Methods for quantifying distributions can be divided into classical measures (mean, standard deviation, etc.) and robust measures (median, interquartile range, etc.).

Classical measures

Advantages: -well known, easily calculated, included in all standard software packages
-statistical properties of these measures can be derived (under certain conditions)
-depend on, and reflect the influence of, all the data

Disadvantages: -sensitive to extreme values (outliers)
-use in classical statistics often depends on restrictive assumptions that are not met

Classical measures are calculated differently (and use different symbols) when applied to populations (where every individual is measured) and to samples (where only a subset of *n* individuals out of the total population of *N* is measured). The quantitative difference is small for any reasonably large *n*.

	Applicable to populations	**Applicable to samples**
Number of measurements	N	n
Individual measurements	x_i	x_i

Mean

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N} \qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(measures central tendency)

Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \qquad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1} = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

(measures dispersion, but in measurement units squared)

Standard deviation

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}} \qquad s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}}$$

(measures dispersion in original measurement units)

Note: for the sample variance and sample standard deviation, three equivalent formulas are shown. The first is conceptually simple but computationally clumsy, since it requires two passes through the data (one to calculate the mean, and the second to sum the squared deviations from the mean). The latter two formulas can be evaluated in a single pass through the data; these "machine formulas" are therefore most often used in calculators and computer software (demonstrating their formal equivalence is left as an exercise for the reader). A machine formula for skewness is also given below.

Skewness

$$\gamma_1 = \frac{1}{\sigma^3} \frac{\sum_{i=1}^{N} (x_i - \mu)^3}{N} \qquad g_1 = \frac{1}{s^3} \frac{\sum_{i=1}^{n} (x_i - \bar{x})^3}{\left(\frac{(n-1)(n-2)}{n}\right)} = \frac{1}{s^3} \frac{\sum_{i=1}^{n} x_i^3 - 3\bar{x}\sum_{i=1}^{n} x_i^2 + 2n\bar{x}^3}{\left(\frac{(n-1)(n-2)}{n}\right)}$$

Positive skewness indicates right skew; negative skewness implies left skew.

Kurtosis

$$\gamma_2 = \frac{1}{\sigma^4} \frac{\sum_{i=1}^{N} (x_i - \mu)^4}{N} - 3 \qquad g_2 = \frac{1}{s^4} \left(\frac{\sum_{i=1}^{n} (x_i - \bar{x})^4}{\left(\frac{(n-2)(n-3)(n-1)}{n(n+1)}\right)} - \frac{3\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right)^2}{(n-2)(n-3)} \right) - 3$$

Positive kurtosis indicates a leptokurtic distribution (narrow peak with long tails); negative implies a platykurtic distribution ("flat" with short tails).

Robust measures (also termed resistant measures)

Advantages:	-insensitive to extreme values
Disadvantages:	-more tedious to calculate -less familiar to readers

Median (robust measure of central tendency)

The median is the 50th percentile (the 0.5 quantile) of the data:

$$median = \begin{cases} x_{(n+1)/2} & if\ n\ is\ odd \\ 1/2\left(x_{(n/2)} + x_{(n/2+1)}\right) & if\ n\ is\ even \end{cases}$$

Trimmed mean (robust measure of central tendency)

The "*q* percent trimmed mean" is the mean of the observations, after *q* percent of the observations have been removed from each end of the distribution. "Trimmed mean", when stated without a percentage, is *usually* the 25 percent trimmed mean (the mean of the middle 50 percent of the data values).

Interquartile range (robust measure of dispersion)

The IQR is the difference between the 75th and the 25th percentiles, that is, the range spanned by the middle 50 percent of the observations:

$$IQR = Q(0.75) - Q(0.25)$$

where the quantiles *Q* are calculated as described in the toolkit on graphically displaying data distributions.

Median absolute deviation (robust measure of dispersion)

The MAD is the median of the distances between each data point and the overall median for the data set:

$$MAD = median(\,d\,) \qquad where \qquad d_i = \left|x_i - median(\,x\,)\right|$$

Beware that exactly the same acronym has also been used as shorthand for a <u>completely different</u> measure, the mean absolute deviation, which is the *mean* absolute value of deviations from the *mean*, rather than the *median* absolute value of deviations from the *median*:

$$MAD = \frac{1}{n}\sum_{i=1}^{n}\left|x_i - \overline{x}\right|$$

Either version of *MAD* is more resistant to outliers than the standard deviation is, but of the two, the *median* absolute deviation is more robust than the *mean* absolute deviation.

Quartile skew coefficient (robust measure of skewness)

The quartile skew coefficient *qs* is the difference between the distances from the median to the upper and lower quartiles, scaled by the IQR:

$$qs = \frac{\left(Q(0.75) - Q(0.5)\right) - \left(Q(0.5) - Q(0.25)\right)}{\left(Q(0.75) - Q(0.25)\right)}$$

Note that the IQR and *qs*, as well as the trimmed mean, can be calculated using percentiles other than the quartiles (such as the 10th and 90th percentiles).

<u>**Characteristics of good descriptive statistics**</u>

lack of bias	(parameters measured on small samples are equally likely to be too high as too low)
robustness	(not unduly influenced by single values, e.g. wild outliers)
efficiency	(individual small samples yield estimates close to true value for whole population)
sufficiency	(use all of the data)
consistency	(as sample gets more complete, parameter converges to the true value for the population)