

Box Plot

In descriptive statistics, a box plot or box plot (also known as a box-and-whisker diagram or plot) is a convenient way of graphically depicting groups of numerical data through their five-number summaries (the smallest observation (sample minimum), lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation (sample maximum)). A box plot may also indicate which observations, if any, might be considered **outliers**. Box plots can be useful to display differences between populations without making any assumptions of the underlying statistical distribution: they are non-parametric. The spacing's between the different parts of the box help indicate the degree of dispersion (spread) and skewness in the data, and identify outliers. Box plots can be drawn either horizontally or vertically.

Outliers

In statistics, an outlier is an observation that is numerically distant from the rest of the data.

They can occur by chance in any distribution, but they *are often indicative either of measurement error or that the population has a heavy-tailed distribution*. In the former case one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high *kurtosis*¹ *and that one should be very cautious in using tool or intuitions that assume a normal distribution*. A frequent cause of outliers is a mixture of two distributions, which may be two distinct sub-populations, or may indicate "correct trial" versus "measurement error"; this is modeled by a mixture model.

In most larger samplings of data, some data points will be further away from the sample mean than what is deemed reasonable. This can be due to incidental systematic error or flaws in the theory that generated an assumed family of probability distributions, or it may be that some observations are far from the center of the data. Outlier points can therefore indicate faulty data, erroneous procedures, or areas where a certain theory might not be valid. However, in large samples, a small number of outliers is to be expected (and not due to any anomalous condition).

Outliers, being the most extreme observations, will include the sample maximum or sample minimum, or both, depending on whether they are extremely high or low. However, the sample maximum and minimum need not be outliers, if they are not unusually far from other observations.

Naive interpretation of statistics derived from data sets that include outliers may be misleading. For example, if one is calculating the average temperature of 10 objects in a room, and most are between 20 and 25 degrees Celsius, but an oven is at 175 °C, the median of the data may be 23 °C but the mean temperature will be between 35.5 and 40 °C. In this case, the median better reflects the temperature of a randomly sampled object than the mean; however, naively interpreting the mean as "a typical sample", equivalent to the median, is incorrect. As illustrated in this case, outliers may be indicative of data points that belong to a different population than the rest of the sample set.

Estimators capable of coping with outliers are said to be robust: the median is a robust statistic, while the mean is not.

¹ See Statistics Glossary