

# Tutorial 1: Review of Basic Statistics

While we assume that readers will have had at least one prior course in statistics, it may be helpful for some to have a review of some basic concepts, if only to familiarize themselves with the notation that will be used in the current book.

To begin our introduction to some of the basic ideas of statistics we will utilize a data set of scores on the Beck Depression Inventory (BDI) for 103 adults (Smith, Meyers, & Delaney, 1998). It is often useful to begin one's examination of a set of data by looking at the distribution of scores. A graph of a frequency distribution indicates the frequency of each of the scores in the data set by the height of the bar plotted over that score. As shown in Figure T.1, the BDI scores range from 0 through 43, with nearly all scores in this range occurring at least once.

Such a distribution of scores would typically be characterized by statistics describing its central tendency and variability, and perhaps supplemented by other measures reflecting the shape of the distribution. While we will focus on the mean, the mode and median are two other useful measures of central tendency which can be obtained basically by counting. The mode, the most frequently occurring score, in this sample is 9, as indicated by the tallest bar in Figure T.1. The median is defined roughly as the middle score, or the score at the 50th percentile. Although there are several slightly different ways of computing the median as well as the other percentiles (e.g., depending upon whether interpolation between, or averaging of, adjacent values is employed; see SAS Institute, 1990, pp. 625–626), it is commonly defined as the minimal score having at least 50% of the scores less than or equal to it. In this sample of 103 scores, the middle score is the score with rank order 52, or in general if  $n$  is the number of scores, the desired rank is  $(n + 1)/2$ . In the current data, there are 52 scores less than or equal to 13 and so 13 may be taken to be the median.

The mean is of course simply the sum divided by the total number of scores,  $n$ . To define this and other statistics, we will use the standard notation of  $Y_i$  to denote the score of the  $i$ th individual on the dependent variable, and will use  $\sum_{i=1}^n$  to denote the summation operator. Thus, denoting the sample mean as  $\bar{Y}$  (read “Y bar”), the verbal definition of:

sample mean equals sum of scores divided by number of scores in sample  
becomes in symbols

$$\bar{Y} = \frac{\sum_{i=1}^n Y_i}{n} = \frac{Y_1 + Y_2 + Y_3 + \cdots + Y_n}{n}$$

For the BDI scores, the sum of the 103 scores is 1622 so the mean is approximately 15.748.

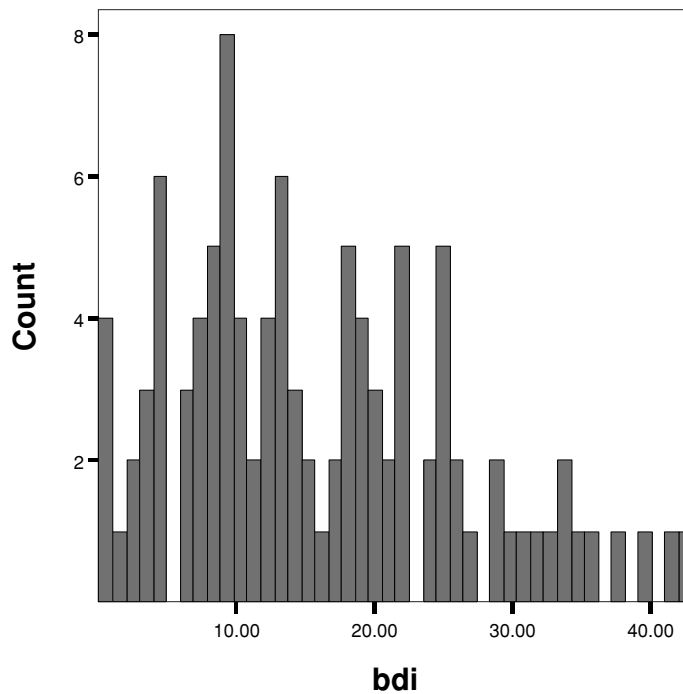


FIG. T.1. Histogram showing frequency of BDI scores.

Before we proceed to define measures of the variability and shape of distributions, it will be useful to make distinctions between different types of distributions of scores. The data collected in an experiment are typically regarded as merely a sample of a larger group of individuals in which the investigator is really interested. The larger group, whether hypothetical or real, constitutes the population. The distribution of the scores in hand is the sample distribution, and the distribution of the larger group of scores which is typically unobserved would be the population distribution.

Characteristics of samples are termed statistics, whereas characteristics of populations are called parameters. We will follow the convention of using Greek letters to denote population parameters and Roman letters to denote sample statistics. Letting the Greek letter mu,  $\mu$ , represent the population mean, we can define the population mean verbally as

population mean equals sum of scores divided by number of scores in population.

For a population having a finite number of scores,  $N_{pop}$ , then the population mean can be defined in a similar fashion to the sample mean

$$\mu = \frac{\sum_{i=1}^{N_{pop}} Y_i}{N_{pop}} = \frac{Y_1 + Y_2 + Y_3 + \cdots + Y_{N_{pop}}}{N_{pop}}$$

The mean is the most familiar example of the mathematical concept of expectation or expected value. The expected value of a random variable, such as  $Y$ , is defined as the sum over all possible values of the product of the individual values times the probability of each. Using “ $\mathcal{E}$ ” to denote expected value, the expected value of  $Y$  is then defined as:

$$\mathcal{E}(Y) = \sum Y_i Pr(Y_i)$$

where the sum is taken over the different values that  $Y$  can assume and the probabilities sum to 1 (we will say a bit more about probabilities and probability distributions below). In a small discrete population, like the numbers on the faces of a die, where all of the possible values are equally likely, the probabilities are just 1 over the number of possible values. Thus, the expected value or average of the values on the faces of a die would be

$$\mathcal{E}(Y) = \sum_{i=1}^{N_{pop}} Y_i Pr(Y_i) = \sum_{i=1}^{N_{pop}} Y_i \frac{1}{N_{pop}}$$

However, in mathematical theory, the population is typically assumed to be infinitely large, and while we will not make use of calculus to rigorously derive such infinite sums, we will nonetheless have occasion to refer to the expected value of such populations, and for example might denote the mean of the values in a normal distribution as the expected value of the scores in the population by writing

$$\mu = \mathcal{E}(Y)$$

In addition to sample and population distributions, there is a third type of distribution that is critically important in the statistical theory relied upon when one uses sample data to make inferences about a population, and that is a *sampling* distribution. A sampling distribution is the distribution of values of a statistic obtained through repeatedly drawing samples of the same size from a population. While one could with sufficient effort empirically approximate a sampling distribution, such distributions typically are unobserved, but nonetheless are critical to the process of drawing inferences. One of the useful functions of sampling distributions is to characterize how sample statistics relate to population parameters. Even though sampling distributions are distributions of values of a statistic instead of distributions of individual scores, it is meaningful to refer to characteristics like the mean of the sampling distribution. It can be shown that if one repeatedly draws simple random samples (where every score has an equal chance of inclusion) from a population, the expected value of the sample means of these samples will equal the population mean. Denoting the mean of the sampling distribution as  $\mu_{\bar{Y}}$ , we can write this principle in mathematical notation as follows:

$$\mu_{\bar{Y}} = \mathcal{E}(\bar{Y}) = \mu$$

This also serves to define one of the desirable properties of statistics. That is, a sample statistic is said to be an *unbiased estimator* of a population parameter when the expected value of the statistic (or, equivalently, the mean of the sampling distribution of the statistic) equals the value of the parameter.

To return after this lengthy digression to the discussion of how to characterize distributions, the most important characteristic of distributions besides their central tendency or location is their spread. It is sometimes said that variability is the most central idea in statistics. As a field, statistics is concerned with the quantification of uncertainty, and uncertainty is the result of variability. The simplest measure of spread or variability of a distribution of scores is the range, or the difference between the maximum and minimum scores. In the BDI data, this can be seen by inspection of the distribution to be 43. Because the range depends upon the two most extreme scores, it is not nearly as stable as other measures of variability which are affected by the distance of every score from the center of the distribution.

In the current text, major emphasis is given to the variance as a measure of variability. The sample variance as a measure of spread is denoted  $s^2$  and is defined as

$$s^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n - 1}$$

Similarly, the population variance is denoted by the square of the Greek letter sigma, i.e.  $\sigma^2$ , which in a finite population consisting of  $N_{pop}$  scores is defined as

$$\sigma^2 = \frac{\sum_{i=1}^{N_{pop}} (Y_i - \mu)^2}{N_{pop}}$$

Thus, the population variance is the average or expected value of the squared deviations from the population mean. The answer to the standard question of why does one have to subtract 1 in the formula for the sample variance is typically expressed in words to the effect of, “So that the sample variance will be an unbiased estimator of the population variance”, or in mathematical notation so that

$$\mathcal{E}(s^2) = \sigma^2$$

Intuitively, one way of thinking of this is that when you are working with a sample where you don’t know the population mean, the sample scores will not be exactly centered around the population mean. Instead, the sample mean will tend to be “pulled” away from the population mean in the direction of the scores in the sample. So, the scores in the sample will be a little closer on the average to the sample mean than to the population mean, and consequently the deviations from the sample mean will tend to be somewhat less than the deviations from the population mean. In the formula for the sample variance if one were to divide the numerator, which is based on deviations from the *sample* mean, by  $n$ , the result would be a bias in the direction of underestimating the population variance. Although we will not take the space to prove it here, it turns out that dividing by  $n - 1$  exactly compensates for this tendency to underestimate the deviations from the population mean.

Besides the range and the variance, a third measure of variability is the standard deviation. The standard deviation is simply the square root of the variance, and has the advantage over the variance of being expressed in the same units as the original scores. Thus, if one’s data consisted of the height of some adolescents measured in feet, the standard deviation would also be expressed in feet whereas the variance would be in square feet.

Having considered ways of measuring the *location* or central tendency, and the *spread* or variability of a distribution, at this point we are ready to consider briefly two ways of characterizing the *shape* of the distribution, namely, skewness and kurtosis. Skewness reflects the asymmetry of the distribution. Positive values of skewness indicate that the distribution has a long tail on the right side (toward the positive end of the real number line), negative values of skewness indicate a distribution with a long tail on the left (pointing toward the negative end of the real number line). A perfectly symmetrical distribution like the bell-shaped normal curve has 0 skewness. For the BDI example in Figure T.1, the central tendency of the distribution is in the lower part of the range of scores with the long tail to the right, and so the distribution is positively skewed.

Kurtosis is an indication of the “peakedness” or flatness of the distribution shape, and also reflects the tails of the distribution. Distributions that have a preponderance of the scores closely

clustered around a central value and more scores in the tails than a normal distribution are said to be leptokurtic or to have positive kurtosis. Distributions that are fairly evenly spread over the entire range are said to be platykurtic, or to have negative kurtosis. Measures of kurtosis are strongly influenced by extreme scores in either tail of the distribution. The standard again is typically taken to be the normal distribution. Thus, leptokurtic distributions are sometimes referred to as “heavy tailed”, i.e. the distribution has a greater proportion of extreme scores than does a normal distribution<sup>1</sup>.

Measures for skewness and kurtosis are not often presented in elementary texts, in part because they can be computationally messy, e.g. the most common index of skewness involves the sum of the third power of deviations from the mean and kurtosis involves the sum of the fourth power of these deviations. Fortunately, standard computer routines such as SPSS Window’s Frequencies routine compute skewness and kurtosis values<sup>2</sup>, as well as the information needed to carry out statistical tests of their significance. In practice, it is sufficient for most purposes to realize that (1) the common measures of shape are driven by the most extreme scores (because raising the largest deviations to the third or fourth power can result in numbers that are huge relative to the sum of many smaller deviations raised to the same power), (2) skewness indicates whether there are more extreme scores in the right tail than in the left (positive skewness) or whether the reverse is true (negative skewness), (3) kurtosis indicates whether there is a greater proportion of extreme scores in either tail than in a normal distribution (positive kurtosis) or whether there is a smaller proportion in the tails than in a normal distribution (negative kurtosis)<sup>3</sup>, and (4) as indicators of shape both skewness and kurtosis are “scale free”, that is, they do not reflect and are not affected by the variance. One intuitive rationale for this last point is that although multiplication of scores by a constant, as in going from feet to inches in a distribution of heights, changes the scale and hence the variability of the scores, it does not change the shape of the distribution in a fundamental way other than scaling. For the BDI data, the skewness is .629 and the kurtosis is -.243, indicating that the right tail of the distribution is longer than the left, but that the distribution has fewer extreme scores than a normal distribution with the same variance.

One way of representing the distribution of scores which has become popular since introduced by Tukey (1977) is the box plot. This display provides information about the central tendency, variability and shape of a distribution using measures that are less influenced by extreme scores than the mean or variance. Although exactly how box plots are drawn is not entirely standard, a box plot provides a graphical representation of a five-number summary of the data: the median, the 25th and 75th percentiles, and either the maximum and minimum scores, or high and low cutoff scores beyond which data might be considered outliers. A box plot for the BDI data is shown in Figure T.2. The line through the middle of the box is at the median, and the two ends of the box represent the 75th and 25th percentiles. The vertical lines coming out from the box are called whiskers, and usually are drawn to extend out to the farthest data point or to a maximum length of 1.5 times the length of the box. The length of the box, or the distance between the 75th and 25th percentiles, is called the interquartile range (IQR), and is a measure of dispersion. If the data were normally distributed the IQR would be approximately 1.35 times the standard deviation. The skewness of the distribution is indicated most obviously by the relative length of the two whiskers. For example, for the BDI boxplot in Figure 2, the positive skewness of the distribution is indicated by the fact that the right whisker is much longer than the left one. Further, one can see that the median is closer to the 25th percentile than to the 75th percentile. Some programs indicate the location of the mean by a plus sign or other symbol in the middle of the box. Although that is not done in Figure T.2, the fact that the mean (15.748) is above the median (13) is one final indication of the positive skewness of the distribution. If you are not familiar with box plots, it may be helpful

### Box Plot of BDI Scores

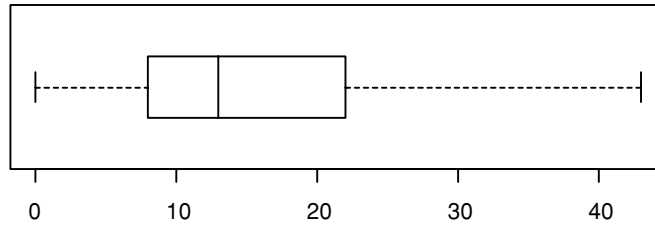


FIG. T.2. Box Plot showing box bounded by the 25th percentile (here a score of 8) on the left and the 75th percentile (a score of 22) on the right, split by a line at the median (13), with whiskers extending out to the minimum score (of 0) on the left and to the maximum score (of 43) on the right.

### BDI Score Frequency Distribution with Box Plot

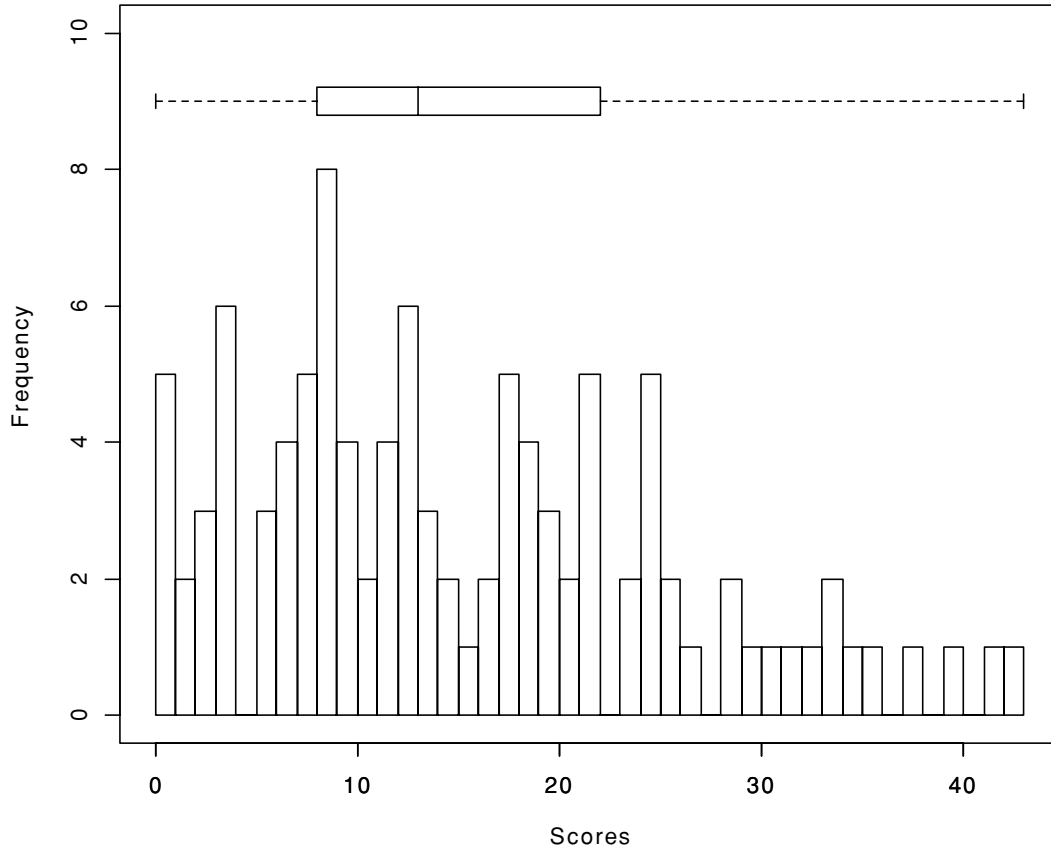


FIG. T.3. Histogram of BDI scores with box plot superimposed.

to see how a box plot summarizes a frequency distribution by superimposing a box plot on a histogram. This is done in Figure T.3, where one can see that here the whiskers extend out to include the entire range of scores in the sample.

To move now from the characterization of distributions, which is typical of *descriptive* statistics, to a review of some of the basic ideas of *inferential* statistics, the theoretically

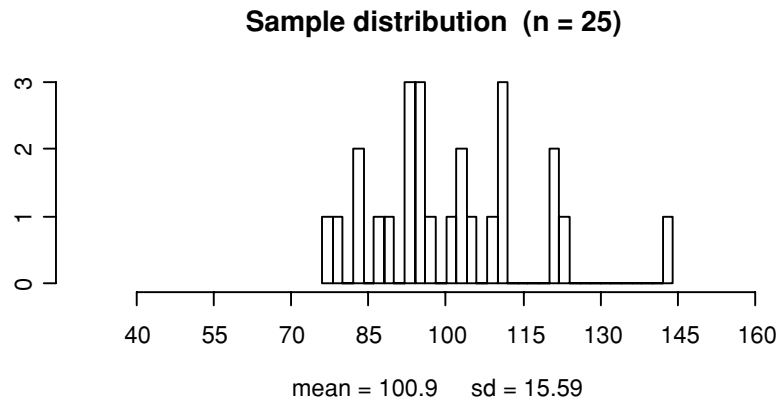


FIG. T.4. Frequency distribution of a sample of 25 scores drawn from a normally distributed population.

demonstrated properties of sampling distributions form the most fundamental basis for statistical inferences. The critical advance in the history of statistics was the derivation of the standard error of the mean, or standard deviation of the sampling distribution of sample means (Stigler, 1986). As Schmidt (1996, p. 121) has noted, even prior to the widespread adoption of hypothesis testing in the 1930's the known theory of the "probable error" of a sample mean informed data analysis. The basic idea of course is that a mean of a group of scores is more stable than an individual score. One can illustrate this graphically by first drawing a sample from a portion of the normal distribution. A sample of 25 scores was drawn from a normally distributed population having a mean of 100 and a standard deviation of 15. The distribution of the individual scores in the sample is shown in Figure T.4.

How does this sample distribution compare to the distribution of the population from which it was drawn? The distribution at the top of Figure T.5 is the population distribution of a normally distributed variable having a mean of 100 and a standard deviation of 15. (We will say more in a moment about the interpretation of such a distribution.) This theoretical distribution might be utilized for example as the approximate form of the population distribution for a standardized intelligence test. The distribution in the middle of the figure is the distribution of the sample of 25 Y scores drawn at random from the population. (A simple random sample is one drawn in such a way that every unit in the population has an equal chance of being included in the sample.) A listing of these (rounded) scores appears in Table T.1 (statistics on these rounded scores differ only slightly from those displayed in Figures T.4 and T.5). Since the point of using a random sample is to have a group of scores that is representative of the population distribution, it should not be surprising that the central tendency of the sample approximates the central tendency of the population and the variability of the sample approximates the variability of the population.<sup>4</sup>

Some comments about the appearance of the distributions are in order. The exact appearance of the graph of the sample distribution will depend in part on arbitrary decisions such as the width of the interval used for grouping scores. Choosing the width so that there are roughly 10–20 intervals is a common rule of thumb, but with small samples even fewer intervals may be desirable and with larger samples more than 20 may be preferable for getting a sense of the shape of the distribution (in the middle of Fig. T.5 the midpoints of successive intervals are 10 units apart resulting in 8 intervals, one of which is empty). Both the population distribution and the sample distribution indicate the probability of scores in a given range. Further, just as the total of the probabilities of all possible outcomes must add up to 1, the total area in a probability distribution also adds up to 1.

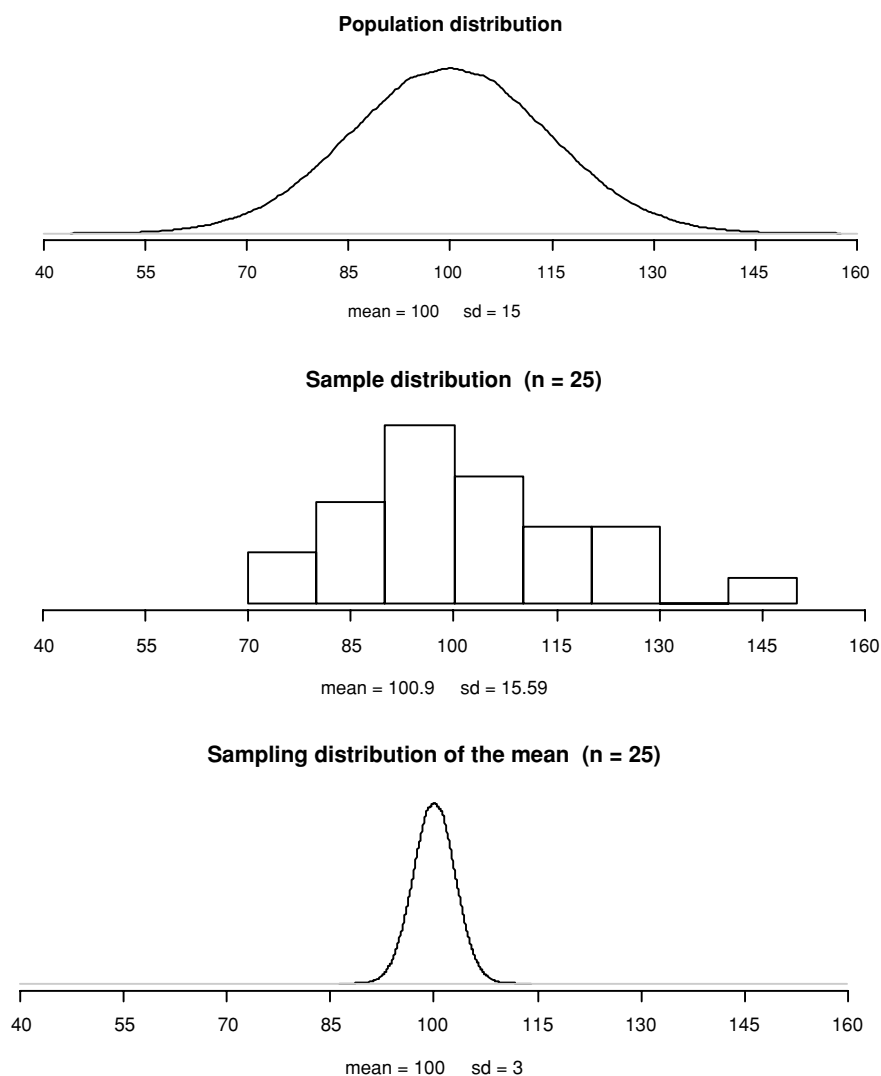


FIG. T.5. Comparing population, sample and sampling distributions.

**TABLE T.1**  
LISTING OF 25 SCORES IN SAMPLE

95	96	103	111	101
92	83	87	121	93
90	124	77	95	122
103	143	93	97	79
82	104	108	111	110

*Note.* Scores were randomly sampled from a normal distribution having a mean of 100 and a standard deviation of 15. Scores shown have been rounded to the nearest integer value.



Regarding the distribution at the bottom of Figure T.5, first note that a sampling distribution is a distribution of values of a statistic, not of individual scores. The values of the statistic are the values that would be observed in repeated “samplings” from the population. The idea is that in any large population there will be an extremely large number of possible samples of a given size that could be drawn. One could construct a sampling distribution empirically by continuing to draw one simple random sample after another from a population, for each computing the value of the statistic of interest, and then creating a distribution of the obtained values of the statistic.<sup>5</sup>

Not surprisingly the mean of the sample means is simply the mean of the population, i.e.  $\mu_{\bar{y}} = \mu$ . It is less clear just what the variability of the sample means will equal. Although increasing the variability of the population would induce greater variability in the sample and hence in the sample mean, it also is intuitively clear that whenever the samples consist of at least a few scores, larger scores will tend to be balanced out by smaller scores so that the sample mean will be fairly close to the center of the entire distribution. Further, it seems reasonable that the larger the sample, the closer the sample mean will tend to be to the population mean.<sup>6</sup> One of the most important results in all of mathematical statistics specifies just how the stability of the sample mean is affected by the size of the sample. Specifically, it is the case that the expected deviation of the sample mean from the population mean decreases as a function of the square root of the sample size. The expected deviation of a sample mean from the population mean is known as the standard error, and is the standard deviation of the sampling distribution. Thus, we can define the standard error of the sample mean,  $\sigma_{\bar{y}}$ , as being a fraction of the population standard deviation,  $\sigma$ , with the key term in the denominator of the fraction being  $\sqrt{n}$ , the square root of the sample size:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

This implies that, even in a sample as small as 25, the sample mean of a simple random sample of the population will have a standard deviation only 1/5 as large as that of the population of individual scores.

This fact about the sample standard deviation takes on great significance in conjunction with the central limit theorem. The central limit theorem says that regardless of the shape of the population distribution from which samples are drawn, independent sampling assures that the sampling distribution of the mean will more and more closely approximate a normal distribution as the sample size increases. As to the question of just how large the sample size needs to be, the typical rule of thumb given is that one can use the normal distribution for making judgments about the likelihood of various possible values occurring when the sample size is 30 or more. A more complete answer would note that it depends in part on the shape of the population distribution. If the population distribution is markedly non-normal, then a larger sample would be required. On the other hand, if the population from which samples are being drawn is itself normally distributed, as is the case in Figure T.5, then the sampling distribution of the mean will be normally distributed for any sample size.

The great benefit of these mathematical results is that they allow one to make precise statements about the probability of obtaining a sample mean within a certain range of values. Prior to the derivation of such results, experimenters who wondered about the stability of the mean obtained in a given study could only replicate the study repeatedly to empirically determine the variability in such a mean. We will illustrate the probability statements that can be made by using the sampling distribution at the bottom of Figure T.5 and the normal probability table appended to the end of this tutorial.

In probability distributions, the chance of values in a given range occurring is indicated by the area under the curve between those values. As you almost certainly recall from your first course in statistics, these probabilities may be determined by consulting a table giving areas under a normal curve. To allow use of a single set of tabled values, the relevant values for a given problem are converted to standard units, and the table provides areas under a “standard normal curve” which is one with a mean of 0 and a standard deviation of 1. The “standard unit” is the standard deviation, and thus to make use of such a table one needs to convert one’s original scores to standard scores or  $z$  scores. In words, a  $z$  score gives the location of a score in terms of the number of standard deviations above or below the mean it is, i.e.

$$z = \frac{\text{score} - \text{mean}}{\text{standard deviation}}$$

One can compute  $z$  scores as a descriptive measure of the relative standing of any score in any distribution regardless of how they are distributed. But one can reasonably make use of the tabled probabilities only if there is justification, such as the central limit theorem or information about the plausible shape of the population distribution, for presuming the scores are normally distributed. The table in the appendix gives the proportion of the area of a normally distributed distribution in the tail or the center of the distribution.

Since the standard deviation of the sampling distribution of the sample mean is the standard error of the mean, the form of the  $z$  score formula in this case is

$$z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$$

The normal distribution table in the Appendix indicates both the proportion of the area under the curve between the mean and a given  $z$  score and also the proportion beyond a given  $z$  score. One of the most useful numbers in this table to note is the area of .3413 between the mean of 0 and a  $z$  score of +1, which because of the symmetry of the normal distribution implies that .6826 of the area under the curve would be between  $z$  scores of  $-1$  and  $+1$ . The area between two positive  $z$  scores, e.g. between  $z = +1$  and  $z = +2$ , is determined by subtracting .3413 from the area between the mean and a  $z$  score of +2, i.e.  $.4772 - .3413 = .1359$ . Note that the areas under the curve between  $-1$  and 0 and between  $-2$  and  $-1$  must be equal to the corresponding areas above the mean because of the symmetry of the normal distribution.

Based on these areas we can make probability statements about the likelihood of obtaining a sample mean between particular values on the scale of the original variable. For example, with a sample size of 25 the probability that the sample mean will be between 97 and 103 is .6826, and the probability the sample mean will be between 94 and 106 is .9544. These are the most familiar values in the normal table and are the basis of saying that about 68% or roughly two thirds of a normal distribution is within one standard deviation of the mean, and approximately 95% is within 2 standard deviations of the mean.

One of the useful applications of such theory is in “interval estimation” or constructing confidence intervals. Suppose instead of having been given the population mean we were trying to estimate the mean IQ in a particular population but only had the data from the sample of 25 individuals shown in Table T.1 and Figure T.4 in hand. If we were willing to assert that the sampling distribution could be reasonably approximated by a normal distribution and also that the population standard deviation was known to be 15 (admittedly a rather implausible scenario), then the normal table could be used to make precise quantitative statements about the relative locations of the sample and population means. Our best “point estimate” of the population mean in such a case would be 100.9, the sample mean. Because the standard error

of this mean is  $\sigma/\sqrt{n} = 15/\sqrt{25} = 3$ , one could say that the population mean is  $100.9 \pm 3$ . For many purposes this may be sufficient, and is what is being communicated when one plots means with “error bars” extending out from the mean the length of 1 standard error. Although one is implicitly communicating something about the likelihood that the range indicated by the error bars will include the population mean, *interval estimation* typically connotes making an explicit probability statement in the form of a confidence interval. A confidence interval specifies how much confidence one should have, i.e. the probability, that a specified range of values will overlap the true population mean. The theory we have reviewed makes clear that with a normally distributed sampling distribution the probability the sample mean will be within two standard errors of the true population mean is roughly .95. More precisely, since a  $z$  score of 1.96 cuts off the 5% most extreme portions of a normal distribution (i.e., .025 in each tail), we can write

$$Pr(\mu - 1.96\sigma_{\bar{Y}} \leq \bar{Y} \leq \mu + 1.96\sigma_{\bar{Y}}) = .95$$

Although the above underscores that the sample mean is the random variable in the equation, it is algebraically equivalent to

$$Pr(\bar{Y} - 1.96\sigma_{\bar{Y}} \leq \mu \leq \bar{Y} + 1.96\sigma_{\bar{Y}}) = .95$$

This second form of the probability assertion is the more relevant to defining confidence intervals because it gives the limits suggested by a given sample mean for the boundaries around the unknown population mean. For example, for the sample in Figure T.4, the upper and lower limits of the confidence interval are computed as  $100.9 \pm 1.96(15/\sqrt{25})$ , or  $100.9 \pm 5.88$ . Thus, the confidence interval for the population mean is  $95.02 \leq \mu \leq 106.78$ . In general, a  $1 - \alpha$  confidence interval can be written

$$\bar{Y} - z_{\alpha/2}\sigma/\sqrt{n} \leq \mu \leq \bar{Y} + z_{\alpha/2}\sigma/\sqrt{n}$$

where  $z_{\alpha/2}$  is the  $z$  score corresponding to a tail probability of  $\alpha/2$ . Confidence intervals are denoted in terms of percentages, where higher percentages correspond to wider intervals, and with the relationship between the percentage and  $\alpha$  being  $100(1 - \alpha)\%$ . For example, to construct a 90% confidence interval,  $\alpha$  would be .10, since  $100(1 - .10)\% = 90\%$ , and the  $z$  score would be that corresponding to a tail probability of .05 or 1.645.

In discussing the logic of hypothesis testing in Chapter 2, we will talk about some common misconceptions like misinterpreting  $p$  values. If anything, the temptation for incorrect thinking about confidence intervals is even more seductive. For example, one must keep in mind that over repeated sampling from the same population, the confidence intervals will move around, whereas the population mean remains fixed. In a large number of replications, 95% of the confidence intervals will include this true population. But it would be incorrect to substitute the numerical limits of the confidence interval we just computed into the probability statement above and to write (or think) that

$$Pr(95.02 \leq \mu \leq 106.78) = .95$$

The problem is all of the terms above are constants, and in particular the population mean is a fixed constant that does not randomly vary over a range of values.

The other major application of the theory of sampling distributions is in hypothesis testing. Here the fact that the population mean is some unknown fixed value is made explicit by embodying alternative conceptions of its possible values into two mutually exclusive hypotheses,

the null and the alternative hypotheses. The null hypothesis is typically stated as an equality and the alternative hypothesis is stated as the complementary inequality. So for a study involving a sample of IQ scores like that in the middle of Figure T.5, one might have hypothesized in advance that the population mean was 110. The null hypothesis ( $H_0$ ) and the alternative hypothesis ( $H_1$ ) would then be written:

$$H_0 : \mu = 110$$

$$H_1 : \mu \neq 110$$

The null hypothesis is tentatively assumed to be true while carrying out a hypothesis test, even though typically it is the hypothesis that the experimenter would like to discredit or “nullify”. Under the same assumptions as were required to construct the confidence interval, namely, that the sampling distribution is normal in form and the population standard deviation is known, one can carry out a one-sample  $z$  test by computing the test statistic:

$$z = \frac{\bar{Y} - \mu_0}{\sigma_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

For example, using the sample data from Figure T.4 again, we would have

$$z = \frac{100.9 - 110}{15/\sqrt{25}} = \frac{-9.1}{3} = -3.03$$

This observed value of the test statistic, which could be denoted  $z_{obs}$ , would be compared to a critical value  $z_{crit}$  determined by table look up to determine whether to reject the null hypothesis. Using  $\alpha = .05$ , since the alternative hypothesis is non-directional one would typically employ a two-tailed test with the region where the null hypothesis would be rejected being the  $\alpha/2$  extreme portion of each of the two tails of the sampling distribution. So again, the tabled value used would be  $\pm 1.96$ , with the decision rule being to reject  $H_0$  if  $|z_{obs}| > 1.96$ . Since the observed  $z$  value here indicates that the sample mean is 3 standard errors below the hypothesized mean, the test statistic is well within the rejection region and we would reject the null hypothesis that the population mean is 110.<sup>7</sup>

As we remarked above, it is rather implausible that one would have knowledge of the population standard deviation in most situations. Instead, it is much more common that one will be estimating the population standard deviation on the basis of sample data at the same time one wants to make an inference from those data regarding the population mean. The ubiquitous  $t$  test solves this problem and yields a test statistic identical in form to the  $z$  test except that the population standard deviation is estimated by the sample standard deviation. That is, one tests the same hypotheses as before but using

$$t = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{\hat{\sigma}/\sqrt{n}} = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}$$

The observed value of this statistic is referred to a value from a table like that in Appendix A.1 of the text to determine its significance. The principal differences between the  $z$  and  $t$  tests from the point of view of the researcher are that the critical value depends on the sample size in the case of the  $t$  test, and these critical values will be larger than the critical  $z$  value for the same  $\alpha$  level. Instead of there being a single standard normal distribution with 0 mean and standard deviation of 1, there is a slightly different  $t$  distribution for each value of  $n$ . The tabled

$t$  distributions all have a zero mean but will typically have a standard deviation a bit larger than 1 (cf. Searle, 1971, p. 48). The form of the  $t$  distribution is essentially indistinguishable from the normal for large  $n$  but the  $t$  is more and more heavy-tailed than the normal the smaller the sample size. The practical implication is that the smaller the sample size, the larger will be the critical  $t$  value that has to be achieved to declare a result significant. The particular  $t$  distribution utilized in a one-sample test is that denoted by  $df = n - 1$ , where  $df$  denotes “degrees of freedom” and corresponds to the denominator term used in computing the sample variance. (The concept of degrees of freedom is explained in considerably more detail in Chapter 3.)

For example, in the case of the illustrative sample data pictured in Figure T.4 where the mean of 100.9 and standard deviation of 15.59 was computed on the basis of a sample size of 25, we would have:

$$t = \frac{\bar{Y} - \mu_0}{\hat{\sigma}_{\bar{Y}}} = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{100.9 - 110}{15.59/\sqrt{25}} = \frac{-9.10}{3.12} = -2.92$$

The critical value here is for a  $t$  with  $n - 1 = 24$   $df$ , or  $t_{crit} = 2.06$  for a two-tailed test at  $\alpha = .05$ . Thus, using the decision rule of rejecting  $H_0$  if  $|t_{obs}| > 2.06$ , we again reject the null hypothesis that  $\mu = 110$  even though the critical value is a bit larger than the value used in the  $z$  test (1.96).

Although we will not take the time to develop other forms of the  $t$  test in detail, we note that all  $t$  tests have the same general form which can be described in words as

$$t = \frac{\text{statistic} - \text{parameter}}{\text{estimated standard error of statistic}}$$

In the case of the one-sample test we just considered the statistic of interest was the sample mean,  $\bar{Y}$ . In the case of a two-group design, the statistic of interest will typically be the difference between the sample means of the two groups, i.e.  $\bar{Y}_1 - \bar{Y}_2$ . The parameter is the expected value of the statistic according to the null hypothesis. The null hypothesis in the two-group will often be that the two populations means,  $\mu_1$  and  $\mu_2$ , are equal which implies that the sampling distribution of the difference in sample means will be centered around  $\mu_1 - \mu_2 = 0$ . If one assumes the population variance of the two groups are equal, then the estimated standard error of the difference in sample means will be

$$\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}$$

where  $n_1$  and  $n_2$  are the sample sizes in groups 1 and 2, and  $s_1^2$  and  $s_2^2$  are the unbiased sample variances in groups 1 and 2, respectively. Thus, the final form of the two-group  $t$  test is

$$t = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\bar{Y}_1 - \bar{Y}_2}} = \frac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \left[ \frac{1}{n_1} + \frac{1}{n_2} \right]}}$$

which will have  $df = n_1 + n_2 - 2$ .

What may not be obvious from the above form of the two-groups test is how the standard error of the difference in means relates to the standard errors of the individual means being compared. One might be tempted to think that because the difference between the two means

is typically a smaller number than either mean that their difference will be less variable than either mean. That this is not true is more evident if we examine what the standard error of the difference in means reduces to in the equal- $n$  case:

$$\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{n}}$$

That is, the standard error of the difference in means is equal to the square root of the sum of the squares of the standard errors of the individual means. Thus, the relation is like that in the Pythagorean theorem, where the length of the hypotenuse is related to the lengths of the legs in a right triangle in a similar way. The implication then is that, rather than being less than the standard error of the individual means, the standard error of the difference in means will be greater than either individual standard error though it will be less than their sum. Methods for arriving at appropriate estimates of the standard errors of combinations of means will be a common concern in the current book, particularly when there is evidence of heterogeneity of variance. Methods based on the  $t$  test formulation for testing specific hypotheses about particular combinations of means will prove particularly useful as a means for dealing with heterogeneity of variance (see Chapter 4, especially p. 163ff.). Further, as will be noted at different points in the main body of the current volume, not just tests of specific comparisons but the various overall tests of interest in multiple group designs can be viewed as generalizations of the simple one- and two-group  $t$  tests we have briefly described here.

## REFERENCES

- Bliss, C. I. (1967). *Statistics in biology: Statistical methods for research in the natural sciences. Volume 1*. New York: McGraw-Hill Book Company.
- DeCarlo, L. T. (1996). On the meaning and use of kurtosis. *Psychological Methods*, 2, 292–307.
- Freedman, D., Pisani, R., & Purves, R. (1998). *Statistics* (3rd ed.). New York: W. W. Norton.
- Harris, R. J. (1997). Reforming significance testing via three-value logic. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.
- Hildebrand, D. K. (1986). *Statistical thinking for behavioral scientists*. Boston: Duxbury Press.
- Hull, C. H., & Nie, N. H. (1981). *SPSS Update 7-9: New Procedure and Facilities for Releases 7-9*. New York: McGraw-Hill Book Company.
- SAS Institute. (1990). *SAS Procedures Guide, Version 6, 3rd Edition*. Cary, NC: SAS Institute.
- Schmidt, F. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods*, 1, 115–129.
- Searle, S. R. (1971). *Linear models*. New York: John Wiley.
- Smith, J. E., Meyers, R. J., & Delaney, H. D. (1998). The Community Reinforcement Approach with homeless alcohol-dependent individuals. *Journal of Consulting and Clinical Psychology*, 66, 541–548.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: Belknap.
- Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

# Notes

- Perhaps the leptokurtic distribution most familiar to students in an elementary statistics course is the  $t$  distribution, which is similar in shape to the normal curve but which for lower degrees of freedom has a larger proportion of scores in the tails of the distribution. This is why values of the  $t$  distribution used as “critical values”, or values cutting off the 5% most extreme portion of the distribution, are larger numerically in absolute value than the values that cut off the corresponding percentage of the normal distribution.
- One can characterize variance, skewness and kurtosis by using the expected value of the deviations from the mean raised to the second, third or fourth powers, respectively. These expected values are referred to as the moments of the distribution. Using this terminology, the variance is called the second moment of the distribution:

$$\sigma^2 = \mathcal{E}(Y - \mu)^2$$

Obviously the variance depends on how far from the mean scores are, but this will be true of the higher moments as well. To come up with measures of skewness and kurtosis which are adjusted for the variance, it is conventional to define them as the expected value of the deviations from the mean raised to the third or fourth power divided by the standard deviation raised to the same power. Thus, the skewness of a population can be defined as

$$\alpha_3 = \frac{\mathcal{E}(Y - \mu)^3}{\sigma^3}$$

and kurtosis as

$$\alpha_4 = \frac{\mathcal{E}(Y - \mu)^4}{\sigma^4}$$

In a normal distribution, the value of these parameters is  $\alpha_3 = 0$  and  $\alpha_4 = 3$ . In a sample, skewness and kurtosis are typically estimated by statistics that adjust for the size of the sample, and are expressed to indicate how far from a normal distribution the shape is (cf. Bliss, 1967, p. 140ff.). For example, SPSS (Hull and Nie, 1981, p. 312) and SAS (SAS, 1990, p. 4) estimate skewness as

$$\frac{N}{N-2} \frac{\sum (Y_i - \bar{Y})^3 / (N-1)}{s^3}$$

and kurtosis as

$$\frac{N(N+1)}{(N-2)(N-3)} \frac{\sum (Y_i - \bar{Y})^4 / (N-1)}{s^4} - \frac{3(N-1)(N-1)}{(N-2)(N-3)}$$

- As DeCarlo (1996) has stressed in a very clearly written explanation of kurtosis, in symmetric distributions positive kurtosis “indicates an excess in either the tails, the center, or both”, but that “kurtosis primarily reflects the tails, with the center having a smaller influence”. In practice, it is the case that distributions with extreme kurtosis often are very asymmetric (see Chapter 3 sections on “Checking for Normality and Homogeneity of Variance” and “Transformations”).
- Students sometimes mistakenly think that because a population or large sample contains many more scores than a small sample it will have a larger variance. Although the observed range of scores may be expected to increase with the number of scores in a group, the sample variance, computed using  $N-1$  as noted above, is an unbiased estimate of the population variance, regardless of the size of the sample.

5. In practice, one could construct the entire sampling distribution exactly only in the case of a very small finite population. For example, if there were only 5 scores in the population and one were using a sample size of 2 and sampling without replacement, then there would be 5 ways of choosing the first score to include in the sample and 4 ways of choosing the second score in the sample. The entire sampling distribution of the sample mean or variance could then be determined by computing such statistics for each of the 20 samples and determining the relative frequency of each possible value (see Hildebrand, 1986, p. 233 for such an example). However, the properties of the theoretical sampling distribution have been derived mathematically for normally distributed populations, and in practice researchers rely on such theoretical results rather than attempting to construct sampling distributions empirically. We will consider an example of an empirical sampling distribution in a context where normality is not assumed in Chapter 2.
6. This same idea was expressed around 1700 by Jacob Bernoulli, one of a family of eminent Swiss mathematicians who became the “father of the quantification of uncertainty”, when he bluntly asserted it was common knowledge that uncertainty decreased as the number of observations increased: “For even the most stupid of men, by some instinct of nature, by himself and without any instruction (which is a remarkable thing), is convinced that the more observations have been made, the less danger there is of wandering from one’s goal” (see Stigler, 1986, pp. 63–65). The more difficult problem that Bernoulli worked on was determining quantitatively *just how much* the uncertainty in estimation of a mean decreased as a result of a given increase in the number of observations. The key principle, later denoted the square root rule, was discovered about 20 years later by Abraham De Moivre, a mathematician from France (De Moivre actually spent most of his life as an expatriate in England after being imprisoned in France for his Protestant beliefs) (Freedman, Pisani, & Purves, 1998, pp. 221–224, 308ff.; Stigler, 1986, pp. 83–84).
7. While some purists might argue that one should not make any conclusions about the direction of the difference given this conventional approach to hypothesis testing, in practice investigators always do and we would argue they should. This can be justified in two ways. First, one might argue that one is using formal hypothesis testing as a screen to determine when it is reasonable to conclude the null hypothesis is false, but that one then still intends to think about the most defensible interpretation. Having decided by a hypothesis test that  $\mu = 110$  is false, one could then conclude outside of the formal hypothesis testing logic that the most rationally defensible interpretation is that  $\mu < 110$ . Second, one could argue that often when one does a two-tailed test one implicitly is simultaneously considering two alternative hypotheses  $H_{1A}: \mu < 110$  and  $H_{1B}: \mu > 110$ . Although typically not made explicit, the implicit decision rule could be said to be that when the observed value of the test statistic falls in the tail of the sampling distribution more consonant with the particular alternative hypothesis specifying that direction, then accept that form of the alternative hypothesis. While in recent years some (e.g. Harris, 1997) have argued for formally codifying this “three-valued logic”, we do not so primarily because the logic does not generalize to tests of null hypotheses involving several group means, e.g.  $\mu_1 = \mu_2 = \mu_3 = \mu_4$ , which will be a major focus of the current volume.



# Appendix

## Table of Proportions of Area under the Standard Normal Curve

