2

Elements of Scientific Data and Tests of Questions

Before we venture into how research might be done, we must discuss some elements of scientific studies, including what kinds of data we might find, what is meant by accuracy and precision of data, and how the nature of data might be revealed by means of frequency distributions. I also need to introduce a few descriptive statistics that will be useful for later discussions. Once we have covered these essentials of data handling, we start our examination of how we might test questions of interest.

2.1 Kinds of Data

We have repeatedly referred above to *variables* without stopping to define what they might be; in our context, variables refer to *properties subject to change*. That may be too general; perhaps we can narrow the idea to *characteristics with respect to which measurements in a sample will vary*.

We collect data to ask, "What is the value of the variable?" Amount of rain, number of leaves per tree, a person's height, class rank, wind velocity, ratio of carbon to nitrogen in soil, concentration of molybdate in seawater, hair color, and electrical resistance all are examples of variables. Each of these can be given a value, which is what we measure or observe.

If we think about the list of variables cited in the preceding paragraph, it may become evident that there are different types of variables. The list seems to include items with rather different properties.

Nominal Data

Nominal data are those that cannot be assigned quantitative properties. Rather, these data can be thought of as classifications, categories, or attributes. Examples might include studies in which we record the number of species of plants, hair color types, country of origin, or whether the subjects are alive or dead.

Nominal data can be examined quantitatively by combining the observations into frequencies. Data that are subject to such grouping are referred to as *enumeration data*. For example, in genetic studies, peas

can be classified into wrinkled and smooth categories, and the number of each category found in samples of n plants can be compared to frequencies expected from a given genetic cross pattern. The frequencies of these groups can be compared using methods discussed in section 3.4.

Ranked Data

Ranked (also called ordinal) data reflect a hierarchy in a classification; they can be ordered or ranked. Examples might be the order of birth among nestlings, or social position of individual lions in a pride. The ranking in this type of data does not imply that the difference between rank 1 and 2, for example, is the same as, or even proportional to, the difference between 3 and 4.

Measurement Data

Measurement data are the most common type of observation. These are characterized by values that are expressible in numeric order, and the intervals between pairs of values are meaningful. The Celsius scale (°C) is of this type. The value "10 °C" is five degrees higher than "5 °C," and the difference between 5 °C and 10 °C is equal to the difference between 20 °C and 25 °C. In this example there is the additional complication that, in fact, this is an arbitrary scale. We cannot say 5 °C is five times the temperature at 1 °C, because 0 °C, the freezing point of water, is only a convenient reference point, not a real zero. Physical chemists use the kelvin scale (K) instead of °C because K provides a true scale in which intervals are multiplicative. For our purposes, the example illustrates that scales of measurement are invariably arbitrary, and we are free to use the scale most appropriate to specific purposes, a topic that we will revisit below when we discuss data transformations.

Measurement data can be continuous or discontinuous. Continuous variables can assume any number of values between any two points. Examples of this type *are* length, area, volume, and temperature. In such data, we could ascertain values to a degree that depends only on the precision of the method of measurement we apply. Discontinuous variables (also called *discrete* or *meristic*) can assume only certain fixed values. Number of fish in a trawl tow, number of young in a nest, or number of teeth in a skull are examples of discontinuous variables. None of these are reasonably expressed—at least in the original data—as "1.6," for example; the data record has to be stated in whole units.

The distinction between continuous and discontinuous variables matters because we analyze or represent these two kinds of data in different ways This distinction may not, however, be as clear as we wish Often it happens that initially discontinuous data are subsequently expressed as continuous data. For example, in fisheries work, each trawl haul cullects only discontinuous data (whole fish), but if we average the number of fish for 10 trawls, we get a number with meaningful decimals. On the other hand, some discontinuous variables are derived from continuous data. For instance, Classes 1–5 for hurricanes (depending on an ascending comparison of wind velocity and other properties) and the

can be classified into wrinkled and smooth categories, and the number of each category found in samples of n plants can be compared to frequencies expected from a given genetic cross pattern. The frequencies of these groups can be compared using methods discussed in section 3.4

Ranked Data

Ranked (also called *ordinal*) data reflect a hierarchy in a classification; they can be ordered or ranked. Examples might be the order of birth among nestlings, or social position of individual lions in a pride. The ranking in this type of data does not imply that the difference between rank 1 and 2, for example, is the same as, or even proportional to, the difference between 3 and 4.

Measurement Data

Measurement data are the most common type of observation. These are characterized by values that are expressible in numeric order, and the intervals between pairs of values are meaningful. The Celsius scale (°C) is of this type. The vaiue "10 °C" is five degrees higher than "5 °C," and the difference between 5 °C and 10 °C is equal to the difference between 20 °C and 25 °C. In this example there is the additional complication that, in fact, this is an arbitrary scale. We cannot say 5 °C is five times the temperature at 1 °C, because 0 °C, the freezing point of water, is only a convenient reference point, not a real zero. Physical chemists use the kelvin scale (K) instead of °C because K provides a true scale in which intervals are multiplicative For our purposes, the example illustrates that scales of measurement are invariably arbitrary, and we are free to use the scale most appropriate for specific purposes, a topic that we will revisit helnw when we discuss data transformations

Measurement data can be continuous or discontinuous *Continuous* variables ran assume any number of values between any two points Examples of this type are length, area, volume, and temperature In such data, we could ascertain values to a degree that depends only on the precision of the method of measurement we apply. Discontinuous, also called *discrete* or *meristic*) can assume only certain fixed values Number of fish in a trawl tow, number of young in a nest, or number of teeth in a skull are examples of discontinuous variables. None of these are reasonably expressed—at least in the original data—as "1.6," for example; the data record has to be stated in whole units.

The distinction between continuous and discontinuous variables matters because we analyze or represent these two kinds of data in different ways. This distinction may not, however, be as clear as we wish. Often it happens that initially discontinuous data are subsequently expressed as continuous data. For example, in fisheries work, each trawl haul collects only discontinuous data [whole fish), but if we average the number of fish for 10 trawls, we get a number with meaningful decimals. On the other hand, some discontinuous variables are derived from continuous data. For instance, Classes 1–5 for hurricanes (depending on an ascending comparison of wind velocity and other properties) and the

Consider one example in which cages hold animals for an experiment. We assign numbers 1, 2, . . . , n to the cages merely as identification labels (a nominal value]. As it turns out, cages nearer the laboratory window are exposed to different light and temperature regimes than those farther from the window. Suddenly we are in a position to transform the data types, because we can use the labels as a proxy measurement for the environmental gradient.

Another example also shows that the type of data may actually be defined by the question we ask.[2] Suppose that for personal reasons, the person who puts numbers on a soccer team's uniforms assigns low numbers to the first-year players. The dispenser of uniforms argues that these numbers are only nominal labels, devoid of quantitative meaning, and in any case, they were assigned at random. The more experienced players complain, saying that the numbers **1–11** do have a meaning, since traditionally these are worn by the starting team (a rank variable of sorts], and that the numbers traditionally refer to position in the field (1 is by custom a goalkeeper's number, e.g.). The players conceive the numbers both as nominal and as ordinal labels. Moreover, the players argue that the assignment of numbers seems unlikely to be random. To test this, a statistician is consulted to settle the issue. The statistician proceeds to treat the uniform numbers as if they were measurements, and does a few calculations to test whether such assignment of low numbers to the first-year players is likely to be due to chance alone. Each of the different viewpoints is appropriately classifying the same data in different ways. Classification of data types thus depends on purpose, rather than being an inherent property.

.2.2 Accuracy and Precision

For data to be as good as possible, they have to be accurate and precise. These two terms are easy to confuse. To distinguish them, let us say that in making a measurement, we want *data that are as close to the actual value as possible*. This is our requirement for *accuracy*. We would also prefer that if we were to repeat our data collection procedure the *repeated values would be as close to each other aspossible*. This is our need for *precision*.

Another way to describe these ideas is to say that a measurement has high accuracy if it contains relatively small systematic variation. It has high precision if it contains relatively small random variation.

Precision will lead to accuracy unless there is a bias in the way we do a measurement. For example, a balance could be precise but miscalibrated. In that case, we would get weights that are repeatable (precise), but inaccurate. On the other hand, the balance could be imprecise in determining weights. In this case occasionally the balance would provide weights that are accurate, but it will not do so reliably, for at the next measurement the weight will be different. Without precision we therefore cannot obtain accuracy. Precision has to do with the quality and resolution of the devices or methods with which we measure variables;

2. Updated from Lord (1953).

accuracy, with how we calibrate the devices or methods once we have obtained precision.

Most measurements we make are going to be approximations. We can indicate the degree of precision (not accuracy) of our measurement by the last digit of the values we report. The implied limit to the precision of our measurement is one digit beyond the last reported digit. If we record a temperature of 4.22 °C, we are suggesting that the value fell somewhere between 4.215 °C and 4.225 °C. If we report that the roundednumber of fish per trawl was 36,000,[3] we imply that the value fell between 35,500 and 36,500. In general, within any one set of measurements, the more nonzero digits, the more precision is implied. Realistic limits to reported precision must be set by the investigator; most researchers report too many digits as significant.

Sokal and Rohlf (1995) suggest an easy rule for quickly deciding on the number of significant figures to be recorded: it is helpful to have between 30 and 300 unit steps from the largest to the smallest measurements to be done. The number of significant digits initially planned can be too low or too high. An example of too few digits is a measurement of length of shells in a series of specimens that range from 4 to 8 mm. Measurement to the nearest millimeter gives only four unit steps. It would be advisable to carry out the measurement with an instrument that provides an additional digit. With a range of length of 4.1–8.2, the new measurements would give 41 unit steps, a more than adequate series. An example of too many digits is to record height of plants that range from 26.6 to 173.2 cm to the nearest 0.1 cm. The data would generate 1,466 unit steps, which is unnecessarily many. Measurement to the nearest centimeter would furnish 146 steps, an adequate number.

There are different procedures to round off figures to report actual precision. I prefer to round upward if the last digit is *greater* than 5. A few numbers will end in 5; to prevent upward or downward biases in rounding in long series of data, rounding of these numbers ending in 5 should be up if the number located before the 5 is odd, and down if it is even. Current software programs may use other alternatives.

2.3 Frequency Distributions

Throughout scientific work we deal with multiple measurements; we can say little about a single datum. A convenient way to gather multiple measurements together is to create a frequency distribution. Frequency distributions present the data in a way that capsulizes much useful information. This device groups data together into classes and provides a way for us to see how frequent (hence the name) each class is.

For example, let's say the values shown on the top of figure 2.1 are data obtained during a study. These values can be grouped into classes, usually referred to as ***bins,*** and then the number of items in each bin (3.3–3.4, 3.4–3.5, etc.) is plotted as in the bottom left panel of figure 2.1. If the frequency plot is irregular and saw-toothed, as is the case in the figure, it is hard to see the emerging pattern of the frequency distribution. We can

3. Actually, it would be clearer to express the rounded number as 3.60×10^4

Original Measurements													
3.5	3.8	3.6	4.3	3.5	4.3	3.6	3.3	4.3	3.9	4.3	3.8	3.7	4.4	4.1
4.4	3.9	4.4	3.8	4.7	3.6	3.7	4.1	4.4	4.5	3.6	3.8	3.8	4.2	3.9

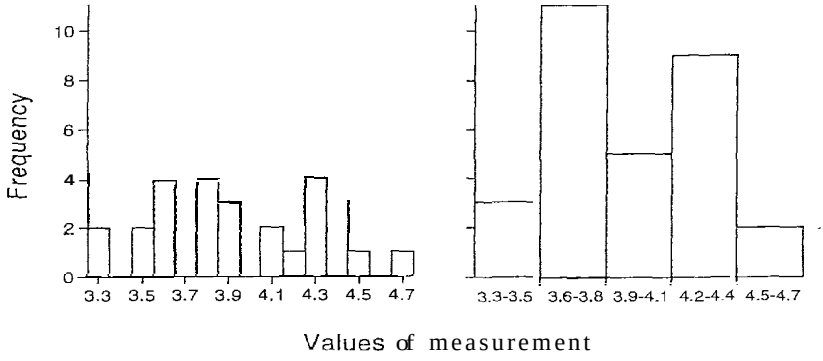


Fig. 2.1 Construction of a frequency distribution. Top: Values for a variable. Bottom left: A first attempt at a histogram. Ticks denote the label of every other bin. Bottom right: Same data treatment, with larger bins (3.3–3.5, 3.6–3.8, etc.).

regroup the measurements into somewhat larger bins **(3.3–3.5, 3.6–3.8,** etc.), as in the bottom right panel. This somewhat larger bin size better reveals the bimodal pattern of the data. Selection of a suitable bin size can convey not only the pattern of the data, but also a fair idea of the smallest significant interval for the variable on the x axis.

Shape of the frequency distribution often depends on sample size. Compare the four frequency distributions of figure 2.2. When the number of measurements is relatively low ($n = 25$; fig. 2.2, top) the distribution appears relatively featureless. It is only as sample number increases that we can place more and more numbers in the same category along the x axis, and the underlying humped shape of the distribution becomes more and more apparent. In many studies we have to deal with sample sizes of 25 or fewer. Discerning the pattern of the distributions with such a relatively low number of observations may be difficult.

Nominal as well as measurement data can be shown as frequencies. For example, for nominal data such as numbers of species of fish caught in trawl hauls, we could make a graph of the number of trawls (i.e., the frequency) in which $0, 1, 2, \ldots, n$ fish species were caught.

The distribution of data of figure 2.2 is fairly symmetrical about its mean; this is not always the case. Many sets of data show considerable skewness. Data with the same number of observations and value of the mean may be quite differently distributed; the upper two distributions in figure 2.3 are fairly symmetrical, but differ in that one is far more variable than the other. The third distribution in figure 2.3 is skewed to the right. If we simply computed the mean, standard deviation (see section 2.4), and so on, for such skewed data, without plotting the frequency distribution, we would have missed some of its major features. Plotting frequency distributions is one of the first things that we ought to do as soon as data become available.

In addition to revealing the central tendency, the scatter of the data around the mean, and whether or not the data are asymmetrical, a frequency distribution may show that there are multiple peaks. In the case

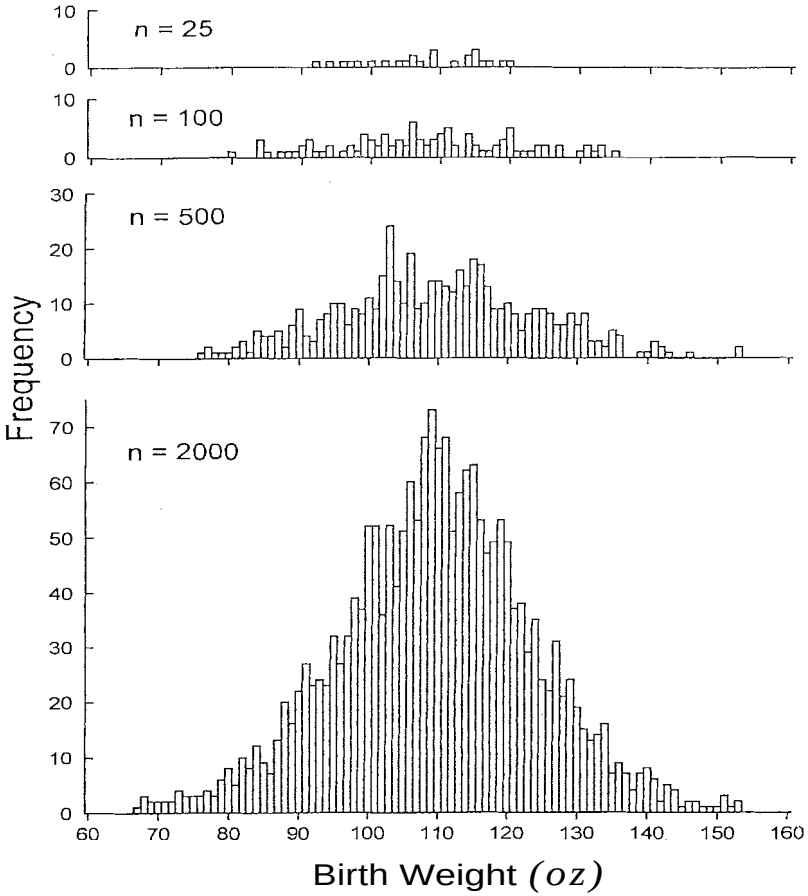


Fig. 2.2 The shapes of frequency distributions of samples depend on the number of observations (n) included. Histograms shows measurements of weights of babies at birth. From ***Biometry,*** 3rd ed., by Sokal and Rohlf © 1995 by W. H. Freeman and Company. Used with permission.

of figure 2.1, for example, we find two modes to the distribution. The bimodal pattern is clearer after we pool size classes along the horizontal axis to eliminate the jagged saw-tooth pattern created by finer subdivision of the variable plotted along the x axis. The bimodality suggests that we might be dealing with two different populations. This is yet another reason why plotting of frequencies is a desirable practice.

Before we learn how to check whether our data show that we have sampled more than one population, we need to acquaint ourselves with some statistics that describe the frequency distributions we have obtained.

2.4 Descriptive Statistics

To describe frequency distributions such as those in figure 2.2, we need to assess the central tendency of the distribution, as well as some indica-

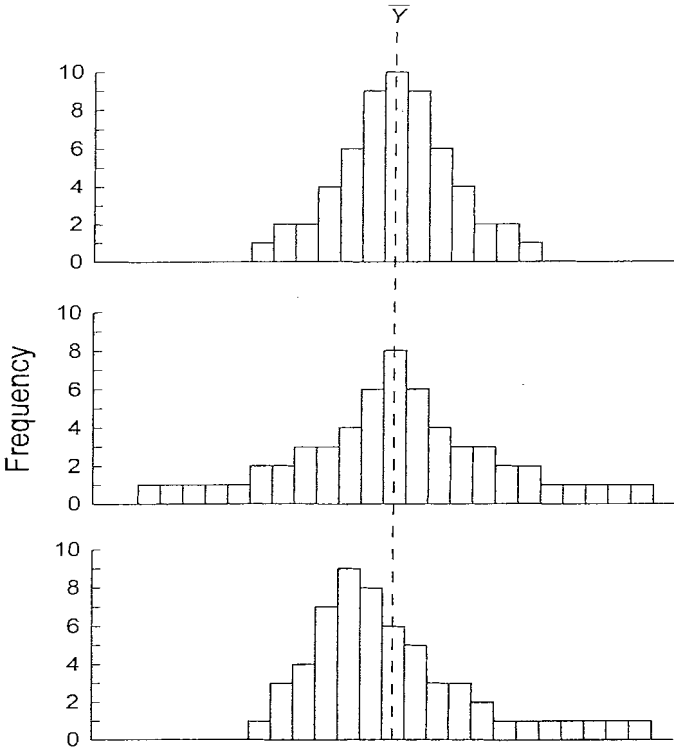


Fig. 2.3 Frequency distributions with same mean ($\bar{Y}$), but different shapes.

tion of how spread out the left and right tails of the distribution may be. There are various ways to quantify central location and spread, each useful for different purposes.

The *mean* is what most people would call the average, and is the most common statistic that describes the tendency to a central location. The mean is intuitively attractive and is appropriate with distributions that are symmetrical and bell-shaped. A disadvantage of the arithmetic mean is that it is markedly affected by extreme values. The *geometric mean* (the antilogarithm of the mean of the logarithms of the measured values) may be useful with measurements whose frequency distributions are skewed (see below). The *mode* is a quick way to judge the most frequent values in a data set, but is seldom used in analyses of scientific data. The *median*, in contrast, is widely used in quantitative analyses, in particular when data fall into frequency distributions that are highly skewed. Most statistical analyses are designed to deal with means, but statistics designed for analysis of medians are increasing (Sokal and Rohlf 1995). It is inherent in the definitions of the various expressions of central tendency that geometric means are less affected by extreme values (outliers) than arithmetic means, while mode and median are unaffected by outliers. The arithmetic mean, median, and mode are numerically the same in symmetrical frequency distri-

Definitions and Formulas for Some Basic Statistics

Central location

Arithmetic mean:[1]

$$\bar{y} = \sum Y_i / n$$

Geometric mean:

$$GM_y = \text{antilog } 1/n \sum \log Y$$

Mode: the mostfrequent category in a frequency distribution

Median: value that is at 50% of n and so divides a distribution into equal portions in data that are ordered numerically—the $(n+1)/2$nd observation

Spread

Range: difference between the smallest and largest values in a sample

Standard deviation:

$$s = \sqrt{(Y_i - \bar{Y})^2 / n - 1)}$$

Coefficient of variation:

$$CV = (s / y) \times 100$$

Standard error of the mean:

$$se_y = s / \sqrt{n}$$

Standard error of the median:

$$se, = (1.2533) \times se,$$

1. Y are the i observations made; the symbol $\sum$ indicates that i values of Y are summed.

butions with single modes. In the asymmetrical distribution shown in figure 2.4, the mode is farthest away from the long shoulder or tail of the distribution, followed by the median, and the arithmetic mean is closest. The geometric mean falls close to the position of the median.

The range is the simplest measure of spread. It usefully shows the bracket of upper and lower values. It does not, however, tell us much about the relative distribution of values in our data. The range is also affected by outliers.

The standard *deviation* is a more desirable measure of spread because it weights each value in a sample by its distance from the mean of the

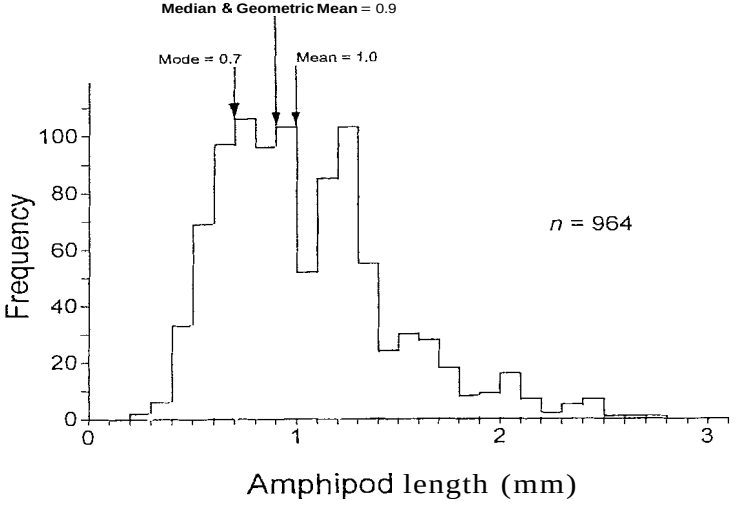
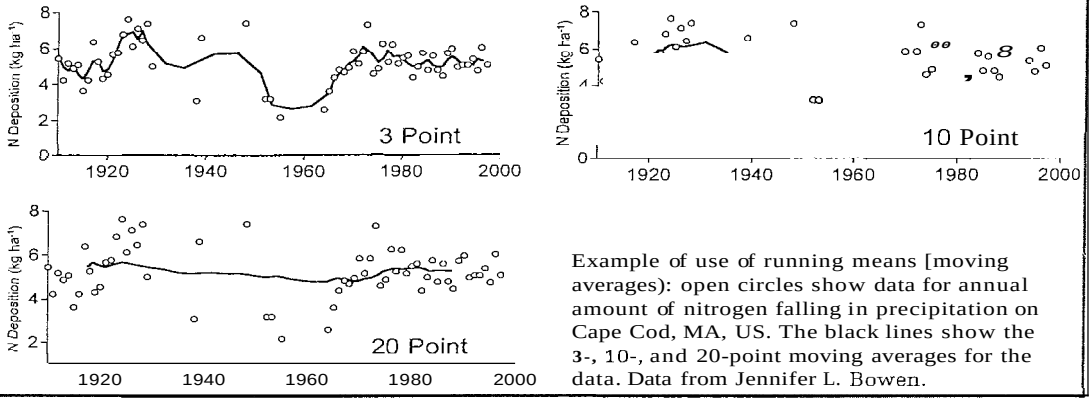


Fig. 2.4 Three measures of central tendency in a skewed frequency distribution; n is the number of observations.

Running Means

If we have a data set collected at intervals of, say, hours, days, centimeters, and so on, we might wish to see if there are trends across the various intervals. Often the variation from one measurement to the next is large enough that it is difficult to discern trends at larger or longer intervals. A simple way to make trends more evident is to use running means. Running means (also called moving averages) are calculated as a series of means from, say, sets of three adjoining values of X; the calculation of the mean is repeated, but for each successive mean we move the set of three values one X value ahead. For example, the first of the running means is $x_1 = (X_1 + X_2 + X_3)/3$, the second is $x_2 = (X_2 + X_3 + X_4)/3$, and so on.

The figure shows a data set, with trend lines calculated as running means of annual deposition of nitrogen in precipitation. Notice that year-to-year variation is smoothed out with 10 point moving averages, and even more so with 20 point moving averages. The latter reveal the longer scale (multidecadal) trends. This procedure gives equal weight to all the X used per mean; for some purposes it might be better to give more weight to the more recent X, for example, in studies of contaminants that could decay through time intervals shorter than the moving average intervals. Berthouex and Brown (1994) and Tukey (1977) discuss this in more detail. The need to smooth out, or "filter," variation at different time or spatial scales has prompted development of the field of statistics referred to as time series analysis. Chatfield (1984) and Brillinger (1981), among many others, provide an introduction to this field.



Example of use of running means [moving averages): open circles show data for annual amount of nitrogen falling in precipitation on Cape Cod, MA, US. The black lines show the 3-, 10-, and 20-point moving averages for the data. Data from Jennifer L. Bowen.

distribution. Let us consider how we might describe variation within a set of data. Suppose we have collected a set of data, and the values are 2, 5, 11, 20, and 22. The mean of the set is 12. We cannot just calculate the average difference between each value and the mean, because the sum of the differences is necessarily zero. We also need to give more importance to large variation (which may be the effect of a source of variation we might want to study) than to small deviations. The solution is to sum the squares of the differences between each observation and the mean. This simultaneously eliminates the sign of the deviations and emphasizes the larger deviations. For the data we have, the differences (or for statisticians, ***deviations***) are −10, −7, −1, +8, and +10. After we square and sum the deviations, we have a value of 314, and dividing this total by the number of observations yields the mean of squared deviations, which for our data is 62.8. To get the values back into the same scale at which we

Alternative Calculation for Variance

Most statistics texts written before the revolution in microcomputers and software show how to calculate the measure of variation in a different way.

$$s = \sqrt{\sum Y^2 - (\sum Y)^2 / n/(n-1)}$$

In the era of mechanical calculators in which I learned how to do science, it was cumbersome to calculate s as given in the box of definitions. Instead, we took the deviations of all measurements as if they extended from zero—this is another way to say we took the actual values—and then squared the values. The sum of squared deviations from zero for our data is 1034. If there had been no deviations, the sum would have been (12 + 12 + 12 + 12 + 12), or using the sum of all values in our data, 3600/5, which is equal to 720. The difference between 1034 and 720 is equal to 314, and is an estimate of variation in the data set. Note that it is the same value as we obtained earlier. Therefore, in general, the mean of squared deviations is

sum of (data)^2 – [(sum of data)^2 / number of data].

This expression is almost the same as what we usually see as the computational formula for s^2, the variance. The variance does weigh the relative magnitudes of deviations in data sets, and is the usual way we describe variation. To undo the effect of squaring, we took the square root of the variance, which provided s, the standard deviation of individual observations within our group of data. With the advent of the computer age, we do not have to worry about computational difficulty, so we use the first version of the formula for s.

did the measurements, we take the square root of the mean of squared deviations, and get the mean deviation.

If you compare the expression for the variance discussed so far with the version given in the box, you will note one discrepancy, which is worth a bit more explanation. In research we take *samples* as a way to obtain statistics (e.g., of X or *s*), which are estimates of the parameters (μ or σ) of a *population* from which the sample was drawn. In the case of the mean, a random sample provides a fair estimate of the population mean: if we have chosen our data by chance, the sample is equally likely to contain larger and smaller values, and the array is representative of the population; hence, we can accept the sample mean as an unbiased estimate of the population mean.

In the case of the variance, however, the sample provides a biased estimate of σ. The variation among the measurements taken refers, of course, to the set of measurements in our sample, so it necessarily is smaller than the variation among values in the population from which the sample was chosen. The estimated variation is therefore corrected for the underestimate of σ. This is best done by expressing variation in terms of *degrees of freedom* (*df*). Few of us really understand *df*, so we have to simply trust the mathematicians. For our present purpose, consider that if we know the mean of the values in a sample, and we know all but one of the values ($n-1$), we can compute the last value. So, if we calculate the mean in the process of calculating *s*, we in a way "use up" one value, that is, a degree of freedom. It turns out that if we divide the sum of squared deviations by ($n-1$) instead of n, we correct for the bias in estimation of σ from samples. Now we have arrived at the expression given in the box and in statistics textbooks.

If we are dealing with data that follow a symmetrical, bell-shaped *normal frequency distribution*, a span of one standard deviation above plus one standard deviation below the mean captures about 68% of the values. A span of $2s$ above and below the mean will include about 95%

of the values, while 3s comprises about 99% of the values.[4] If we want to get an idea of the relative size of the variation represented by the standard deviation, we can divide it by the mean and multiply by 100 to obtain the coefficient of *variation*. The coefficient of variation is especially useful for comparing variation of means that differ considerably in magnitude.

We can sample apopulation more than once, and take multiple measurements on each occasion. We can then calculate the arithmetic mean for each sample. Those means themselves have a hequency distribution, usually with a smaller variability than that for the individual measurements. We can calculate the standard deviation of the means to quantify how variable they are. This new statistic is called the standard *error* of the mean, $se_{\bar{y}}$, an important statistic that enables us to compare means. The $se_{\bar{y}}$ is the measure of variation we want in most instances, since in practice we most often intend to compare means rather than individual observations.

2.5 Distributions and Transformations of Data

Use of the mean, standard deviation, or standard error presupposes that we are dealing with "normally" distributed data. "Normal" is a misnomer we are stuck with, since, as discussed below, many data sets fail to follow the so-called norm. Normal distributions occur in situations where (1) many factors affect the values of the variables of interest; (2) the many factors are largely independent of each other, so the effects of the factors on the variable are additive; and (3) the factors make approximately equal contributions to the variation evidenced in the variable.

It is a wise precaution to check the frequency distribution of data before doing any calculations. For most purposes, it is enough to see arough bell shape to the distribution. Recall that for the number of observations we usually have, we should not expect a perfect bell-shaped distribution (fig. 2.2, bottom). Tests are available to ascertain whether we have a normal distribution (see, e.g., Sokal and Rohlf 1995, chap. 6). One easy method is to plot the cumulative hequencies on a probability plot; in such plots, normal distributions appear as straight lines (fig. 2.5, top). Note that the obviously nonnormal distributions of figure 2.5 (middle left and bottom left) show small but systematic deviations from the straight lines (middle right and bottom right).

Often we have to deal with data that are not normally distributed. If we wish to estimate the central tendency and variation in our data, our best option is to recast the data in such a way that the transformed data become normally distributed (section 3.6). Some might feel uncomfortable about such apparent sleight of hand. Recall, however, that all scales are arbitrary, and that the nature of data depends on the purpose of the researcher. We might be familiar with transformed scales without knowing it; pH units are expressed on a log scale, for example. Here we are merely recasting values in ways that fit our purpose. Useful arithmetical

4. These statements are shorthand for the idea that if we were to repeat the sampling many times, and we were to recalculate the standard deviation again, the values would have a 95 or 99% probability of falling within the range of values of 2 or 3 standard deviations.

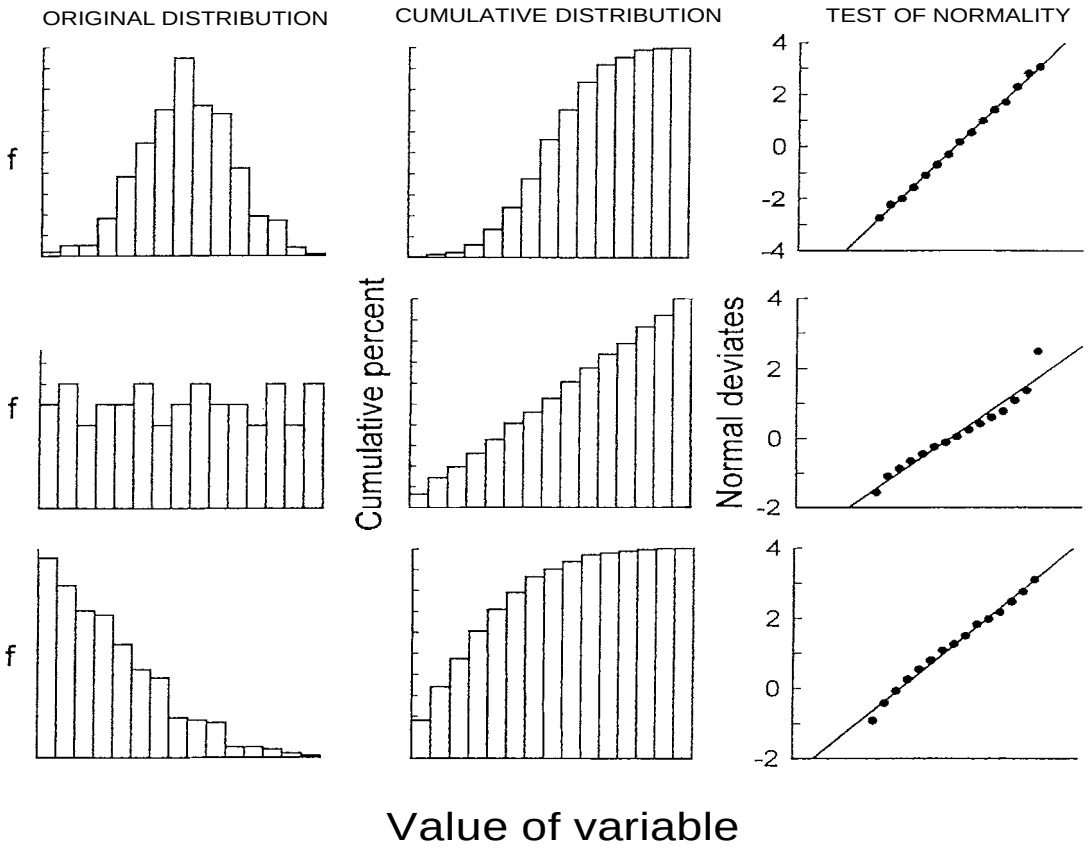


Fig. 2.5 Graphical check for normality of three different data sets.

operations that lead to normality of transformed data are the logarithmic, square root, and inverse sine transformations.

The *logarithmic transformation* is the most common of all transformations. Log transformations are especially apt in the rather common case of distributions that are strongly skewed to the right (fig. 2.6, top), that is, where there are more frequent observations at low values, or zero may be the most frequent observation. There is some disagreement among statisticians as to what to do with values of zero; some prefer transformations such as $Y' = \log(Y + 1)$, but others suggest omitting zero values. We are therefore free to choose.

Log transformations are possible with any of the types of logs. We use $\log_{,,}$ but $\log_2$ can also be useful. Transformation to $\log_2$ allows us to express the frequency in bins that double at each interval. Doublings per interval is an intuitively appealing way to display data of this sort.[5]

5. This type of transformation has received much attention in the ecological literature and has acquired a glossary all its own. The distributions are described as *log-normal*. The bins have been called ***octaves,*** after a fancied parallel to mu-

Distributions Other Than Normal

In the text, we casually refer to symmetrical, bell-shaped frequency distributions as the normal distribution. The normal distribution is just one among many random frequency distributions that describe data collected under various conditions. Other distributions include the following:

(Positive) Binomial. This is the distribution of events that can occur, or not, in samples of a definite size taken from a very large population. Example: number of males in families of a given size. The number of boys in 100 families of 3 children is 12, 36, 38, and 14, for 0, 1, 2, and 3 boys. The variance of a binomial is always less than the mean.

Poisson. This is the distribution (named after an eighteenth-century mathematician) of large samples of events in which one of the alternatives is much more frequent than the other, and the frequency of occurrences is constant. The mean equals the variance in this distribution. Example: lumber of flaws in parts for Mercedes Benz automobiles, or number of Prussian soldiers kicked to death by horses. The chance of flaws or deaths is rare, and cases of flaws or deaths are more or less unconnected to one anothers' occurrence.

Hypergeometric. This is the distribution of events sampled from a finite population without replacement Example frequency ot marked fish collected from a population Into which we released a given number of marked fish

There are many sampling situations in which distributions of data are far from random. The commonest outcome of sampling surveys is to find that data depart from randomness, and are clumped. Clumped distributions have an excess of observations at a tail of the distribution (we have called these skewed distributions, e.g., fig. 2.4). For such cases, different distributions can be used, as follows.

Negative binomial. This is similar to Poisson, but for the more common case in which probability of occurrence is not the same. For example, if the Prussian soldiers counted were to include cavalry and infantry, the risk would differ systematically with different exposure to horses. In this distribution the mean is always much smaller than the variance. The distribution was discovered by a certain de Montmort about 1700, and the name comes from mathematical details of little interest to the rest of us.

Logarithmic. This occurs in skewed data distributions with some relatively large values on the right tail of the distribution. Log transformations convert these to near-normal distributions (see section 3.6).

Square root transformations tend to convert data taken as counts (insects per leaf, worms per sample of soil, nests per tree, e.g.) to normal distributions (fig. *2.6,* middle). Such data may be Poisson rather than normally distributed, such that the magnitude of the mean is related to that of the variance. A square root transformation usually makes the variance independent of the mean. If there are zeroes in the data, it is necessary to use a slightly different transformation, for example, $\sqrt{Y + 0.5}$. Square root transformations have effects similar to, but less powerful than, those of log transformations.

Inverse sine transformations are useful to normalize percentage or proportional data (fig. *2.6,* bottom). This operation makes the mean independent of the variance for percentage data, which are characteristically binornial in nature. Inverse sine transformations of percentages or proportions make variances independent of means. Percentages are also

---

sical octaves, in which each octave corresponds to a doubling in the frequency of vibration of a note. Actually, musical octaves were derived from the eight notes of a musical scale. "Doublings" or, as Williams (1964) proposed, "doublets" might have been a more descriptive term

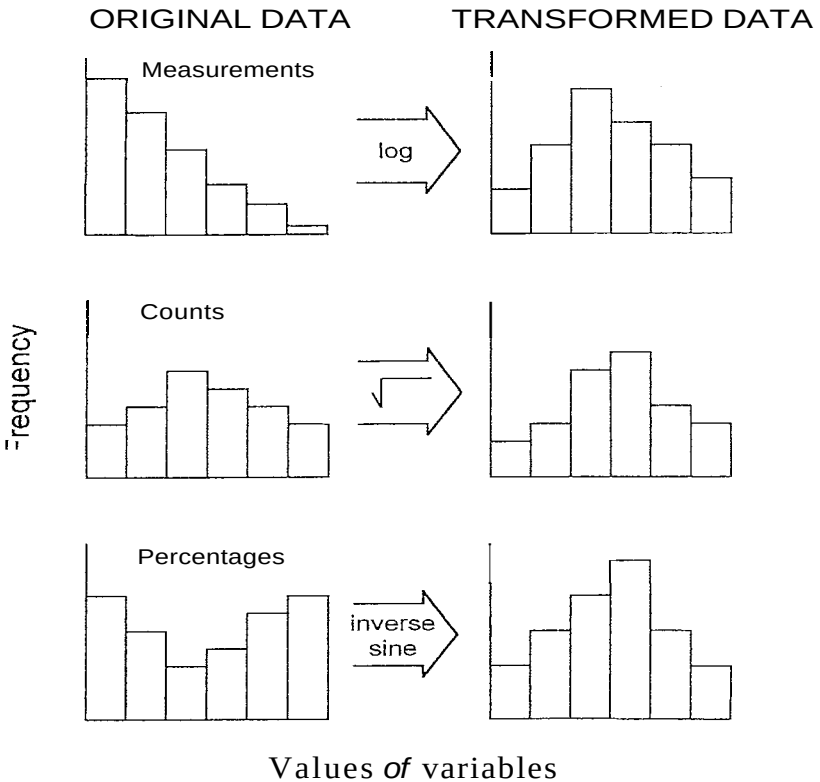


Fig. 2.6 Nonnormal frequency distributions, and transformations (top, logarithmic; middle, square root; bottom, inverse sine) to convert data to normal distributions.

curtailed at the tails of the distributions, unlike normal distributions. The inverse sine transformation expands the range near 0 and 100, thus making the distribution nearer to normal.

Box–Cox transformations are useful if we have no a priori reason to select any other transformation that provides the closest approximation to normality in the recast values. The calculation is best done on a computer. For a quick rule of thumb (Sokal and Rohlf 1995) try a series of transformations, $1/\sqrt{Y}$, $\sqrt{Y}$, $\ln Y$, $1/Y$, for samples skewed to the right, and the series of transformations Y^2, Y^3, . . . , for samples skewed to the left.

2.6 Tests of Hypotheses

We can now return to the question of how to check whether our data belong to one population or to more than one. Suppose we are studying a variable (say, oxygen content of water), make many observations at two sites, and produce a frequency distribution. The frequency distribution

Let the Data Speak First

Once we have a data set, it is a good idea to really try to understand the data before plunging them into statistical tests now temptingly easy to do using software packages. We can let the data speak to us by means of a few manipulations.

A plot of frequency distributions (or box plots; see fig. 9.5) will let us perceive whether there is a central tendency in the data, if the data are skewed, what the left and right tails of the distributions are like, whether extreme values or outliers are present, or if there are apparent differences among data from different treatments or samples. If we have data collected across a gradient (such as time, space, or dosage), a plot of the data versus the gradient will reveal trends or identify outliers that could be either errors or telling extremes.

A plot of means versus variances can tell us whether variation changes with size of the mean. This is useful for several reasons, one being that this could tell us if the data meet assumptions of statistical tests to be applied.

Fairly simple data manipulations early on will provide us with a clear sense of what our data are really like, as well as suggest how we might test the data, and what further manipulations, such as transformations, might be needed for data analysis. Chapters 3 and 4 make evident why these initial data manipulations might be worthwhile.

is shown diagrammatically in figure 2.7; the curves are continuous and rounded simply because we intend to show what would happen if we were to make many, many observations. The shape is in contrast to the step-shaped distributions characteristic of real samples, in which we inevitably have a limited number of observations.

We are interested in ascertaining whether the values of oxygen content of water at one site differ from those measured at the other site. How likely is it that the mean concentrations at the two sites are the same? The usual approach is to ask what is referred to as the null *hypothesis*, that is, the hypothesis that there is *no* effect, in our case that the population means from the two sites are *not* different.

Statistical tests allow us to calculate how likely it is that the question (or hypothesis) that we are testing is true. By convention, we usually test whether there are no differences between the data sets we are examining. The tests can, of course, yield a continuous range of probabilities, from highly likely to rather unlikely that it is true that there are no differences between our data sets. How do we decide that something in this continuum is meaningful? We need some clearer benchmarks, and hence researchers have decided, arbitrarily, on "significance" levels. These are usually given as 1 in 20 (probability, or $P = 0.05$), or 1 in 100 ($P = 0.01$) that the differences are larger than expected due to chance. These levels are spoken of as "significant" and "highly significant" and are often symbolized by adding "*" or "**" following the value of the testing criterion calculated by the test used.

Statisticians use the term "significant" in a way that should not be confused with our usual notion of the word. Results of research might be "statistically significant" but not necessarily of profound consequence or interesting. For research purposes, "statistically significant" means only that the probability of a difference as large as we found by effects of chance factors alone is less than one of the predetermined thresholds (0.05 or 0.01).

Earlier we made the point that tests of hypotheses are the hallmark of empirical science, but such tests are not as straightforward as they might

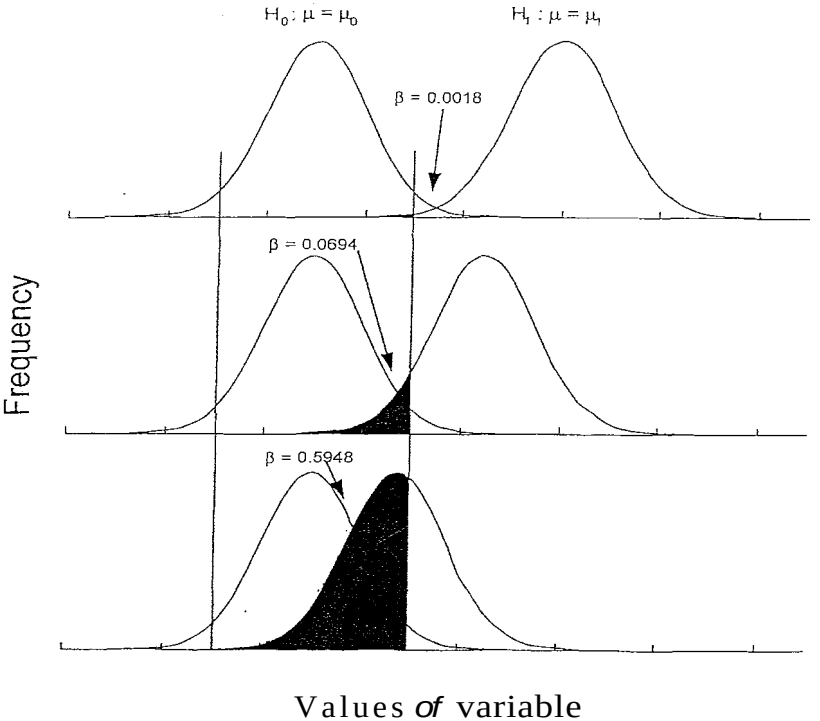


Fig. 2.7 Test of hypotheses. H_0 and H_1 are null and alternative hypotheses, respectively; β (black area) is the probability of committing a Type II error (accepting an untrue hypothesis). Power $(\beta - 1)$ diminishes as the means approach each other.

seem. Consider the null hypothesis, which we can refer to as H_0, which we wish to test (fig. 2.7). In its simplest form, the hypothesis can be true or false, and our test can accept it or reject it. If it is in fact true but our test rejects it, we make what we refer to as *Type* I *error*. If our test accepts the hypothesis but it is false, we make a *Type* II *error*. How do we deal with these two undesirable outcomes?

By convention, we test whether the likelihood of the difference being significant is either 1 in 20 (the probability level, P or a = 0.05) or 1 in 100 ($\alpha = 0.01$). These values, as already noted, are called the *significance levels* of the tests: they are the probability that a result arose by chance alone. If we conclude that a result is significant at the probability level of 0.05, we are saying that either the result is as we claim, or a coincidence arose with odds of 1 in 20. That the possibility of coincidence is real is shown by a confession of a distinguished agricultural statistician, who once found a quite significant difference at $a = 0.001$, only to learn later that an assistant had forgotten to apply the treatments.

If in a test of a hypothesis we commit a Type I error, we are giving up information that is true. To reduce the possibility of committing such an error, we can of course be more stringent in our test, that is, increase the level of significance at which we run the test. Unfortunately, there are limits to this stringency. If we demand less uncertainty (i.e., move the

It ain't as much the things we don't know that gets us into trouble. It's the things we know ain't so.

Artemus Ward

vertical line to the right in fig. 2.7), we increase the probability of committing a Type 11 error (shown by the black area of fig. 2.7). That is likely to be a worse outcome, since then we would be accepting as true something that is false. It is generally preferable to err on the side of ignorance rather than to accept false knowledge.

Because for most purposes Type 11 errors are more egregious than Type I errors,[6] statisticians suggest that statistical tests be run at the 0.05 or 0.01 levels of probability (the level of Type I error we are willing to commit), rather than at higher α levels. The level or probability of a Type II error is denoted as β. Note in figure 2.7 that β increases as two means come closer to each other. This says that the probability of committing a Type 11 error increases. The *power* of a statistical test is $(1 - \beta)$ and refers to the probability of rejecting the null hypothesis when it is false. Note how the power of the test diminishes in figure 2.7 as the two means approach each other.

Discussion of levels of significance brings up a common problem, that of multiple comparisons. In large studies it is often possible to test many comparisons. For example, in surveys of cancer rates, one might be tempted to compare incidence of cancers of the skin, ovary, liver, and so on, in many types of subpopulations (women under 40 vs. women over 40 years of age, males who exercise daily vs. those who exercise weekly vs. males who do not exercise, women who bathe in freshwater lakes vs. those who swim only in the ocean vs. those who do both, etc.). Where we run such multiple comparisons, we will inevitably find that some of the comparisons turn out to be "statistically significant," even though the differences might be due to chance alone. The tests we use in such cases all have a level of probability, say, 1 out of 20; this means that we *expect* that in 1 out the 20 tests we are performing we *will,* erroneously, find a "significant" difference. And we will.

Another problem with multiple tests is that in any given study there are only so many degrees of freedom. Each degree of freedom "entitles" the researcher to make one comparison. The number of comparisons done, however, should not exceed the number of degrees of freedom. If they do, this means that we are not really testing the differences at the significance levels we think we are, but rather at lower levels of probability.

Results that are not statistically significant *do* not *prove* that the data we are comparing *are similar*. Scientific tests of the kind we are discussing are not designed to prove that something is true, because there is a real possibility that we might incur a Type I1 error by seeking to prove something. Thus, tests characteristic of empirical science differ from the unambiguous "proofs" possible within tautologies such as geometry and mathematics.

6. Harvey Motulsky pointed out to me that generalities such as this might prevent us from being aware of the consequences of the two kinds of error. He suggests some cases where Type I errors are trivial and Type II errors bad (e.g., in screening compounds for new drugs, a Type I error means one just does one more test, but a Type II error might mean missing a new drug). In other cases Type I errors may be fatal and Type II errors trivial [releasing a new drug for a disease already treated well by an existing drug). Those are exceptional cases; what is important is to understand the two kinds of error.