/2} + x_{(n+2)/2}}{2}$$

Formular 27

The lines extending upwards and downwards from each box - the “whiskers”- are drawn from the end of the box to the last data point within an adjustment factor k= 1.5, which represents 1.5 times the interquartile range IQR of the box. [13, 31]. In those cases, the whiskers are not extending the minimum and maximum values, these data are considered and marked as outlier values [26, 28, 30, 31]. The adjustment factor k= 1.5 is used to calculate boundaries for what constitutes mild outliers and in the case of a k= 3.0 for extreme outliers. [25, 26, 21, 29, 31]

lower inner fence: $Q_{0.25}$ – 1.5 IQR

Formular 28

upper inner fence: $Q_{0.75} + 1.5$ IQR

Formular 29

The k= 3.0 rule is quite conservative, implying that far out values can be comfortably declared as outliers when the data are assumed to come from random normal samples.

lower outer fence: $Q_{0.25} – 3.0$ IQR

Formular 30

upper outer fence: $Q_{0.75} + 3.0$ IQR

Formular 31

Observations flagged as outside require further study to determine their causes.[21] The number that accompanies an outlier data point is known as the case identification number. [29] The k= 1.5 and k= 3.0 rule should not be used alone to declare outside observations as defective. [21] The lines at the end of the “whiskers” are often terminated “fences” with a horizontal line. [13] Each whisker contains 24.651% of the distribution. [28] Individual observations outside the “fences” are drawn as separate points on the plot, these observations are outliers. [13, 21]. These fences are "imaginary values" that usually do not occur within the empirical data. They are only used to obtain the upper and lower "adjacent values". [31] Box plots display asterisks on the graph to indicate when data sets contain outliers. [6, 9] For a normal distribution, observations outside the “fences” are expected about 0.7 % in the case of k= 1.5, so individual points outside the fences are generally considered to be outliers [13] or mild outliers [33]. Mild outliers are shown with circles. [33] Observations outside the “fences” are expected about 0.002 % in the case of k= 3.0, so individual points outside the fences are generally considered to be extreme outliers. [33] Cases that are extreme outliers are shown with asterisks. [33] Even when data are not normally distributed, a box- plot can be used because it depends on the median and not the arithmetic mean of the data. [30] It is important to identify unusual and problematic aspects of the data. At the very least, the presence of outside values should lead the analyst to inspect these observations more closely. The only real drawback of a box- plot is that it is fairly insensitive to multiple modes within the data. But beyond this limitation, the box- plot crams a great deal of information into a concise and easily understood visual display. Because of this, it probably is the second most frequently used graphical method for data, behind only the histogram in popularity [31]

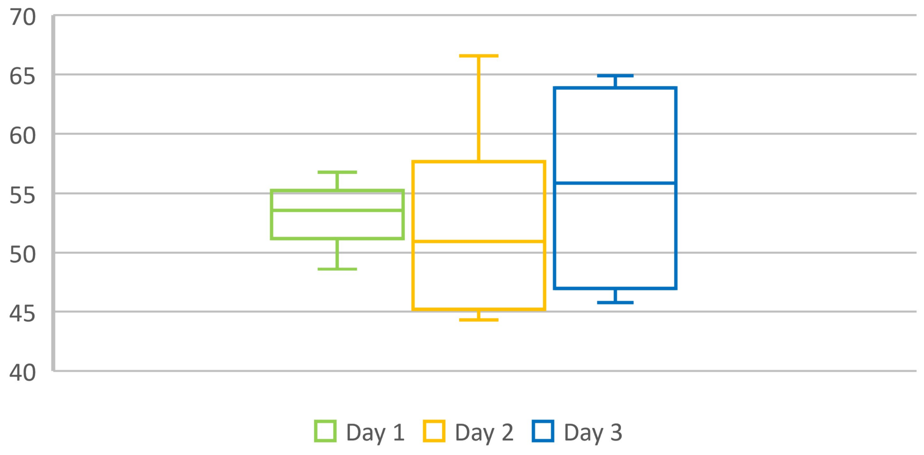

Figure 6 displays a Box plot of thiamphenicol data by day. [13]

An advantage of the boxplot, in comparison to the histogram, is that identification of outliers is based on statistical methods rather than subjective ‘‘eyeballing.’’ [29]

Statistical Tests to identify Outliers

In contrast to the previous described informal methods, the formal methods are test- based methodologies that usually require a statistical test to test a hypothesis. [35] Hypothesis testing is a statistical analysis that uses sample data to assess two mutually exclusive theories about the properties of a population. Statisticians call these theories the null hypothesis and the alternative hypothesis. A hypothesis test assesses a sample statistic and factors in an estimate of the sample error to determine which hypothesis the data support. [9] This paper deals with tests on the basis of the Gaussian-distribution. Therefore, a formal method to examination whether the distribution is normal will be described.

Preliminary Considerations

Before describing individual tests, it is useful to consider what action should be taken on the basis of outlier tests. [9] A positive outcome from an outlier test is best considered as a signal to investigate the cause; usually, outliers should not be removed from the data set solely because of the result of a statistical test. However, experience suggests that human or other error is among the most common causes of extreme outliers, as described. This experience has given rise to fairly widely used guidelines for acting on outlier tests on analytical data, based on the outlier testing and inspection procedure included in ISO 5725 Part 2 for processing interlaboratory data. The main features are:

1. Test at the 95% and the 99% confidence level. [13, 40]
2. All outliers should be investigated and any errors corrected.
3. Outliers significant at the 99% level may be rejected unless there is a technical reason to retain them.
4. Outliers significant only at the 95 % level should be rejected only if there is an additional, technical reason to do so.
5. Successive testing and rejection are permissible, but not to the extent of rejecting a large proportion of the data.

This procedure leads to results which are not seriously biased by rejection of chance extreme values, but are relatively insensitive to outliers at the frequency commonly encountered in measurement work. Note, that this objective can be attained without outlier testing by using robust statistics where appropriate. It is important to remember that an outlier is only “outlying” in relation to some prior expectation. If the data were e.g. Poisson distributed, many valid high values might be incorrectly rejected because they appear inconsistent with a normal distribution. It is also crucial to consider whether outlying values might represent genuine features of the population. It follows that outlier testing needs careful consideration where the population characteristics are unknown or, worse, known to be nonnormal. The most important role of outlier testing is to provide objective criteria for taking investigative or corrective action. Outlier tests are also used in some circumstances to provide a degree of robustness. [13]

The Probability Plot Correlation Coefficient Test

The probability plot correlation coefficient (PPCC) is used as a test statistic of Gaussian- distribution. This methodology is a test statistic on the basis of the linearity of the relationship between two variables. The null hypothesis for the PPCC- test is that the data are normally distributed with the PPCC-test statistic the correlation coefficient r. The PPCC defined as the product moment correlation coefficient between ordered observation xi and the yi. [18, 39]

$$PPCC(x,y) = r(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Formular 32

Where PPCC = 1 the data are perfectly normal distributed, while PPCC = 0 indicates no correlation and following on that no normal distribution. The PPCC is compared to a critical value cv for a specified level of significance α and sample size n. If the PPCC is less than the critical value (cv), the null hypothesis that the data is normal distributed can be rejected. Statistical tables typically give critical values (cv); but approximated values as function of n and a significance level of α= 0.05 are given by the formular mentioned below:

$$cv(n, \propto = 0{,}05) \approx 1{,}0063 - \frac{0{,}1288}{\sqrt{n}} - \frac{0{,}6118}{n} + \frac{1{,}3505}{n^2}$$

[18] Formular 33

The PPCC- test provides in conjunction with the associated probability plot, a quantitative and graphical representation of goodness- to- fit. The advantages of the PPCC- test is that the test statistic is conceptually easy to understand. It combines two fundamentally simple concepts: the probability plot and the correlation coefficient. [39] Where the data does not fit a theoretical distribution, nonparametric statistical tests should be used. Nonparametric tests require fewer assumptions about the data and as they do not rely on the underlying distribution they are often referred to as distribution-free. Nonparametric tests can be applied to all distributions.

Generalized Extreme Studentized Deviate Test for Outliers

Many outlier tests exist, this essay is focused on the Generalized Extreme Studentized Deviate Test (ESD- Test). The generalized ESD- test [36] is a generalization of Grubbs-test. [37] The ESD- test can be used to detect one or more outliers in a data set that follows an approximately normal distribution. [36, 30] Manoj and Kannan compared different methods for detecting outliers and found that generalized ESD- test is better than Grubbs' and Dixons' tests. [38] The primary limitation of many tests is that the suspected number of outliers, k, must be specified exactly. If k is not specified correctly, this can distort the conclusions of these tests. On the other hand, the generalized ESD test only requires that an upper bound for the suspected number of outliers be specified. Given the upper bound of outliers r, the generalized ESD- test essentially performs r separate tests: a test for one outlier, a test for two outliers, and so on up to r outliers.

The generalized ESD- test is defined for the hypothesis:

Null hypothesis; H0: There are no outliers in the data set

Alternative Hypothesis; H1: There are up to r outliers in the data set

If $R_i > \lambda_i$, then the null hypothesis is rejected.

The test statistic is computed as follows:

$$R_i = \frac{\mathrm{max_i}|\mathrm{x_i} - \bar{\mathrm{x}}|}{\mathrm{S}}$$

Formular 34

With $\bar{x}$, x_i and S denoting the arithmetic mean the sample and sample standard deviation S, respectively. The suspected extreme observation has to be removed and then the test statistic $R_1, R_2, \dots, R_r$ has to be recomputed with n - 1 observations. This procedure has to be successively repeated until r observations have been removed, respectively tested by the test statistic λ_i. The corresponding critical value for the test statistics can calculate as follows:

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t^2{}_{p,n-i-1})(n-i+1)}} \quad i = 1,2,\dots,r$$

Formular 35

Where $t_{p,v}$ is the 100p percentage point from the t- distribution with v which is the degrees of freedom (n-i-1) (Whereupon i represents the removed outlier and n the total number of values) the and the significance level α

$$p = 1 - \frac{\alpha}{2(n-i+1)}$$

Formular 36

The number of outliers is determined by finding the largest i such that $R_i > \lambda_i$

Simulation studies by Rosner indicate that this critical value approximation is very accurate for n ≥ 25 and only reasonably accurate for n ≥ 15.

> Note: That although the generalized ESD- test is essentially Grubbs test applied sequentially, there are a few important distinctions: The generalized ESD- test makes appropriate adjustments for the critical values based on the number of outliers being tested for that the sequential application of Grubbs test does not. Trying to use Grubbs test sequentially could stop at the wrong iteration and declare no outliers. [35, 36]

To improve the robustness of the generalized ESD -test a modification of the original version was tested. The arithmetic mean $\bar{x}$ was replaced by the median $\tilde{x}$ and the result showed an increased efficacy of the outlier detection observation. The test statistic is computed as follows:

$$R_i = \frac{max_i|x_i - \tilde{x}|}{S}$$

Formular 37

With $\tilde{x}$, x_i and S denoting the median the sample and sample standard deviation, respectively. Simulations have illustrated that the modified version of the test, performance is robust compared to the classical generalized ESD -test. [38]

Challenges of Using Outlier Hypothesis Tests

When performing an outlier test, the analyst needs to choose a procedure based on the number of outliers or specify the number of outliers. Other methods, such as the Tietjen- Moore Test, require the analysing person to specify the number of outliers. There are two problems that can occur when the analyst specifies the incorrect number in a data set. A kind of masking occurs when too few outliers specified. The additional outliers that exist can affect the test so that it detects no outliers. Conversely, a kind of swamping occurs when too many outliers are specified. In this case, the test identifies too many data points as being outliers. Because of these problems, it is necessary to stay categorical critical in the application of outlier tests.

Philosophy about Finding Outliers

The philosophy based on the use of the in- depth knowledge about all the variables when analysing data. Part of this knowledge is knowing what values are typical, unusual, and impossible. When the analyst has an in- depth knowledge, it is often best to use the more straightforward, visual methods. At a glance, data points that are potential outliers will pop out under a knowledgeable gaze of the analyst. Consequently, the use of boxplots, histograms, and old- fashioned data sorting should be the first step. These simple tools provide often enough information to find unusual data points for further investigation. [9]

> *Note: The "Anscombe's quartet" as an example for the superiority of visual methods:*
>
> *Anscombe's quartet comprises four data sets that have nearly identical simple descriptive statistics, yet have very different distributions and appear very different when graphed. Each dataset consists of eleven (x, y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data when analysing it, and the effect of outliers and other influential observations on statistical properties. He described the article as being intended to counter the impression among statisticians that "numerical calculations are exact, but graphs are rough" [41]*

It can be critical to use the Z_{score} and hypothesis tests to find outliers because of their various complications. Using outlier tests can be challenging because they usually, assume the data follow the normal distribution or like the ESD- test the t-distribution. Additionally, the existence of outliers makes the Z_{score} less extreme.

These methods for identifying outliers are sensitive to the presence of outliers. Fortunately, as long as researchers use a simple way to display unusual values, a knowledgeable analyst is likely to know which values need further investigation.

"In my view, the formal statistical tests and calculations are overkill because they can't definitively identify outliers." Jim Frost [9]

Ultimately, analysts must investigate unusual values and use their expertise to determine whether they are legitimate data points. Statistical procedures do not know the subject matter or the data collection process and cannot make the final determination. The analyst should not include or exclude an observation using only the results of a hypothesis test or statistical measure. The analyst should not necessarily remove one or all outliers. [9] Random variation generates occasional extreme values by chance; these are part of the valid data and should generally be included in any calculations. [13] Outliers can be very informative about the subject area and data collection process. It is vital to understand how outliers occur and whether they might happen again as a normal part of the process or study area.

Statistical Analyses on the basis of alternative Methods

What has to be done when the outliner cannot legitimately remove, but they violate the assumptions of the applied statistical analysis? The analyst wants to include them but do not want them to distort the results. There are various statistical analyses applicable for that problem. These statistical analyses are the nonparametric hypothesis tests which are robust to outliers. For these alternatives to the more common parametric tests, outliers will not necessarily violate their assumptions or distort their results.

References

[1] A. Holmes, Barbara Illowsky, Susan Dean: Introductory Business Statistics; 2018 Rice University. OpenStax Rice University 6100 Main Street MS-375 Houston, Texas 77005,

[2] Ausreißer, last accessed on 13.01.2023, https://de.wikipedia.org/wiki/Ausrei%C3%9Fer

[3] Wie reich ist Heilbronn?, last accessed on 13.01.2023, https://phonk-magazin.de/heilbronn-reichste-stadt-deutschlands/

[4] Im Land der Lügen: Wie uns Politik und Wirtschaft mit Zahlen manipulieren | Marktcheck SWR , last accessed on 13.01.2023, https://www.youtube.com/watch?v=PC1Dw1lfLtI

[5] Heilbronn mit bundesweit höchstem Pro-Kopf-Einkommen, last accessed on 13.01.2023, https://www.swr.de/swraktuell/baden-wuerttemberg/heilbronn/heilbronn-einkommensstaerkste-stadt-deutschlands-100.html

[6] Outlier, last accessed on 13.01.2023, https://en.wikipedia.org/wiki/Outlier

[9] Jim Frost, MS: Hypothesis Testing, Statistic by Jim Publishing, 2020

[10] 68–95–99.7 rule, last accessed on 13.01.2023, https://en.wikipedia.org/wiki/68%E2%80%9395%E2%80%9399.7_rule

[11] Studentization, last accessed on 13.01.2023, https://en.wikipedia.org/wiki/Studentization

[12] Shiffler, R. E. Maximum Z Score und Outliners, The American Statistician, February 1988, Vol. 42, No1, 79-80

[13] Stephen L. R. Ellison, Vicki J. Barwick, Trevor J. Duguid Farrant: Practical Statistics for the Analytical Scientist: A Bench Guide (Valid Analytical Measurement) – 15. November 2009, Royal Society of Chemistry; 2nd ed. 2010 Edition

[14] Q–Q plot, last accessed on 13.01.2023, https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot

[15] Normal probability plot, last accessed on 13.01.2023, https://en.wikipedia.org/wiki/Normal_probability_plot#Other_distributions

[16] Normal Probability Plot, last accessed on 13.01.2023, https://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm

[17] Probit, last accessed on 13.01.2023, https://en.wikipedia.org/wiki/Probit

[18] Testing normality in cinemetrics, last accessed on 13.01.2023, https://nickredfern.wordpress.com/2009/02/19/testing-normality-in-cinemetrics/

[19] Stephen W. Looney and Thomas R. Gulledge, Jr.: Use of the Correlation Coefficient with Normal Probability Plots; The American Statistician; Vol. 39, No. 1 (Feb., 1985)

[20] Ludwig Fahrmeir, Christian Heumann, Rita Künstler, Iris Pigeot, Gerhard Tutz: Statistik; 8. Auflage; Springer- Verlag 2016

[21] Boris Iglewicz, David C. Hoaglin: Volume 16: How to Detect and Handle Outliers, 1993 by ASQC

[22] A. Holmes, Barbara Illowsky, Susan Dean: Introductory Business Statistics; 2018 Rice University.

[23] Karanjit Singh, Shuchita Upadhyaya: Outlier Detection: Applications and Techniques, January 2012, International Journal of Computer Science Issues 9(3)

[24] Outliers explained: a quick guide to the different types of outliers, last accessed on 13.01.2023, https://towardsdatascience.com/outliers-analysis-a-quick-guide-to-the-different-types-of-outliers-e41de37e6bf6

[25] Upper and Lower Fences, last accessed on 13.01.2023, https://www.statisticshowto.com/upper-and-lower-fences/

[26] Box Plot, last accessed on 13.01.2023, https://www.itl.nist.gov/div898/handbook/eda/section3/eda337.htm

[27] A. Holmes, Barbara Illowsky, Susan Dean: Statistics, OpenStax Rice University, 6100 Main Street MS-375, Houston, Texas 77005, 2020 Texas Education Agency (TEA)

[28] Box Plot Explained with Examples, last accessed on 13.01.2023, https://statisticsbyjim.com/graphs/box-plot/#more-19849

[29] Fabrice I. Mowbray, Susan M. Fox-Wasylyshyn and Maher M. El-Masri: Univariate Outliers: A Conceptual Overview for the Nurse Researcher, Canadian Journal of Nursing Research, 2019, Vol. 51(1) 31–37

[30] Steven Walfish: A Review of Statistical Outlier Methods, Pharmaceutical Technology Nov 2, 2006

[31] William G. Jacoby: Statistical Graphics for Univariate and Bivariate Data (Quantitative Applications in the Social Sciences) SAGE Publications, Inc; Illustrated Edition, 15. April 1997

[32] Tabachnick, B. G., & Fidell, L. S. (2013). Using multivariate statistics (6th ed.). Boston, MA: Pearson.

[33] Polit, D. F. (2010). Statistics and data analysis for nursing research (2nd ed.). Boston, MA: Pearson.

[34] Field, A. P., & Miles, J. (2010). Discovering statistics using SAS: (and sex and drugs and rock n roll). Thousand Oaks, CA: SAGE.

[35] Sehar Saleem, Maria Aslam and Mah Rukh Shaukat: A REVIEW AND EMPIRICAL COMPARISON OF UNIVARIATE OUTLIER DETECTION METHODS, Pak. J. Statist., 2021 Vol. 37(4), 447-462

[36] Generalized ESD Test for Outliers, last accessed on 13.01.2023, https://www.itl.nist.gov/div898/handbook/eda/section3/eda35h3.htm

[37] Generalized Extreme Studentized Deviate Test, last accessed on 13.01.2023, https://real-statistics.com/students-t-distribution/identifying-outliers-using-t-distribution/generalized-extreme-studentized-deviate-test/

[38] Mufda Jameel Alrawashdeh: An adjusted Grubbs' and generalized extreme studentized deviation, Demonstratio Mathematica 2021; 54: 548–557

[39] Filliben, J. J. The Probability Plot Correlation Coefficient Test for Normality, Technometrics, pp. 111-117. (February 1975)

[40] Amtl. Sammlung § 64 LFGB: Planung und statistische Auswertung von Ringversuchen zur Methodenvalidierung; September 2006

[41] F. J. Anscombe: Graphs in Statistical Analysis, The American Statistician, Vol. 27, No. 1 (Feb., 1973), pp. 17-21

[42] Thomas, Kickhahn: Statistik für Naturwissenschaftler für Dummies; 2017; Wiley-VCH Verlag GmbH

[43] L. Fahrmeir, u.a.: Statistik. Der Weg zur Datenanalyse, 8. Auflage; Springer- Verlag, Mai 2016

[44] Online Statistics Education: A Multimedia Course of Study (http://onlinestatbook.com/). Project Leader: David M. Lane, Rice University.

[45] Terence Tao: Universelle Gesetze; Spektrum der Wissenschaft; Januar; 2014

[46] Gesetz der großen Zahlen, last accessed on 13.01.2023, https://de.wikipedia.org/wiki/Gesetz_der_gro%C3%9Fen_Zahlen

[47] Irasianty Frost: Statistische Testverfahren, Signifikanz und p-Werte; Springer Fachmedien Wiesbaden GmbH 2017

[48] Cox, D. R. (2006), Principles of Statistical Inference, Cambridge University Press.

[49] Hartmut Schiefer, Felix Schiefer: Statistik für Ingenieure, Springer Fachmedien Wiesbaden GmbH 2018

[50] Heidrun Matthäus; Wolf- Gert Matthäus: Statistik und Excel; Springer Fachmedien Wiesbaden; 2016

[51] Herbert Stocker: Methoden der empirischen Wirtschaftsforschung; Universität Innsbruck Sommersemester 2020