**Outlier Analysis**

**Introduction:**

An outlier is an observation that lies an abnormal distance from other values in a random sample from a population. In a sense, this definition leaves it up to the analyst (or a consensus process) to decide what will be considered abnormal. Before abnormal observations can be singled out, it is necessary to characterize normal observations. It needs close attention else it can result in wildly wrong estimations.

Outliers can drastically change the results of the data analysis and statistical modeling. There are numerous unfavorable impacts of outliers in the data set:

1. It increases the error variance and reduces the power of statistical tests
2. If the outliers are non-randomly distributed, they can decrease normality
3. They can also impact the basic assumption of Regression, ANOVA and other statistical model assumptions.

Causes of Outliers:

1. Poor data quality / contamination
2. Low quality measurements, malfunctioning equipment, manual error
3. Correct but exceptional data

Examination of the data for unusual observations that are far removed from the mass of data. These points are often referred to as outliers. There are two graphical techniques for identifying outliers [scatter plots](http://www.itl.nist.gov/div898/handbook/eda/section3/scattera.htm) and [box plots](http://www.itl.nist.gov/div898/handbook/eda/section3/boxplot.htm).

Statistical ways to Remove Outliers:

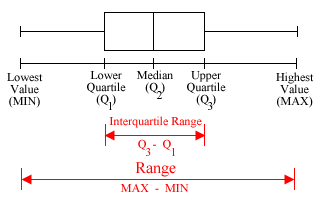
1. Boxplot Method
2. Grubbs’Test

**Boxplot Method:**

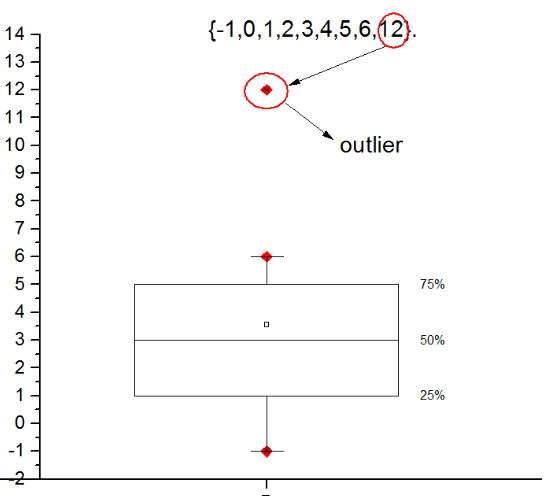
The box plot is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th [percentiles](http://www.itl.nist.gov/div898/handbook/prc/section2/prc252.htm)). If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the interquartile range or IQ.

A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median. The following quantities (called fences) are needed for identifying extreme values in the tails of the distribution:

1. Lower inner fence: Q1 – 1.5IQ
2. Upper inner fence: Q3 + 1.5IQ
3. Lower outer fence: Q1 - 3IQ
4. Upper outer fence: Q3 + 3.5IQ



A point beyond an inner fence on either side is considered a mild outlier. A point beyond an outer fence is considered an extreme outlier.



In the above figure red point is beyond the outer fence and considered as outlier.

**Grubbs’Test:**

Grubbs' test is used to detect a single [outlier](http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm) in a univariate data set that follows an [approximately normal](http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h.htm#Normality) distribution. Grubbs' test is based on the assumption of [normality](https://en.wikipedia.org/wiki/Normal_distribution).  Grubbs’Test is also known as the maximum normed residual test. Grubbs' test is defined for the hypothesis:

H0: There are no outliers in the dataset

Ha: There is exactly one outlier in the dataset.

The Grubbs' test statistic is defined as:



Y¯ and s denoting the sample mean and standard deviation, respectively

Grubbs' test can be used to answer the following questions:

1. Is the maximum value an outlier?
2. Is the minimum value an outlier?

**Care to be taken:**

Outliers should be investigated carefully. Often they contain valuable information about the process under investigation or the data gathering and recording process. Outliers can be a great source of information. Deviating from the norm could be a signal of suspicious activity, breaking news, or an opportunistic or catastrophic event.

Before considering the possible elimination of these points from the data, one should try to understand why they appeared and whether it is likely similar values will continue to appear. Of course, outliers are often bad data points.

**Additional information:**

Research topic: You can replace outliers with NA and use imputation method to impute missing values assuming outliers are NA. You can use mean median or KNN imputation to fill missing values. This way you can save valuable information of other variables instead of removing whole observation.

**Applications**:

1. Fraud Detection (Credit card, telecommunications, criminal activity in e-Commerce)
2. Customized Marketing (high/low income buying habits)
3. Medical Treatments (unusual responses to various drugs)

**Interview Questions:**

1. What do you understand by outliers and inliers?
2. What are the methods to screen outliers?
3. How can outlier values be treated?