Consequences of inappropriate detection and removal of outliers in statistics

Georgi Petkov
Institute of Population and Human Studies,
Bulgarian Academy of Science, Sofia, Bulgaria

Abstract: *In statistics, the presence of outliers in the data set could wrongly distort the estimation of the mean. In addition, the extreme values increase the variability and consequently, the power of the statistical methods decreases. However, there are disagreements in the literature both about what the nature of outliers is, and about how to deal with them when doing further statistical analyses. A lot of conventional procedures for both detecting and dealing with outliers are discussed. The effect of increase the probability of error of the first order type is demonstrated with two simple simulations. The general conclusion is that an information outside of the data set is necessary for a correct decision. This information could come only from the human expertise of the researchers of the specific domain of interest. The importance of the topic for outliers is discussed, the need of deeper analyses, accompanied with many simulation studies, is argued.*

Keywords*: OUTLIERS, STATISTICS, DATA SCIENCE, SIGNIFICANCE LEVEL, STATISTICAL POWER*

1. Introduction

In statistics, outliers are typically defined as data points that are far away from the other observations (for ex. [1]). Often they could produce large discrepancies between statistical estimations of some parameters and their true values. There is a huge number of scientific work analyzing the nature, the reasons for existence and the possible tools for dealing with outliers (for ex. [2], [3]). However, there are still controversies about the adequacy of some of the technics for removing data, especially those widespread in the social sciences.

The first problem is that there are not stable mathematical definitions about what is outlier. One should rely on expert knowledge in the specific domain in order to define which values are impossible. In addition, some outliers could reflect imprecisions of the measuring process. Thus, both the deep domain knowledge and the expertise in the methodology for measurement are crucial for correct detection and deal with outliers.

Some of the mistakes in data analyses reflect also a controversy about the meaning of the concept of outlier. One interpretation is that outlier are impossible values. For example, if one measures the height of a sample of people and there is a value of 18.2 meters somewhere in the data, this is obviously a product of a mistake. 18.2 meters is an impossible value for a height of a person. Maybe it is a result of a typo error, or something else, but it is certainly an error. However, a value of 2.40 meters is not necessarily an outlier. According to a different understanding of the concept, however, every value that differs significantly from the other data points is defined as an outlier. Thus, even the value of 2.40 meters, which could be possibly the height of a very tall person, should be mathematically processed in the same way as the impossible values.

It is unproductive to argue for the word meaning, however it is essential to have consistency between the understanding the nature of the outliers and consequent decisions what to do with them. If one think the outliers as impossible values, she should, of course, remove them from the data set. In this case, the difficulty is in the process of detecting them. One should prove strictly their impossibility on the basis of a deep domain knowledge. Otherwise, deleting valid empirical data would produce wrong conclusions. From the other side, if one think outliers as just very extreme values, it is easier to detect them mathematically but it is a question how to deal with them.

2. How big is the impact of outliers on statistical estimations and conclusions

There are two main types of problems that may arise from the impossible and extreme values in the data set. First, the statistical estimation of the parameters could become too imprecise. Second, the statistical procedures for making conclusions could lose power because of the increase of the variance of the observed variables.

The estimation of the mean of a set of values is sensitive to the extreme values. For example, the mean of the numbers 1, 2, 3, 4, 5, 6, 7, 8, and 9 is 5 but if ther is a value of 100 in the data set, then the mean (of the numbers 1, 2, 3, 4, 5, 6, 7, 8, 9, 100) becomes 14.5, almost three times larger. Obviously, if the value of 100 has come across by a typo error for example, our statistical conclusions would be invalid if we do not remove it before doing analyses.

However, the second problem – increase of the variance of the observed variables and the consequent decrease of the statistical power of the used methods, could produce logical fallacy. One may argue that increasing of the statistical power is always an honorable goal, thus it is always right to use tools to do it. Therefore, it would be better if we remove the extreme values from our data set. This is a wrong conclusion. It is definitely better to use more powerful statistical methods; there is no doubt on this. Otherwise, even following strictly the standards for statistical significance (for ex. $p<0.05$ in the behavioral and other studies), the number of the wrong conclusions could produce very large problems, like for example the replication crisis in social psychology (see for ex. [4]). Increasing the statistical power, however, does not mean simply removing some data points from our sample. It means improving the procedures for collecting data; increasing the sample size, etc. It should be done *before* the actual data collection. Otherwise, jest decreasing artificially the variance of the observed variable would definitely make worse both our estimations and the adequacy of the used statistical methods. It will produce larger number of type one errors (incorrectly rejection of the null hypothesis). This prediction could be checked with a simple simulation.

Simulation 1

In order to test how decreasing the variance impact the first order error type, a simulation on Python was performed. Two sets form exponents of a same normal distribution N(0, 1) were randomly generated. Then a t-test between the respective two samples of numbers was performed and it was detected whether the significance value $p<0.05$ or not. If $p<0.05$ then a wrong rejection of the null hypothesis was recorded. The simulation was performed 10000 times thus stabilizing the first order error rate. The procedure was repeated for various sample sizes (10, 30, and 100) and for various criteria for removing data (removing all data above 2, 2.5, and 3 standard deviations respectively).

Results:

The results from the simulation are summarized in Table 1:

It is seen from the results that deleting some data points above a certain threshold for the standard deviation, undesirable increases the probability for a wrong rejection of the null hypothesis (type I error rate). It increases more when more data are altered, in particular if use lower thresholds or when the sample sizes are larger.

Summarizing, the increase of the statistical power by just removing extreme values is a very wrong procedure. It is definitely better to use large sample sizes and precise measuring instruments

in order to improve the statistical power, but this should be done before collecting the data. Otherwise, this is not a valid procedure. It causes artificial increase of the rate of the more dangerous error in statistics – possible wrong rejection of the null hypothesis.

Table 1: *Percentage of the incorrectly rejection of null hypothesis (first order error rate) when varying sample size and the threshold for removing data*

	Sample size 10	Sample size 30	Sample size 100
Removing data above 2 st. devs.	0.1172	0.1214	0.1350
Removing data above 2.5 st. devs.	0.1064	0.1060	0.1189
Removing data above 3 st. devs.	0.0487	0.0965	0.1113

3. Methods for detecting outliers

After demonstrating the inappropriateness of just removing data outside of a fixed number of standard deviations, let us turn back to the various understandings of the concept of outlier. According to [5], who follows the definition of an outlier as any data point, which is far from the other data, there are three types of outliers: (1) typos; (2) values for elements that are not part of the population of interest; and (3) natural extreme values.

An example of a typo was already given – if one records the height of people and there is a value of 18.2 meters, there is no other reasonable explanation then it is wrongly typed.

It is often more difficult to detect a value that is incorrectly recorded because the respective element is not from the studied population. If, for example, one studies the effect of a certain educational method on the student's knowledge, it is possible that a certain student with disabilities or with language difficulties had been measured in one of the groups. Actually, the research question had been probably about normal developed students only and in such a case, the respective result could be logical treated as coming from a different population, which is not under the research interest. However, it is often difficult to precise such argumentation. Anyway, the decision whether a certain outlier is from this type, should be taken very carefully, in accordance with the domain knowledge and without pure mathematical considerations.

Finally, some data points should be extreme by a definition. Is a certain variable is distributed normally for example, about 2/3 of the data should be outside the interval mean$\pm$1.st.dev; about 5% - outside of the interval mean$\pm$2.st.dev, etc. They should not be removed, otherwise, the estimation of the true variance would be incorrect and as a consequence, the probability for the first type error will increase (as demonstrated in Simulation 1).

Following [5] there are several methods for detecting extreme values that could be thought as outliers:

(1) Just sort data and look on the both ends.

(2) Make a graph and look whether some outliers pop-out.

(3) Calculate the z-values, sort them and observe the numbers outside (-3, 3).

(4) Calculate the quartiles Q1 (25%), Q2 (the median) and Q3 (75%) and then make fences around Q1 and Q2 with length 1.5*median and 3*median. All data outside the inner fences could be thought as extreme values.

(5) Use hypothesis testing method, for example the Grubb's test [6].

The list above is sorted by complexity of the method. The problem is that often as simple a method is, as reliable it is. For example, The Grubb's test requires initially to mark how many outliers are expected. However, one can not know this, actually the goal is to find them. Thus, the method does not work correctly with an unknown number of outliers. At the same time, the list is sorted inversely by automaticity of the decision. As simpler methods for detecting the extreme values is used, as larger is the necessity of a top-down domain knowledge in order to make a decision what to do with the respective extreme value.

4. Methods for removing outliers

As was mentioned, depending on how one views on the outliers – incorrectly presented data or just extreme values, the higher difficulty is either on defining a certain value as an outlier, or on deciding what do to with it. If the outlier is a wrong value, one should prove that it is there either by a typo or that the respective element is outside of the studied population. After the eventual proof of this, the decision is simple – just remove it from the data set. From the other side, if one think any extreme value as an outlier, then its detecting is relatively simple but the decision how to deal with it could be very difficult.

If we follow the second view, then may look what Frost [7] proposes to do with the already detected outliers. First, it could be very usable to analyze why it is an outlier at all. This could give us a valuable information about the process of collecting the data and as well about the sampling process. After that, it is time for a decision.

(1) If this is obviously and undoubtable a typos, then remove it.

(2) If this is obviously and undoubtable a value from an element, which is not a member of the studied population, then remove it.

(3) If this is a natural extreme value that should be there, then leave it and do not do anything more about it. The percentage of the values above the thresholds of 1, 2, 3, and more standard deviations could be easily estimated and could be compared with the respective theoretically expected percentages. If one cannot decide whether a certain value is naturally extreme, better leave it. Removing should be done only after finding undoubtable reason for it.

(4) Use alternative statistical methods for analyzing data. There are three main types for this: use non-parametric tests, use transformation of the data, and bootstrapping.

Non-parametric tests

Whereas statistical mean is very sensitive to extreme values, the median is not. Then, if there are extreme outliers, which we cannot remove because there is no obvious reason for it, it seems reasonable to compare the medians of the respective studied groups, instead of the means. However, the problem is that usually there are not well defined theoretical sample distributions for the medians. Anyway, almost all of the parametric tests have developed their non-parametric versions. For full range, see Table 2 (adapted and supplemented form [7]):

Table 2: *Pairs of parametric and non-parametric tests for the same research design*

	Parametric test using means	Non-parametric test based on medians
Estimation of one mean	One-sample t-test	1-sample Sign, 1-sample Wilcoxon
Comparing two groups	Two-sample t-test	Mann-Whitney test
Comparing more than two groups.	One-way ANOVA	Kruskal-Wallis, Mood's median test
Looking for a relationship	Pearson's correlation	Spearman's correlation

As it is seen, many of the analyses could be performed both by using parametric and non-parametric tests and often different results would be obtained. Isn't this a problem?

Sometimes the mean is not a good estimation of the central tendency but the decision of this is not a mathematical task. Consider for example, the following small sample of monthly incomes: 9 times 1 000 and a single 1 000 000 value. Nine people receive 1000 each and one person receives 1 000 000. The last single value is an outlier, following the "any extreme value" definition. However, it is a possible (and probably true) value and we cannot remove it from the data set. What is the better estimation of the central tendency – the mean of almost 100 000 or the median of 1000? The answer of this question is not obvious! The answer is that it depends! It depends on the purpose for estimation the central tendency. From the point of view of the social minister for example, whose goal is to deal with misbalances of the income, the mean is much misbalanced because of the outlier and he probably would prefer to deal with the median. From the point of view of the financial minister, however, who wants to predict the budged, the mean is a perfect estimation of the central tendency.

Thus, the decision of whether to use parametric or non-parametric test should be based mainly on the decision whether the mean or the median is the central tendency measurement of interest. But this is a task from the respective domain, it is not a mathematical one. Thus, one should not choose to use non-parametric test just because there are outliers in the data set.

Even if the mean-median choose decision is not restricted from the specific domain, one should consider some additional mathematical restrictions: For small sample sizes, usually the parametric tests require normal distributions and equality of variance, whereas the non-parametric tests are still sensitive to the equality of variance but may deal with non-normal distributions. However, for large sample sizes these requirements are not valid for the parametric tests. The Central Limit Theorem cares about the normality of the distributions, whereas the tests become less sensitive to eventual inequality of the variance. As a conclusion, it seems better to use non-parametric tests instead of parametric ones only in the case of small sample sizes, non-normal distributions, and equal variances. At the same time, it should be always preferable to plan experimental designs with larger sample sizes, thus achieving enough power for conclusions. It seems that a paradox arises if we chose non-parametric tests solely on the outliers consideration. The non-parametric tests are better only when the sample sizes are small and the distributions are non-normal. But in these cases the uncertainty of whether a certain value is outlier or not, also increases.

Summarizing, the chose between parametric and non-parametric tests should be based on the domain specific knowledge about which measure – mean or median – is a better representatives of the central tendency. There is no arguable reason to make this decision on the basis of existence or absence of extreme outliers.

Transformation of data

One often use technique for data manipulation is to initially perform some algebraic operations on the variables, most often non-linear ones, for example to take logarithm or exponent. There are domain-specific and mathematical consideration for this. For example, it is well known in psychology that often the psychological sensation of a certain stimulus is a logarithmic (or exponential) function of the physical intensity of the respective stimulus [8]. In such a case, it is reasonable to perform the respective mathematical operation and to deal with the transformed variable. From the other side, often the arguments for a certain operation are purely statistical. For example, it is known that reaction time is distributed non-normally with a long tail on right [9]. That is why some authors [9] recommend to normalize such distributions by simply logarithm. As a side effect, the influence on the mean of any very big outliers will decrease.

It is questionable, however, whether this non-domain but purely mathematically based transformation would not produce any spurious statistical results. One doubt of the adequacy on logarithmic reaction time pre-processing method is expressed in [10]. This question is additionally tested empirically with the simulation that follows in the same context. It is checked whether the probability for wrongly rejection of the null hypothesis will be affected by logarithmic transformation of the variables. The sample size would vary.

Simulation 2

For the purpose of the simulation, performed on Python, two sets of random variables were generated from a normal distribution N(0, 1). Both they were taken with their absolute value in order to avoid logarithms of negative numbers. On the next step, two additional variables were created being natural logarithms of the first two. Like in the previous simulation, a t-test was performed and any incorrect rejections of the null hypothesis were recorded. The simulation was repeated 10000 times and total percent of the wrongly rejected null hypotheses was reported. The simulation was repeated by varying the sample size – 10, 30, and 100.

Results

For sample size 10, the error type I rate was 0.63. For sample size 30 – 0.550; for sample size 100 – 0.505. Thus, we may conclude that the simple logarithmic transformation potentially increases the type I error rate but this happen with small sample sizes only.

Anyway, further investigation of the eventual mistakes that could happen with arbitrary algebraic operations on data are needed.

Bootstrapping

One relatively rarely used in psychology and social sciences, but seeming promising technique, is bootstrapping. Briefly speaking, by using bootstrapping, we replace the theoretical sample distribution with an empirically created one by using our data. For a very simplified example, assume our data set consists of four data points: 1, 2, 3, and 5. We may recombine these points with replacement. Thus, some possible variations of the data set are: (1) 1, 2, 3, 5; (2) 1, 3, 3, 3; (3) 2, 3, 3, 5, etc. Of course, the method is used mainly on relatively large data sets. Usually huge number of permutations (tens of thousands) with replacement are used. Every single permutation could be viewed as a possible data set and the distribution of the parameter estimations among all these permutations form our sample distribution.

The three conditions for bootstrapping are [7]:

1. The bootstrap method has an equal probability of randomly drawing each original data point for inclusion in the resampled datasets.

2. The procedure can select a data point more than once for a resampled dataset. This property is the "with replacement" aspect of the process.

3. The procedure creates resampled datasets that are the same size as the original dataset.

There are a lot of advantages by using bootstrapping method. There are no any restrictions and requirements for normal or another distributions, for equality of variance, etc. There are not any in advance expectations about the sample distribution. The method works perfectly for confidence intervals estimations.

However, there are still some assumptions to be met in order to obtain valid results from bootstrapping. The most important of them is that our data represent well the population. Otherwise, if there are some biases, they will reflect the results after bootstrapping as well. This arises a logical question about the outliers. If a certain outlier is due to typos or another type of mistake, then bootstrapping is a

good method to eliminate its impact. However, if it is a principal problem in the way of collecting data (for example, if it is a bias to do typos in non-random direction), or in the measurements methods, or anyway in the design of the investigation, then the bootstrapping method will probably not improve the results.

5. Conclusions

In statistics, the presence of outliers in the data set could wrongly distort the estimation of the mean or other parameters. In addition, the extreme values increase the estimation of the variance of certain variables and consequently, the power of the statistical methods used decreases. However, there are disagreements in the literature both about what the nature of outliers is, and about how to deal with them when doing further statistical analyses.

There are at least three main sources of outliers. They may be due to typos and if one can prove this by arguing that such values are impossible, then without doubt they should be removed from further analyses.

The outliers may come from elements of the sample that doesn't come from the studied population. If such is the case, again, these values could be removed together with all other values from the same source.

Sometimes, however, there are extremely large or extremely small values in our data set that could not be ignored without reason. It is a bad research practice in some groups for behavioral studies just to cut all data above 2, 2.5, or 3 standard deviations from mean. This procedure is justificated by the seeking for higher power of the statistical methods. It is true that low-powered studies are one of the cornerstones for the replication crisis in social psychology [4] for example, and without a doubt, they should be avoided. The increase of the power, however, should be performed with methods outside of the particular data. This means, the procedure should be improved as much as possible; the measuring instruments should be tuned; the sample sizes should be increased; etc. All of these good research practices however should be done before collection of data. After it, removing data points without clear reason outside form the data themselves, has the same effect as skew of the sample. Removing data is in fact a skew of the data that happens after, not before collecting the data.

The effect of increase the probability of first order type is demonstrated with two simple simulations. In the first one, the simple removing data from two samples from the same distribution – exponent of a standard normal one, causes significant increase of the first order error rate. As much data are distorted, as higher the increase. In addition, as higher the sample size, as higher the effect. In the second simulation, it is demonstrated how another widespread procedure for dealing with outliers – namely, taking logarithm of the variable, causes also a small increase of the probability for type I error.

A lot of conventional procedures for both detecting and dealing with outliers are discussed. As a general rule of thumb, as objective is a certain procedure, as less reliable it is. This is a widespread phenomena. The correct detecting and dealing with outliers requires domain knowledge outside form the data. Thus, it is better simply to sort the values and to look them visually on a graph, than to use hardcore procedures for detecting outliers. This is because the data themselves doesn't content information about the specific research domain, nor about the procedure for obtaining them. Such information, however, is necessary for a correct conclusion.

Three of the techniques require special attention – using non-parametric tests, transforming the raw data, and bootstrapping. Unfortunately, the general conclusion is that again just using complex statistical methods cannot displace the expert knowledge in the domain. The non-parametric tests seem mathematically better than their parametric partners only when the sample sizes are small, the distributions are non-normal, and the variances among the groups are almost equal. However, exactly in these cases all methods have relatively small power nevertheless of the presence or absence of outliers. In addition, the main argument for choosing parametric or non-parametric test should be again top-down knowledge and good understanding of the research question, Namely, it is a matter of interest whether the mean or the median better represents the central tendency. This question is domain based and is orthogonal from the pure mathematical considerations.

The same arguments could be said for any algebraic transformation of the data like logarithm, exponent, etc. The choose should depend on the domain specific knowledge, not of the requirements of a specific statistical test. Obviously, if certain requirements of a certain test are not met, one can not arbitrarily distort data just to fulfill them. The methods use should be adapted to the data, not vice versa.

Bootstrapping seems promising technique for estimations of parameters, confidence intervals, etc. It doesn't rely on prior theoretical assumptions for the sample distribution. Thus, using bootstrapping seems very good method for decreasing the influence of a single extreme outlier. However, any biases embedded in the research design or measuring process, would reflect the results from bootstrapping too. Thus, again it seems that even bootstrapping is good when we know about outliers (but in this case we can just remove them), and is not perfect when we do not know them. But knowledge should always come from the domain and outside of the data set.

Finally, probably the worst possible bad scientific practice is to use different methods for dealing with outliers and to report only the "successful" ones. Obviously, in such a case, a huge number of wrong rejections of the null hypothesis would be reported.

The topic for outliers seems promising for further deeper analyses, accompanied with a lot of simulation studies. Improving methodology, especially psychology and other social science, is of a huge importance for establishment of uncontroversial knowledge.

References

1. Outliers (7 April 2022) In *Wikipedia*, https://en.wikipedia.org/wiki/Outlier
2. Ilyas, I. F., & Chu, X. (2019). *Data cleaning*. Morgan & Claypool.
3. Kazil, J., & Jarmul, K. (2016). *Data wrangling with python: tips and tools to make your life easier*. O'Reilly Media, Inc.
4. Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, 39(4), 202.
5. Frost, J. (2022) Five Ways to Find Outliers in Your Data. In *Statistics by Jim*. https://statisticsbyjim.com/basics/outliers/
6. Grubbs, F. E. (February 1969). Procedures for detecting outlying observations in samples. *Technometrics*. **11 (1)**: 1–21.
7. Frost, J. (2022) Guidelines for Removing and Handling Outliers in Data. In *Statistics by Jim*. https://statisticsbyjim.com/basics/remove-outliers/
8. Weber–Fechner law (29 March 2022) In *Wikipedia*, https://en.wikipedia.org/wiki/Weber%E2%80%93Fechner_law
9. Lachaud, C. M., & Renaud, O. (2011). A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model. *Applied Psycholinguistics*, 32(2), 389-416.
10. Changyong, F. E. N. G., Hongyue, W. A. N. G., Naiji, L. U., Tian, C. H. E. N., Hua, H. E., & Ying, L. U. (2014). Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, 26(2), 105.