Lagunita is retiring and will shut down at 12 noon Pacific Time on March 31, 2020. A few courses may be open for self-enrollment for a limited time. We will continue to offer courses on other online learning platforms; visit http://online.stanford.edu.

Bookmark this page

Understanding Outliers

Learning Objective: Summarize and describe the distribution of a quantitative variable in context: a) describe the overall pattern, b) describe striking deviations from the pattern.

Learning Objective: Relate measures of center and spread to the shape of the distribution, and choose the appropriate measures in different contexts.

Understanding Outliers

We just practiced one way to 'flag' possible outliers. Why is it important to identify possible outliers, and how should they be dealt with? The answers to these questions depend on the reasons for the outlying values. Here are several possibilities:

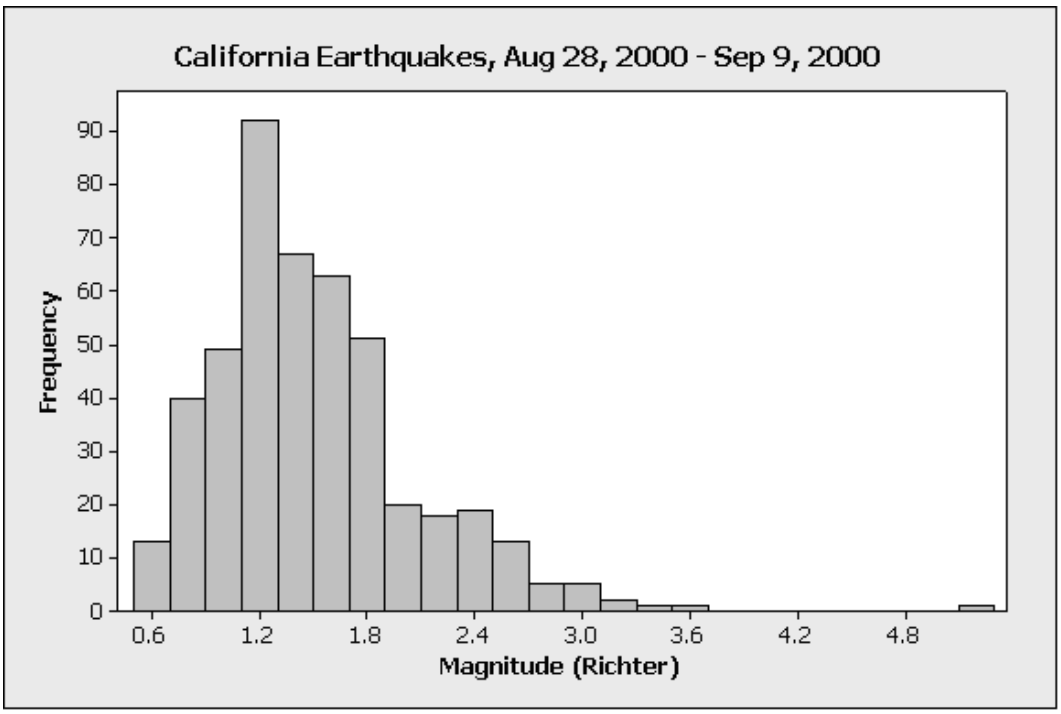
1. Even though it is an extreme value, if an outlier can be understood to have been produced by **essentially the same sort of physical or biological process** as the rest of the data, and if such extreme values are expected to **eventually occur again**, then such an outlier indicates something important and interesting about the process you're investigating, and it **should be kept** in the data.

2. If an outlier can be explained to have been produced under fundamentally **different** conditions from the rest of the data (or by a fundamentally different process), such an outlier **can be removed** from the data if your goal is to investigate only the process that produced the rest of the data.

3. An outlier might indicate a **mistake** in the data (like a typo, or a measuring error), in which case it **should be corrected if possible or else removed** from the data before calculating summary statistics or making inferences from the data (and the reason for the mistake should be investigated).

Here are examples of each of these types of outliers:

1. The following histogram displays the magnitude of 460 earthquakes in California, occurring in the year 2000, between August 28 and September 9:



Identifying the outlier:

On the very far right edge of the display (beyond 4.8), we see a low bar; this represents one earthquake (because the bar has height of 1) that was much more severe than the others in the data.

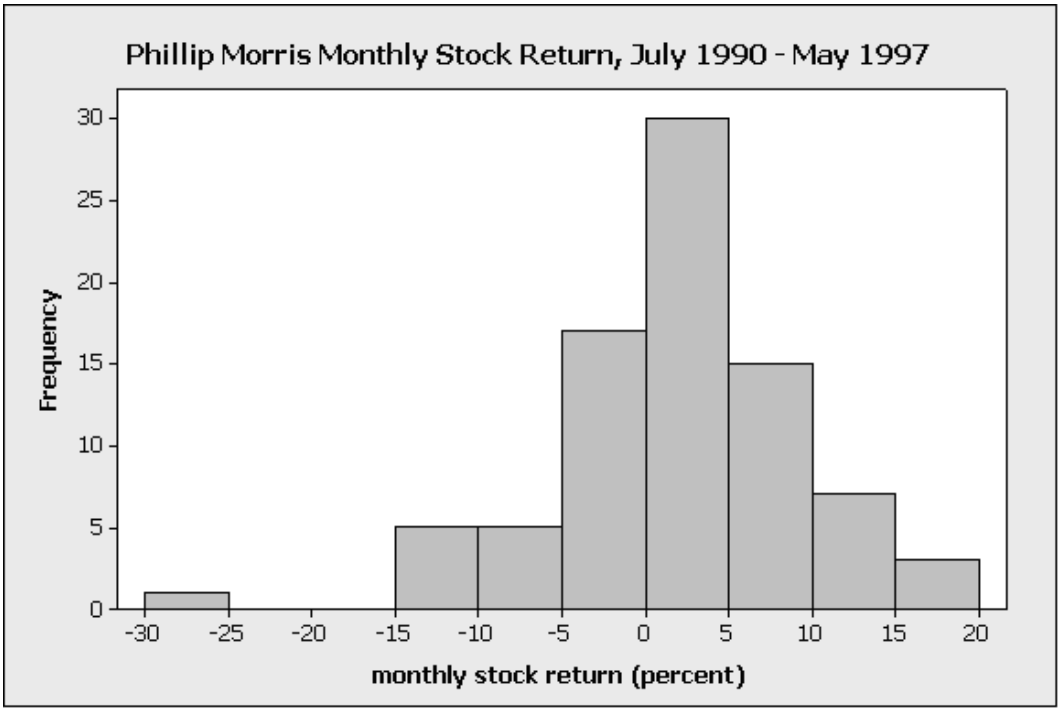
Understanding the outlier:

In this case, the outlier represents a much stronger earthquake, which is relatively rarer than the smaller quakes that happen more frequently in California.

How to handle the outlier:

For many purposes, the relatively severe quakes represented by the outlier might be the most important (because, for instance, that sort of quake has the potential to do more damage to people and infrastructure). The smaller-magnitude quakes might not do any damage, or even be felt at all. So, for many purposes it could be important to keep this outlier in the data.

2. The following histogram displays the monthly percent return on the stock of Phillip Morris (a large tobacco company) from July 1990 to May 1997:



Identifying the outlier:

On the display, we see a low bar far to the left of the others; this represents one month's return (because the bar has height of 1), where the value of Phillip Morris stock was unusually low.

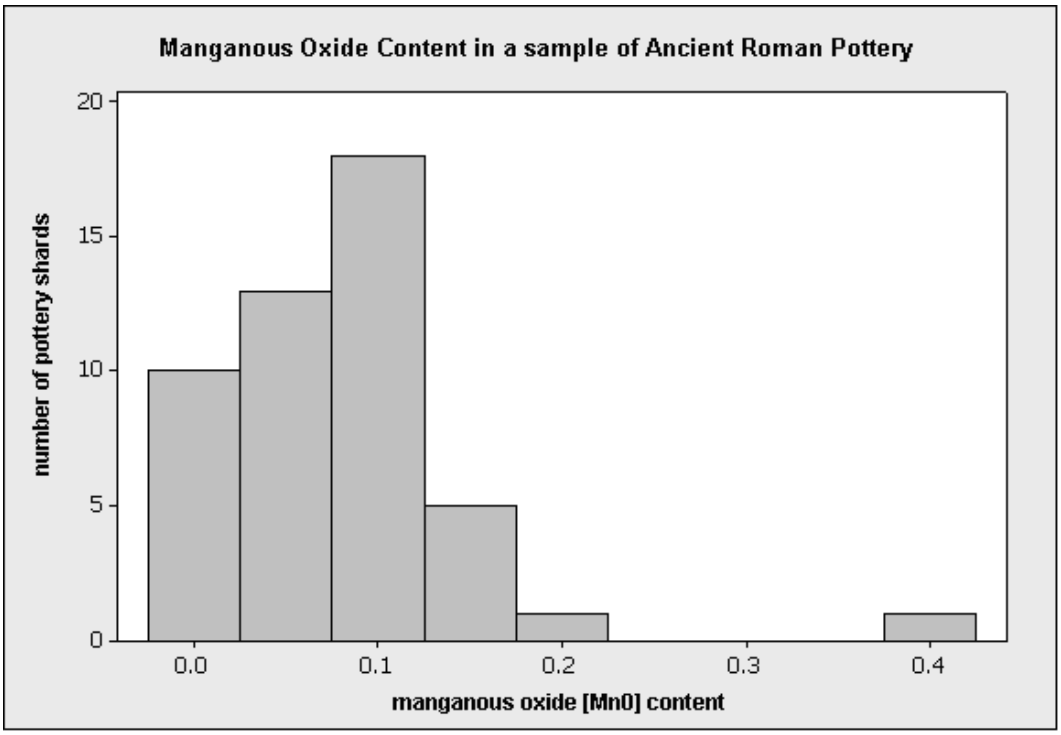
Understanding the outlier:

The explanation for this particular outlier is that, in the early 1990s, there were highly-publicized federal hearings being conducted regarding the addictiveness of smoking, and there was growing public sentiment against the tobacco companies. The unusually low monthly value in the Phillip Morris dataset was due to public pressure against smoking, which negatively affected the company's stock for that particular month.

How to handle the outlier:

In this case, the outlier was due to unusual conditions during one particular month that aren't expected to be repeated, and that were fundamentally different from the conditions that produced the values in all the other months. So in this case, it would be reasonable to remove the outlier, if we wanted to characterize the 'typical' monthly return on Phillip Morris stock.

3. When archaeologists dig up objects such as pieces of ancient pottery, chemical analysis can be performed on the artifacts. The chemical content of pottery can vary depending on the type of clay as well as the particular manufacturing technique. The following histogram displays the results of one such actual chemical analysis, performed on 48 ancient Roman pottery artifacts from archaeological sites in Britain:



Identifying the outlier:

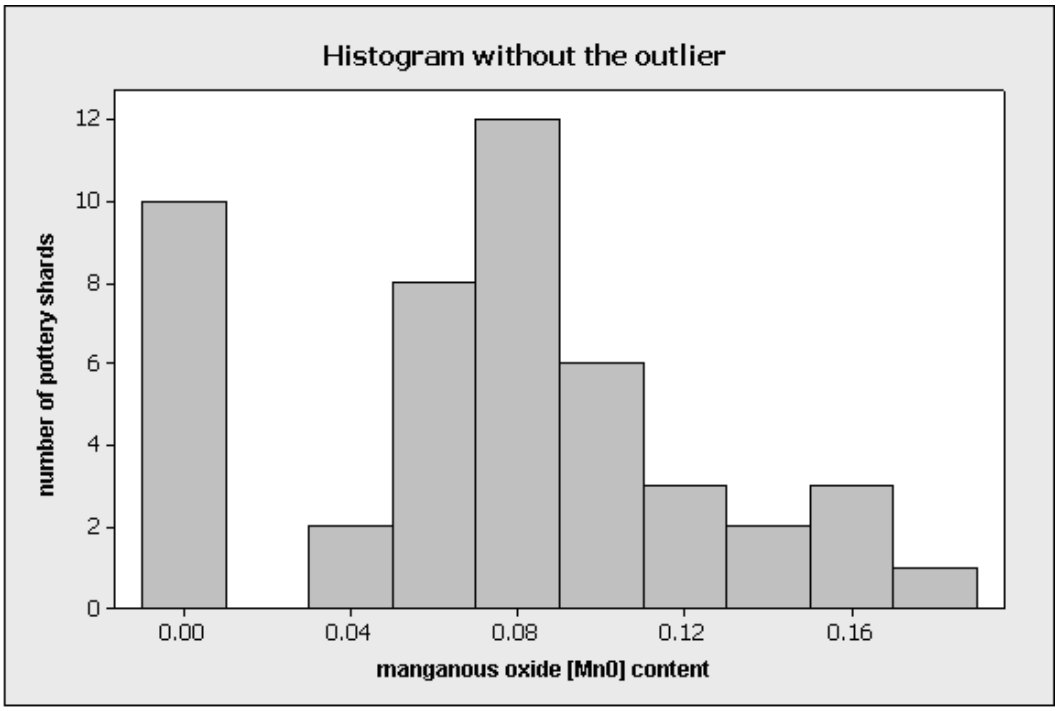
On the display, we see a low bar far to the right of the others; this represents one piece of pottery (because the bar has a height of 1), which has a suspiciously high manganous oxide value.

Understanding the outlier:

Based on comparison with other pieces of pottery found at the same site, and based on expert understanding of the typical content of this particular compound, it was concluded that the unusually high value was most likely a typo that was made when the data were published in the original 1980 paper (it was typed as ".394" but it was probably meant to be ".094").

How to handle the outlier:

In this case, since the outlier was judged to be a mistake, it should be removed from the data before further analysis. In fact, removing the outlier is useful not only because it's a mistake, but also because doing so reveals important structure that was otherwise hidden. This feature is evident on the next display:



When the outlier is removed, the display is re-scaled so that now we can see the set of 10 pottery pieces that had almost no manganous oxide. These 10 pieces might have been made with a different potting technique, so identifying them as different from the rest is historically useful. This feature was only evident after the outlier was removed.

Let's Summarize

- The range covered by the data is the most intuitive measure of spread and is exactly the distance between the smallest data point (min) and the largest one (Max).
- Another measure of spread is the inter-quartile range (IQR), which is the range covered by the middle 50% of the data.
- IQR = Q3 - Q1, the difference between the third and first quartiles. The first quartile (Q1) is the value such that one quarter (25%) of the data points fall below it, or the median of the bottom half of the data. The third quartile is the value such that three quarters (75%) of the data points fall below it, or the median of the top half of the data.
- The IQR should be used as a measure of spread of a distribution only when the median is used as a measure of center.
- The IQR can be used to detect outliers using the 1.5(IQR) criterion. Outliers are observations that fall below Q1 - 1.5(IQR) or above Q3 + 1.5(IQR).

Open Learning Initiative

Unless otherwise noted this work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

© All Rights Reserved