A statistical population is a set of entities concerning which statistical inferences are to be drawn, often based on a random sample taken from the population. For example, if we were interested in generalizations about crows, then we would describe the set of crows that is of interest. Notice that if we choose a population like all crows, we will be limited to observing crows that exist now or will exist in the future. Probably, geography will also constitute a limitation in that our resources for studying crows are also limited.

Population is also used to refer to a set of potential measurements or values, including not only cases actually observed but those that are potentially observable. Suppose, for example, we are interested in the set of all adult crows now alive in the county of Cambridgeshire, and we want to know the mean weight of these birds. For each bird in the population of crows there is a weight, and the set of these weights is called the population of weights.

Standard deviation is a widely used measurement of variability or diversity used in statistics and probability theory. It shows how much variation or "dispersion" there is from the "average" (mean, or expected/budgeted value). A low standard deviation indicates that the data points tend to be very close to the mean, whereas high standard deviation indicates that the data are spread out over a large range of values.

Technically, the standard deviation of a statistical population, data set, or probability distribution is the square root of its variance. It is algebraically simpler though practically less robust than the average absolute deviation.[1][2] A useful property of standard deviation is that, unlike variance, it is expressed in the same units as the data. Note, however, that for measurements with percentage as unit, the standard deviation will have percentage points as unit.

In addition to expressing the variability of a population, standard deviation is commonly used to measure confidence in statistical conclusions. For example, the margin of error in polling data is determined by calculating the expected standard deviation in the results if the same poll were to be conducted multiple times. The reported margin of error is typically about twice the standard deviation – the radius of a 95 percent confidence interval. In science, researchers commonly report the standard deviation of experimental data, and only effects that fall far outside the range of standard deviation are considered statistically significant – normal random error or variation in the measurements is in this way distinguished from causal variation. Standard deviation is also important in finance, where the standard deviation on the rate of return on an investment is a measure of the volatility of the investment.

The standard error of a method of measurement or estimation is the standard deviation of the sampling distribution associated with the estimation method.[1] The term may also be used to refer to an estimate of that standard deviation, derived from a particular sample used to compute the estimate.

For example, the sample mean is the usual estimator of a population mean. However, different samples drawn from that same population would in general have different values of the sample mean. The standard error of the mean (i.e., of using the sample mean as a method of estimating the population mean) is the standard deviation of those sample means over all possible samples (of a given size) drawn from the population. Secondly, the standard error of the mean can refer to an estimate of that standard deviation, computed from the sample of data being analysed at the time.

A way for remembering the term standard error is that, as long as the estimator is unbiased, the standard deviation of the error (the difference between the estimate and the true value) is the same as the standard deviation of the estimates themselves; this is true since the standard deviation of the difference between the random variable and its expected value is equal to the standard deviation of a random variable itself.

In practical applications, the true value of the standard deviation is usually unknown. As a result, the term standard error is often used to refer to an estimate of this unknown quantity. In such cases it is important to be clear about what has been done and to attempt to take proper account of the fact that the standard error is only an estimate. Unfortunately, this is not often possible and it may then be better to use an approach that avoids using a standard error, for example by using maximum likelihood or a more formal approach to deriving confidence intervals. One well-known case where a proper allowance can be made arises where Student's t-distribution is used to provide a confidence interval for an estimated mean or difference of means. In other cases, the standard error may usefully be used to provide an indication of the size of the uncertainty, but its formal or semi-formal use to provide confidence intervals or tests should be avoided unless the sample size is at least moderately large. Here "large enough" would depend on the particular quantities being analysed.

In statistics, a histogram is a graphical representation, showing a visual impression of the distribution of data. It is an estimate of the probability distribution of a continuous variable and was first introduced by Karl Pearson [1]. A histogram consists of tabular frequencies, shown as adjacent rectangles, erected over discrete intervals (bins), with an area equal to the frequency of the observations in the interval. The height of a rectangle is also equal to the frequency density of the interval, i.e., the frequency divided by the width of the interval. The total area of the histogram is equal to the number of data. A histogram may also be normalized displaying relative frequencies. It then shows the proportion of cases that fall into each of several categories, with the total area equaling 1. The categories are usually specified as consecutive, nonoverlapping intervals of a variable. The categories (intervals) must be adjacent, and often are chosen to be of the same size.[2]

Histograms are used to plot density of data, and often for density estimation: estimating the probability density function of the underlying variable. The total area of a histogram used for probability density is always normalized to 1. If the length of the intervals on the x-axis are all 1, then a histogram is identical to a relative frequency plot.

In statistics, correlation and dependence are any of a broad class of statistical relationships between two or more random variables or observed data values.

Familiar examples of dependent phenomena include the correlation between the physical statures of parents and their offspring, and the correlation between the demand for a product and its price. Correlations are useful because they can indicate a predictive relationship that can be exploited in practice. For example, an electrical utility may produce less power on a mild day based on the correlation between electricity demand and weather. Correlations can also suggest possible causal, or mechanistic relationships; however, statistical dependence is not sufficient to demonstrate the presence of such a relationship.

Formally, dependence refers to any situation in which random variables do not satisfy a mathematical condition of probabilistic independence. In general statistical usage, correlation or co-relation can refer to any departure of two or more random variables from independence, but most commonly refers to a more specialized type of relationship between mean values. There are several correlation coefficients, often denoted ρ or r, measuring the degree of correlation. The most common of these is the Pearson correlation coefficient, which is sensitive only to a linear relationship between two variables (which may exist even if one is a nonlinear function of the other). Other correlation coefficients have been developed to be more robust than the Pearson correlation, or more sensitive to nonlinear relationships.[1][2][3]

The correlation coefficient ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between X and Y perfectly, with all data points lying on a line for which Y increases as X increases. A value of −1 implies that all data points lie on a line for which Y decreases as X increases. A value of 0 implies that there is no linear correlation between the variables.

More generally, note that (Xi − X)(Yi − Y) is positive if and only if Xi and Yi lie on the same side of their respective means. Thus the correlation coefficient is positive if Xi and Yi tend to be simultaneously greater than, or simultaneously less than, their respective means. The correlation coefficient is negative if Xi and Yi tend to lie on opposite sides of their respective means.