Identifying Outliers in Data

Consider the following unidimensional data – test scores: 25, 29, 3, 32, 85, 33, 27, 28 where both 3 and 85 appear to be "outliers." That is, they do not appear to be consistent with the other data. 3 seems unusually low and 85 seems unusually high. How do we identify outliers in data? Here are a few guidelines because there really are no hard and fast rules.

On what basis was it determined which scores are outliers in the above data set? It depends upon the judge's criteria. Data points that are typically more than $\pm 3\sigma$ (see SLH document ***Range, Normal Distribution, and Standard Deviation*** for a definition) from the mean are often rejected as outliers; oftentimes the criteria might be $\pm 1\sigma$. In the case of normally distributed data, the three-sigma rule means that roughly 1 in 22 observations will differ by twice the standard deviation or more from the mean, and 1 in 370 will deviate by three times the standard deviation. Criteria for rejection vary. On what basis were the 3 and 85 rejected? Consider the following:

Mean = 32.75
Standard Deviation (σ) = 23.13

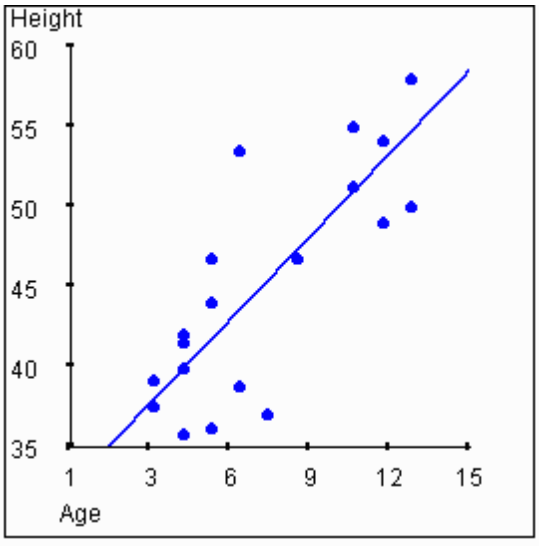
$Mean - 1\sigma = 32.75 - 23.13 = 9.62$ (minimum "normal" value)
$Mean + 1\sigma = 32.75 + 23.13 = 55.88$ (maximum "normal" value)

3 is below the -1σ cutoff
85 is above the $+1\sigma$ cutoff

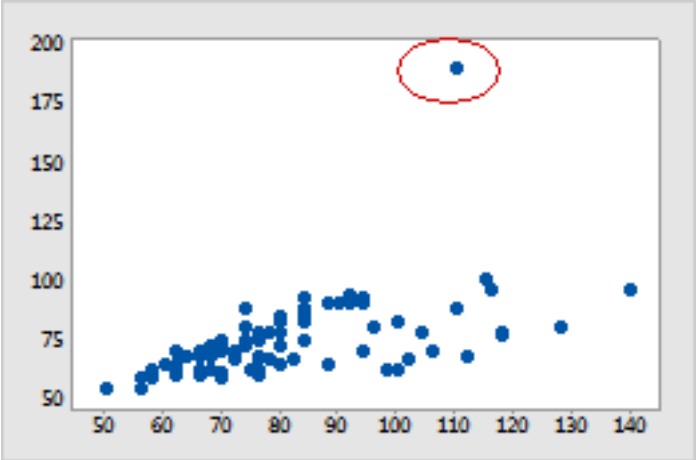
Ergo, both 3 and 85 are considered outliers by the criterion that "normal values" differ from the mean by $\pm 1\sigma$. Would either or both 3 and 85 be considered outliers if the criterion was $Mean \pm 2\sigma$?

What causes outliers? It depends upon the type of data being collected. For instance, if the outliers are in academic test data, it could be that a student had a bad night prior to the test. Perhaps the student was sick or there was a recent death in the family, or perhaps the student made a consistent mistake when placing answers on a Scantron or Opscan form. Such things can and do happen in the real world. Because it's not always easy to know the cause (and even more work for the teacher to prepare make-up exercises), many teachers drop one or more lowest scores as a matter of fairness. Students understand the situation and generally appreciate this act of both kindness in a teacher.

If the outliers are to be made clear from the analysis of a graph (multidimensional data), they are evidenced by the fact that they do not appear to "fit" the other data. Consider for instance a chart of height versus age. It's a well-known fact that height increases somewhat uniformly with age from about 1 to 15 years of age, but there are exceptions. Look at the 6-year-old who stands 55" tall. This person is unusually tall for that age. Look at the 7.5-year-old who stands only 37" tall – this person is unusually short for that age. These are clearly outliers who might have an endocrine problem (excess or deficit of a growth hormone). Biological data such as these tend to have lots of variation and correlation coefficients are used to tell about their degree of consistency (See Correlation Coefficients in *Student Lab Handbook.)*



When plotting physical data, outliers are typically obvious such as in the graph shown to the right. Look at the datum in the graph with the red circle around it. It clearly is an outlier because it does not seem to conform to the trend of the other data.



Outliers in other physical data (in say lab data, x vs. y) can have problems similarly despite the fact that data are typically mathematically precise (conforming to linear, power, exponential distribution, etc.). Outliers just don't seem to "fit" the trend of the data. These outliers can often be readily explained:

1. The data collection process was carelessly done.
2. A measuring device or sensor was inadvertently misread.
3. A measuring device might not have been properly calibrated or set up in this instance.
4. There was a physical fault in the measuring device in this instance.
5. Data might have been inaccurately recorded.
6. Data might have been inaccurately plotted.

Such outlying data might actually be recollected in lab situations. This is better than rejecting them out of hand and merely eliminating them. Who knows? Perhaps you've discovered an interesting scientific phenomenon! It's therefore best to recheck suspected data when this might be the case.

In the article Chi-Square Test for Goodness of Fit, you will learn how to deal with the natural scatter of data that might or might not lead you to the conclusion that data fit a particular model with a given degree of certainty.