**Outliers**

**5 Ways to Find Outliers in Your Data**

Finding outliers depends on subject-area knowledge and an understanding of the data collection process. While there is no solid mathematical definition, there are guidelines and statistical tests you can use to find outlier candidates.

Outliers are a simple concept—they are values that are notably different from other data points, and they can cause problems in statistical procedures.

**Method 1: Sorting Your Datasheet to Find Outliers**

Sorting your datasheet is a simple but effective way to highlight unusual values. Simply sort your data sheet for each variable and then look for unusually high or low values.

**Method 2: Graphing Your Data to Identify Outliers**

Boxplots, histograms, and scatterplots can highlight outliers.

univariate outliers: Boxplot (interquartile method with fence), histograms

Bivariate/multivariate outliers: scatterplots

The scatterplot with regression line shows how most of the points follow the fitted line for the model. However, the circled point does not fit the model well.

This type of outlier can be a problem in regression analysis. Given the multifaceted nature of multivariate regression, there are numerous types of outliers in that realm.

**Method 3: Using Z-scores to Detect Outliers**

Z-scores can quantify the unusualness of an observation when your data follow the normal distribution. Z-scores are the number of standard deviations above and below the mean that each value falls.

The further away an observation’s Z-score is from zero, the more unusual it is. A standard cut-off value for finding outliers are Z-scores of **+/-3** or further from zero.

However, if your data don’t follow the normal distribution, this approach might not be accurate.

Note that Z-scores can be misleading with small datasets because the maximum Z-score is limited to (n−1) / √ n.\*

Also, note that the outlier’s presence throws off the Z-scores because it inflates the mean and standard deviation as we saw earlier. Notice how all the Z-scores are negative except the outlier’s value. If we calculated Z-scores without the outlier, they’d be different! Be aware that if your dataset contains outliers, Z-values are biased such that they appear to be less extreme (i.e., closer to zero).

**Method 4: Using the Interquartile Range to Create Outlier Fences**

You can use the interquartile range (IQR), several quartile values, and an adjustment factor to calculate boundaries for what constitutes minor and major outliers. Minor and major denote the unusualness of the outlier relative to the overall distribution of values. Major outliers are more extreme. Analysts also refer to these categorizations as mild and extreme outliers.

The IQR is the middle 50% of the dataset. It’s the range of values between the third quartile and the first quartile (Q3 – Q1). We can take the IQR, Q1, and Q3 values to calculate the following outlier fences for our dataset: lower outer, lower inner, upper inner, and upper outer. These fences determine whether data points are outliers and whether they are mild or extreme.

Values that fall inside the two inner fences are not outliers.

To calculate the outlier fences, do the following:

1. Take your IQR and multiply it by 1.5 and 3. We’ll use these values to obtain the inner and outer fences. For our example, the IQR equals 0.222. Consequently, 0.222 \* 1.5 = 0.333 and 0.222 \* 3 = 0.666. We’ll use 0.333 and 0.666 in the following steps.
2. Calculate the inner and outer lower fences. Take the Q1 value and subtract the two values from step 1. The two results are the lower inner and outer outlier fences. For our example, Q1 is 1.714. So, the lower inner fence = 1.714 – 0.333 = 1.381 and the lower outer fence = 1.714 – 0.666 = 1.048.
3. Calculate the inner and outer upper fences. Take the Q3 value and add the two values from step 1. The two results are the upper inner and upper outlier fences. For our example, Q3 is 1.936. So, the upper inner fence = 1.936 + 0.333 = 2.269 and the upper outer fence = 1.936 + 0.666 = 2.602.

The IQR method is helpful because it uses percentiles, which do not depend on a specific distribution. Additionally, percentiles are relatively robust to the presence of outliers compared to the other quantitative methods.

**Method 5: Finding Outliers with Hypothesis Tests**

You can use hypothesis tests to find outliers. Many outlier tests exist, but I’ll focus on one to illustrate how they work. In this post, I demonstrate **Grubbs’ test**, which tests the following hypotheses:

* Null: All values in the sample were drawn from a single population that follows the same normal distribution.
* Alternative: One value in the sample was not drawn from the same normally distributed population as the other values.

If the p-value for this test is less than your significance level, you can reject the null and conclude that one of the values is an outlier. The analysis identifies the value in question.

If you use Grubbs’ test and find an outlier, don’t remove that outlier and perform the analysis again. That process can cause you to remove values that are not outliers.

**Challenges of Using Outlier Hypothesis Tests: Masking and Swamping**

When performing an outlier test, you either need to choose a procedure based on the number of outliers or specify the number of outliers for a test. Grubbs’ test checks for only one outlier. However, other procedures, such as the **Tietjen-Moore Test**, require you to specify the number of outliers. That’s hard to do correctly! After all, you’re performing the test to find outliers! Masking and swamping are two problems that can occur when you specify the incorrect number of outliers in a dataset.

* Masking occurs when you specify too few outliers. The additional outliers that exist can affect the test so that it detects no outliers. For example, if you specify one outlier when there are two, the test can miss both outliers.
* Conversely, swamping occurs when you specify too many outliers. In this case, the test identifies too many data points as being outliers. For example, if you specify two outliers when there is only one, the test might determine that there are two outliers.

**Philosophy about Finding Outliers**

As you saw, there are many ways to identify outliers. The philosophy is that you must use your in-depth knowledge about all the variables when analyzing data. Part of this knowledge is knowing what values are typical, unusual, and impossible.

I find that when you have this in-depth knowledge, it’s best to use the more straightforward, visual methods. At a glance, data points that are potential outliers will pop out under your knowledgeable gaze. Consequently, I’ll often use boxplots, histograms, and good old-fashioned data sorting! These simple tools provide enough information for me to find unusual data points for further investigation.

Typically, I don’t use Z-scores and hypothesis tests to find outliers because of their various complications. Using outlier tests can be challenging because they usually assume your data follow the normal distribution, and then there’s masking and swamping. Additionally, the existence of outliers makes Z-scores less extreme. It’s ironic, but these methods for identifying outliers are actually sensitive to the presence of outliers! Fortunately, as long as researchers use a simple method to display unusual values, a knowledgeable analyst is likely to know which values need further investigation.

In my view, the more formal statistical tests and calculations are overkill because they can’t definitively identify outliers. Ultimately, analysts must investigate unusual values and use their expertise to determine whether they are legitimate data points. Statistical procedures don’t know the subject matter or the data collection process and can’t make the final determination. You should not include or exclude an observation based entirely on the results of a hypothesis test or statistical measure.

At this stage of the analysis, we’re only identifying potential outliers for further investigation. It’s just the first step in handling them. If we err, we want to err on the side of investigating too many values rather than too few.

**Not all outliers are bad and some should not be deleted.** In fact, outliers can be very informative about the subject-area and data collection process. It’s important to understand how outliers occur and whether they might happen again as a normal part of the process or study area.

**Guidelines for Removing and Handling Outliers in Data**

Outliers can be very informative about the subject-area and data collection process.

It’s essential to understand how outliers occur and whether they might happen again as a normal part of the process or study area. Unfortunately, resisting the temptation to remove outliers inappropriately can be difficult. Outliers increase the variability in your data, which decreases statistical power. Consequently, excluding outliers can cause your results to become statistically significant.

Outlier identification is just the first step. Deciding how to handle outliers depends on investigating their underlying cause.

In broad strokes, there are three causes for outliers:

* data entry or measurement errors,
* sampling problems and unusual conditions, and
* natural variation

**Data Entry and Measurement Errors and Outliers**

Errors can occur during measurement and data entry. During data entry, typos can produce weird values.

These types of errors are easy cases to understand. If you determine that an outlier value is an error, correct the value when possible. That can involve fixing the typo or possibly remeasuring the item or person. If that’s not possible, you must delete the data point because you know it’s an incorrect value.

**Sampling Problems Can Cause Outliers**

Inferential statistics use samples to draw conclusions about a specific population. Studies should carefully define a population, and then draw a random sample from it specifically. That’s the process by which a study can learn about a population.

Unfortunately, your study might accidentally obtain an item or person that is not from the target population. There are several ways this can occur. For example, unusual events or characteristics can occur that deviate from the defined population. Perhaps the experimenter measures the item or subject under abnormal conditions. In other cases, you can accidentally collect an item that falls outside your target population, and, thus, it might have unusual characteristics.

If you can establish that an item or person does not represent your target population, you can remove that data point. However, you must be able to attribute a specific cause or reason for why that sample item does not fit your target population.

**Natural Variation Can Produce Outliers**

The previous causes of outliers are bad things. They represent different types of problems that you need to correct. However, natural variation can also produce outliers—and it’s not necessarily a problem.

All data distributions have a spread of values. Extreme values can occur, but they have lower probabilities.

If your sample size is large enough, you’re bound to obtain unusual values. In a normal distribution, approximately 1 in 340 observations will be at least three standard deviations away from the mean. However, random chance might include extreme values in smaller datasets! In other words, the process or population you’re studying might produce weird values naturally. There’s nothing wrong with these data points. They’re unusual, but they are a normal part of the data distribution.

If the extreme value is a legitimate observation that is a natural part of the population you’re studying, you should leave it in the dataset. I’ll explain how to analyze datasets that contain outliers you can’t exclude shortly!

**Guidelines for Dealing with Outliers**

Sometimes it’s best to keep outliers in your data. They can capture valuable information that is part of your study area. Retaining these points can be hard, particularly when it reduces statistical significance! However, excluding extreme values solely due to their extremeness can distort the results by removing information about the variability inherent in the study area. You’re forcing the subject area to appear less variable than it is in reality.

When considering whether to remove an outlier, you’ll need to evaluate if it appropriately reflects your target population, subject-area, research question, and research methodology.

If the outlier in question is:

* A measurement error or data entry error, correct the error if possible. If you can’t fix it, remove that observation because you know it’s incorrect.
* Not a part of the population you are studying (i.e., unusual properties or conditions), you can legitimately remove the outlier.
* A natural part of the population you are studying, you should not remove it.

**Reminder:**

* When you decide to remove outliers, document the excluded data points and explain your reasoning. You must be able to attribute a specific cause for removing outliers.
* Another approach is to perform the analysis with and without these observations and discuss the differences. Comparing results in this manner is particularly useful when you’re unsure about removing an outlier and when there is substantial disagreement within a group over this question.

**Statistical Analyses that Can Handle Outliers**

What do you do when you can’t legitimately remove outliers, but they violate the assumptions of your statistical analysis? You want to include them but don’t want them to distort the results. Fortunately, there are various statistical analyses up to the task. Here are several options you can try.

* Nonparametric hypothesis tests are robust to outliers. For these alternatives to the more common parametric tests, outliers won’t necessarily violate their assumptions or distort their results.
* In regression analysis, you can try transforming your data or using a robust regression analysis available in some statistical packages.
* Finally, bootstrapping techniques use the sample data as they are and don’t make assumptions about distributions.

These types of analyses allow you to capture the full variability of your dataset without violating assumptions and skewing results.