

Biostatistics: DESCRIPTIVE STATISTICS: 1, AVERAGES

1. Introduction

Statistical analysis is commonly concerned with evaluating samples taken from one or more populations. In the case where a sample comprises several individuals or observations, it is usual to calculate a single representative value that is a consensus observation, sometimes referred to as the 'central tendency'. Often statistical tests assess whether this value is likely to be representative of the population as a whole, or whether the consensus values from two or more samples differ.

2. Mean of a set of observations of a continuous variable

The simplest descriptive statistic is the arithmetic mean, or simply the 'mean'. This is calculated as the average of a series of observations of a continuous variable (ie one where any value can occur). If a sample consists of several observations $x_1 \dots x_n$, then the mean is calculated as:

$$\text{mean } x = \Sigma(x_1 \dots x_n)/n$$

where the expression $\Sigma(x_1 \dots x_n)$ is the sum of all of the **n** observations in the series $x_1 \dots x_n$. (If this notation is difficult for you, refer to the helpsheets on **variables and parameters** and **equations**.)

Here is an example data set, with a sample comprising ten observations

Observation	1	2	3	4	5	6	7	8	9	10
Value	1.69	1.55	2.36	1.73	0.89	1.39	1.79	2.58	1.21	2.10

$$\Sigma(x_1 \dots x_n) = 17.29$$

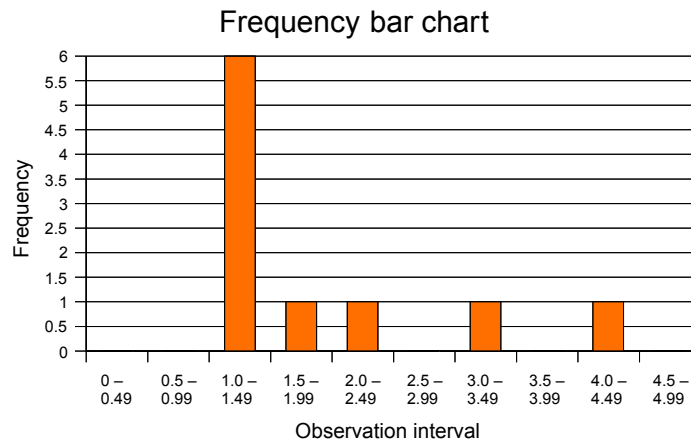
$$n = 10$$

$$\text{mean } x = \Sigma(x_1 \dots x_n)/n = 1.73$$

We introduced the mean as the simplest descriptive statistic, suggesting that there are other ways to represent the 'average' for a sample. So what determines when it is appropriate to calculate a mean value, and more importantly when not to? Consider a second set of observations:

Observation	1	2	3	4	5	6	7	8	9	10
Value	1.37	1.45	1.23	1.67	3.19	1.39	1.41	1.27	2.10	4.24

If we group observations into intervals of 0.5, six of the ten fall in the interval 1.0 to 1.49. Yet if we calculate a mean value using the formula given above, the value is 1.93, which falls outside the interval containing most of the observations. This is because three values are greater than 2, with the highest being 4.24. If we plot the distribution of these observations, it is clear that the high values have a disproportionate effect on the calculation of the mean:



Such a distribution of observations is termed 'skewed', where most of the observations are clustered in one part of the range but there are a few points that form a 'tail' to one side or other of the main group. This is opposed to a symmetrical distribution, such as the classical normal distribution, where most of the points cluster close to the middle of the overall range, and outliers are distributed equally on either side.

This demonstrates that a mean is an ideal representation of the average value for group of observations with a normal distribution, or a good approximation to it, but fails to represent the average value for the sample if the distribution of observations is skewed or otherwise markedly different from normality.

The mean of a set of observations in a spreadsheet can be calculated easily using the 'AVERAGE' function. The expression '=AVERAGE(C1:C9)' calculates the mean of the values in cells C1 to C9 of column C, whilst '=AVERAGE(B5:F5)' calculates the mean of the values B5 to F5 in row 5.

3. Median of a set of observations

In discussing the mean, we indicated that this statistic is appropriate where a the observations in a sample of scalar data are distributed normally, but not if the distribution is skewed. A quantity termed the 'median' provides a better measure of the average value under such conditions.

The median is simply the 'middle' value in a data series. If the n observations in the series $x_1 \dots x_n$ are ordered from the smallest to largest values, the median is the value of the mid-point of the ordered series if n is an odd number, or the mean of the two values

either side of the mid-point if n is an even number. Using the skewed data set as an example:

Observation	1	2	3	4	5	6	7	8	9	10
Value	1.37	1.45	1.23	1.67	3.19	1.39	1.41	1.27	2.10	4.24
Ordered series	1.23	1.27	1.37	1.39	1.41	1.45	1.67	2.10	3.19	4.24

In this case, $n = 10$ is an even number, so that the median is the mean of values 5 and 6 (highlighted), which gives a value of 1.43. Compare this with the mean of 1.93, which lies closest to value 8 in the ordered series.

If you calculate the median of a normal distribution, it is very close to the value of the mean (and should be identical for a large sample from a truly normal distribution). For the first data set, considered in the calculation of the mean, the median value is 1.71 whilst the mean was 1.73.

4. Mode

Mean and median are appropriate representations of the 'average' value for a set of observations of a continuous variable, that is one that can take any value. Neither is suited to category data, for instance where an event or response is scored on a scale from 'high' to 'low' or 'easy' to 'difficult'. Imagine ten observers are asked to rate a particular task on a five-point scale from 1 for 'very difficult' to 5 for 'very easy':

Observer	1	2	3	4	5	6	7	8	9	10
Rating	5	3	5	2	3	4	5	4	2	5

Taking a mean of the ratings gives a value of 3.8, which is best expressed as 'just a bit short of easy'. Although this value is quite close to one of the category scores, it suggests a resolution that is inappropriate – the numerical values are simply labels and have no intrinsic significance. In this instance, the most suitable way to express a representative value is the mode. This is the category in which most observations or scores lie – in this case four of the ten observers rated the task as 'very easy' (score of 5).