Science Online

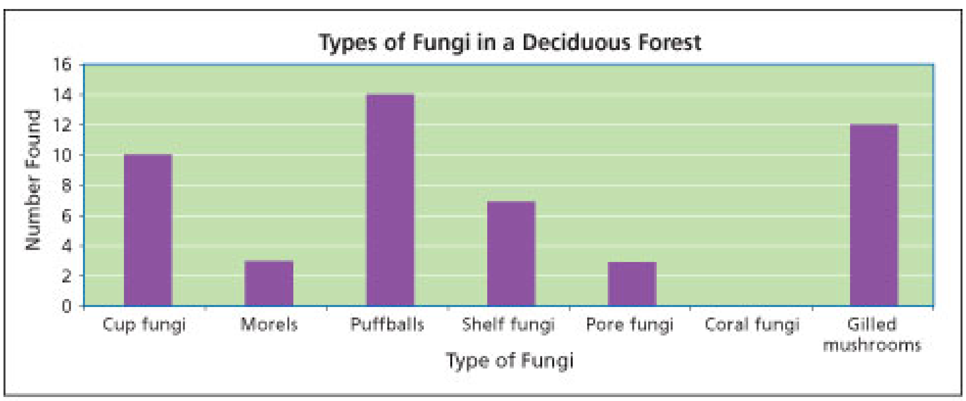
data presentation and analysis

Data are the objective information collected during the course of an experiment and may take the form of measurements or descriptive observations. Data analysis is the conversion of the raw data into useful information from which a researcher can draw conclusions.

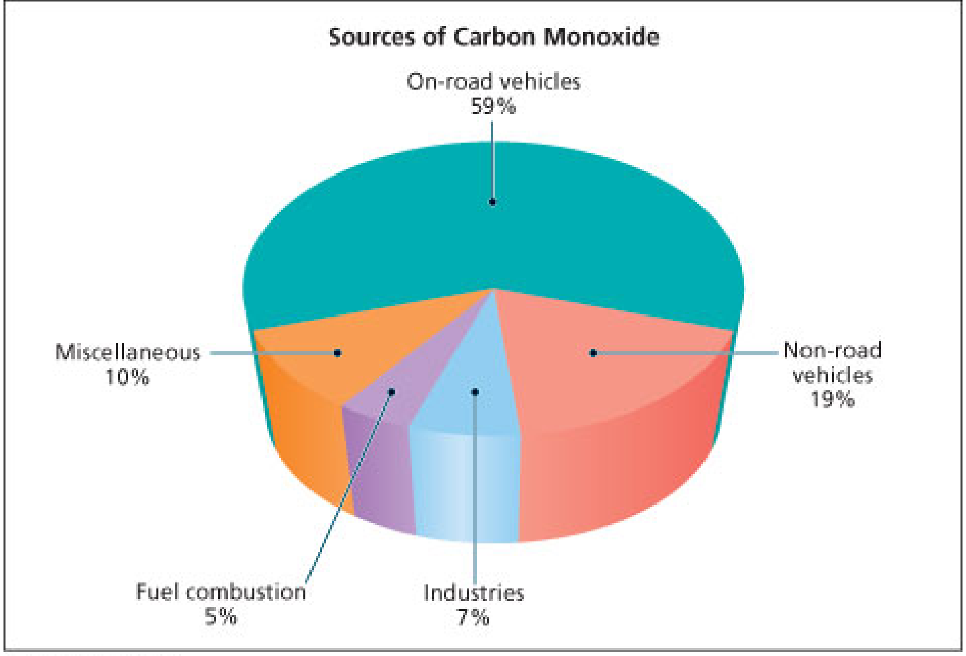
Graphs

Charts, tables, and graphs are useful ways to present data. Selecting the most appropriate method for presenting data facilitates analysis by accentuating patterns. Charts are often used for raw data collection; the researcher records the information directly as lists, notes on a diagram or on a map, or on other defined areas of a chart. After recording all the data on a chart, the researcher may compile the information in table format or in a spreadsheet. Tables consist of rows and columns that contain combined data or information. Generating a table facilitates graphing, especially when the researcher uses computer software to generate the graphs or pictorial representations. The use of the many different forms of graphs to present data reveals trends or patterns that might not be as apparent when using a chart or a table to present the information.

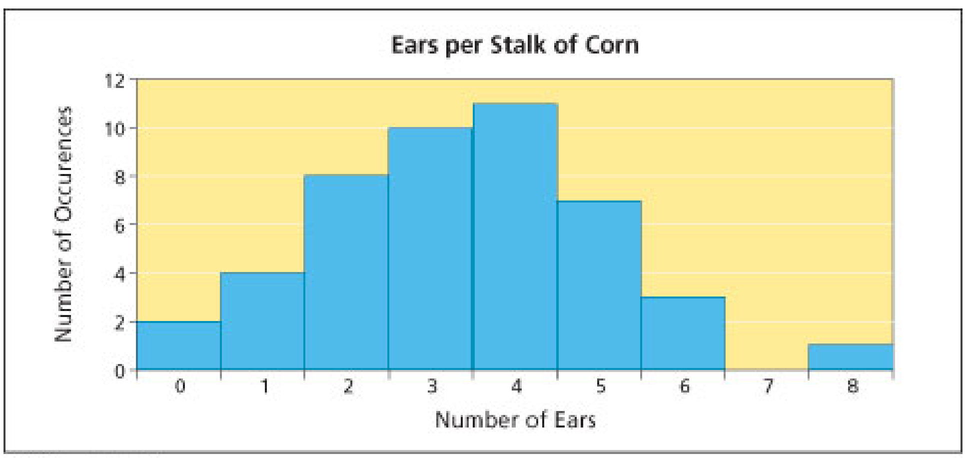
Bar graphs depict quantitative data from discontinuous categories. The horizontal axis contains the distinct categories, and the vertical axis shows the number of occurrences. Vertical bars shaped like rectangles extend upward to the demarcation for the number of units associated with that data set. For example, a researcher surveying a defined area of a deciduous forest for different types of fungal species might depict data in the form of a bar graph. One can look at the graph below and immediately see that puffballs were the most common type of fungus in the observed area of forest, and that coral fungi were not observed at all.



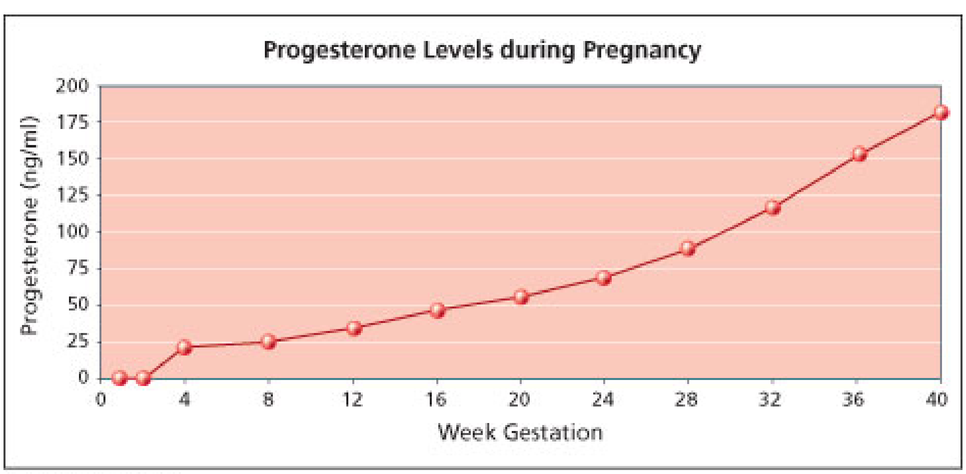
Pie charts or pie graphs show the relative frequencies of subdivisions of a set of data. The area of each "slice" of the circular pie represents the percentage of the whole that a particular subdivision makes up. For example, an environmental scientist may report the sources of carbon monoxide emissions in a particular city in the form of a pie chart like the one below. This allows one to see at a glance that the major source of emissions is on-road vehicles, with nonroad vehicles (such as boats or construction equipment and machinery) contributing the second highest amount.



Histograms also depict frequencies, but they reveal how the frequency of the data relates to another particular variable. One can study a histogram to see the distribution of the frequencies over the entire range of the variable. For example, a botanist might use a histogram to show the frequency of number of ears per stalk for a new strain of corn. The number of ears on a stalk would be indicated on the *x* axis, and the rectangular bars would extend up to the number of occurrences for each number. From the histogram depicted below, one can see that four ears per stalk occurs most frequently and that the numbers ranged from zero to eight ears per stalk.

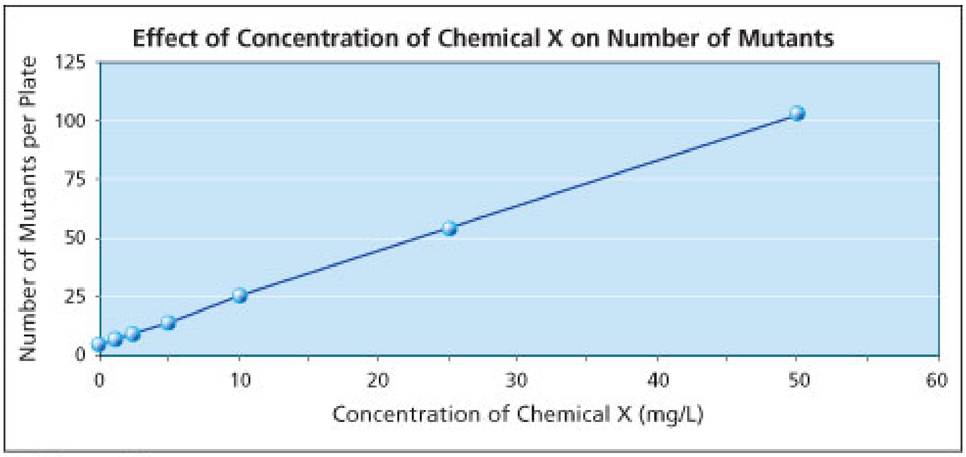


Line graphs reveal relationships between two sets of continuous data and are especially useful for displaying data over time. The data points are plotted and connected by lines that may reveal informative trends. For example, the line graph below shows that circulating levels of the hormone progesterone progressively increase during gestation.



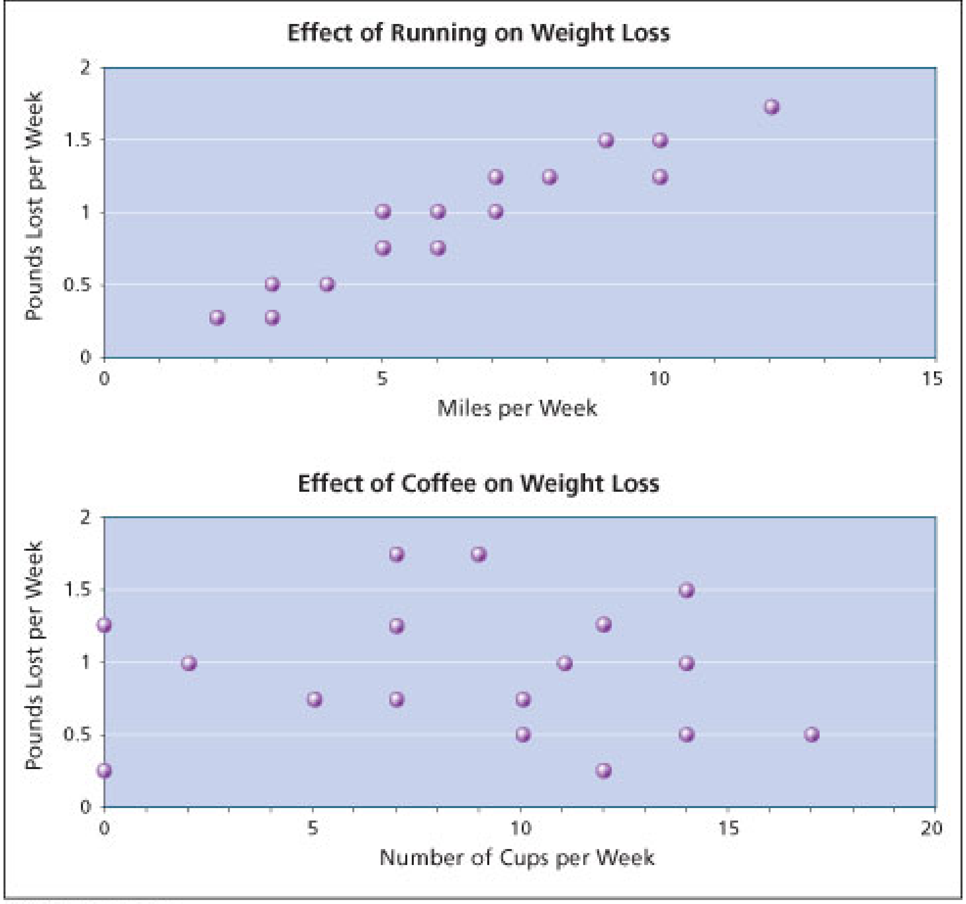
One can infer values for unobserved data by extending the line of a line graph beyond the last data point. Scientists often use this process, called extrapolation, to make predictions about the future effect of a trend if it continues along the same pathway. For example, if data show that the number of fish species in a particular lake is declining at a rate of one per year, and the lake currently supports 17 different species of fish, then one could use extrapolation to predict that if the current trend continues, within 17 years the lake will be devoid of fish. Interpolation is the process of determining a value at an interval between two observed data points. For example, if the ichthyologist responsible for counting the number of fish species in the lake recorded data for every year between 1993 and 2007 with the exception of the year 2002, then one could use interpolation to infer the number of fish that were present in the lake in 2002 from the data obtained before and after that time point.

In a scientific experiment, the independent variable is the factor that the experimenter intentionally manipulates, and the dependent variable is what the experimenter evaluates for change. If the experiment is carefully controlled, meaning the only difference between experimental systems is the independent variable, then the general direction and the steepness of the line drawn between data points demonstrate how changing the independent variable affects the dependent variable, if at all. The independent variable is graphed on the horizontal axis, which is referred to as the *x* axis. The dependent variable is graphed on the vertical axis, which is referred to as the *y* axis. In the line graph, the dependent variable is the number of mutant bacterial colonies that grow on a plate of medium and is a function of the concentration of the chemical mutagen added to the medium, at least for the range of concentrations tested.



One must be careful, however, in concluding from a line graph that one variable is causing the measured variable to change. Some plots may reveal a trend or relationship without one of the variables directly influencing the other. The variables change together in a way not expected on the basis of chance alone, but one does not necessarily cause the other to change. In some cases the researcher cannot ensure that only a single factor is altered during an experiment. Data collection might involve observing a phenomenon that occurs in nature and cannot be manipulated. For example, the average temperature of the ocean may vary at different times of the day, but the researcher cannot experimentally control the many factors that cause this effect (e.g., the intensity of the radiation from the Sun as it reaches the water, oceanic circulation patterns). In this case, the time of day does not cause the change in temperature—other factors that also change over time cause the resulting temperature variations—but plotting the data on a line graph will still reveal the temporal patterns in temperature fluctuations.

Scatter plots are similar to line graphs but are used to determine whether and how one variable relates to another. After the data are entered onto the scatter plot, a computer can generate a line of best fit, also called a regression line, through the data points. If the data points lie on a reasonably straight line, they are said to be correlated. A widely scattered pattern indicates that no correlation exists between the two factors. If the *y* values increase as the *x* values increase, then the variables have a positive correlation, and if the *y* values decrease as the *x* values increase, the variables are said to be negatively correlated. To illustrate this, from the scatter plots (generated from hypothetical data), one can conclude that running larger distances correlates with greater weight loss, whereas the number of cups of coffee someone drinks each week does not appear to be correlated to weight loss.



Statistical Analysis

Scientists use statistical analysis to interpret numerical information obtained during the course of an experiment. Statistics is a branch of mathematics devoted to collecting data and extracting meaningful information from those data. Because the data obtained during an experiment must be analyzed afterward, the researcher must carefully design an experiment that will allow one to perform an efficient and meaningful statistical analysis. Statistics helps the experimenter do this, for example, by indicating how many data must be collected in order to draw reasonable conclusions from the results. Another purpose of statistical analysis is to determine the reliability of a data set. Ideally, the information obtained from the observed or tested sample reflects information that is true for a larger population. In order to determine whether a sample accurately depicts the behavior or a result characteristic for the larger population, one must first be able to summarize the experimental data. The mean and standard deviation are two important statistical descriptions of a sample data set.

The mean (χ) is the average of a sample of numerical values and can be calculated by the following formula:

$$\chi = \frac{\sum_{i=1}^{i=n} x_i}{n}$$

Simply, the mean equals the sum of all the numerical values in the data set divided by the number of values (n). As the sample size increases, that is, as the number of values in the data set increases, the mean of the sample approaches the mean of the population (μ) that the sample represents. (Statisticians use Roman letters to indicate sample values, referred to as statistics, and use Greek letters to represent population variables, referred to as parameters.)

Standard deviation (s) and variance (s^2) indicate the spread of the population, in other words, the degree to which the individual data points differ from the mean. A small standard deviation means that all the measured variables are very close in value. The standard deviation of a sample (s) of n data is an estimate of the standard deviation of the entire population (σ). The variance is its square.