Statistical Outliers [1]

Siddharth Kalla [2]64.7K reads

Statistical outliers are data points that are far removed and numerically distant from the rest of the points. Outliers occur frequently in many statistical analyses and it is important to understand them and their occurrence in the right context of the study to be able to deal with them.

An outlier can be a chance phenomenon, measurement error or due to an experimental error [3]. It can also occur in special cases that have a heavy tail distribution, in which cases the assumption of a normal distribution [4] may not hold.

Certain statistical estimators are able to deal with statistical outliers [5] and are robust, while others cannot deal with them. A typical example is the case of a median [6], that can deal with outliers well, since it would not matter whether the extreme point is far away or near the other data points, as long as the central value is unchanged.

The mean [7], on the other hand, is affected by outliers as it increases or decreases in value depending on the position of the outlier.

One should be careful while dealing with outliers and not mistake them for experimental errors or exceptions at all times. outliers can indicate a different property and may indicate that they belong to a different population.

Many times, outliers should be given special attention till their cause is known, which is not always random or chance. Therefore a study needs to be made before an outlier is discarded.

Statistical outliers are common in distributions that do not follow the traditional normal distribution. For example, in a distribution with a long tail, the presence of statistical outliers is more common than in the case of a normal distribution.

In case of a normal distribution, it is easy to see that at random, about 1 in 370 observations will deviate by more than three times the standard deviation [8] from the mean [7]. This ratio decreases drastically for more distant values. Therefore if there is a more than frequent case of data away from the mean, then the cause needs to be examined.

For example, if out of 1000 data points, 5 points are at a distance of four times the standard deviation or more, then these outliers need to be examined.

Links

[1] https://explorable.com/statistical-outliers
[2] https://explorable.com/users/siddharth
[3] https://explorable.com/experimental-error
[4] https://explorable.com/normal-probability-distribution
[5] http://en.wikipedia.org/wiki/Outlier
[6] https://explorable.com/calculate-median
[7] https://explorable.com/arithmetic-mean
[8] https://explorable.com/measurement-of-uncertainty-standard-deviation