3

Basic Statistics: A Review

3.1 Preview

This chapter reviews the fundamental statistical concepts and methods that are needed to understand the more sophisticated multivariable techniques discussed in this text. Through this review, we shall introduce the statistical notation (using conventional symbols whenever possible) employed throughout the text.

The broad area associated with the word *statistics* involves the methods and procedures for collecting, classifying, summarizing, and analyzing data. We shall focus on the latter two activities here. The primary goal of most statistical analysis is to make *statistical inferences*—that is, to draw valid conclusions about a *population* of items or measurements based on information contained in a *sample* from that population.

A *population* is any set of items or measurements of interest, and a *sample* is any subset of items selected from that population. Any characteristic of that population is called a *parameter*, and any characteristic of the sample is termed a *statistic*. A statistic may be considered an estimate of some population parameter, and its accuracy of estimation may be good or bad.

Once sample data have been collected, it is useful, prior to analysis, to examine the data using tables, graphs, and *descriptive statistics*, such as the sample mean and the sample variance. Such descriptive efforts are important for representing the essential features of the data in easily interpretable terms.

Following such examination, statistical inferences are made through two related activities: *estimation* and *hypothesis testing*. The techniques involved here are based on certain assumptions about the probability pattern (or *distribution*) of the (*random*) variables being studied.

Each of the preceding key terms—*descriptive statistics, random variables, probability distribution, estimation,* and *hypothesis testing*—will be reviewed in the sections that follow.

3.2 Descriptive Statistics

A *descriptive statistic* may be defined as any single numerical measure computed from a set of data that is designed to describe a particular aspect or characteristic of the data set. The most common types of descriptive statistics are measures of *central tendency* and of *variability* (or *dispersion*).

The central tendency in a sample of data is the "average value" of the variable being observed. Of the several measures of central tendency, the most commonly used is the sample mean, which we denote by $\bar{X}$ whenever our underlying variable is called X. The formula for the sample mean is given by

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

where n denotes the sample size; $X_1, X_2, \ldots, X_n$ denote the n measurements (or observed values) of X; and $\sum$ denotes summation. The sample mean $\bar{X}$—in contrast to other measures of central tendency, such as the median or mode—uses in its computation all the observations in the sample. This property means that $\bar{X}$ is necessarily affected by the presence of extreme X-values, so in some cases it may be preferable to use the median instead of the mean.

Measures of central tendency (such as $\bar{X}$) do not, however, completely summarize all features of the data. Obviously, two sets of data with the same mean can differ widely in appearance (e.g., an $\bar{X}$ of 4 results both from the values 4, 4, and 4 and from the values 0, 4, and 8). Thus, we customarily consider, in addition to $\bar{X}$, measures of variability, which tell us the extent to which the values of the measurements in the sample differ from one another.

The two measures of variability most often considered are the *sample variance* and the *sample standard deviation*. These are given by the following formulas when considering observations $X_1, X_2, \ldots, X_n$ on a single variable X:

$$\text{Sample variance} = S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad (3.1)$$

$$\text{Sample standard deviation} = S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2} \qquad (3.2)$$

The formula for S^2 describes variability in terms of an average of squared deviations from the sample mean—although $(n-1)$ is used as the divisor instead of n, due to considerations that make S^2 a good estimator of the variability in the entire population.

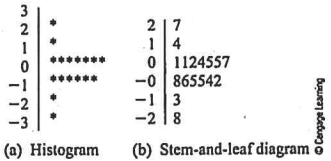
A drawback to the use of S^2 is that it is expressed in squared units of the underlying variable X. To obtain a measure of dispersion that is expressed in the same units as X, we simply take the square root of S^2 and call it the sample standard deviation S. Using S in combination with $\bar{X}$ thus gives a fairly succinct picture of both the amount of spread and the center of the data, respectively.

When more than one variable is being considered in the same analysis (as will be the case throughout this text), we will use different letters and/or different subscripts to differentiate among the variables, and we will modify the notations for mean and variance accordingly. For example, if we are using X to stand for age and Y to stand for systolic blood pressure, we will denote the sample mean and the sample standard deviation for each variable as $(\overline{X}, S_X)$ and $(\overline{Y}, S_Y)$, respectively.

All statistical analyses begin with an examination of descriptive statistics computed from the data set at hand. However, often the most direct and revealing way to examine the data is to make a series of plots. We describe three types of simple but useful plots: histograms (especially stem-and-leaf versions), schematic plots, and normal probability plots.

Suppose that we have collected data on the amount of error that occurs in measurements taken with a particular type of instrument. We think the error may be related to the age of the instrument; therefore, readings are taken with 17 instruments of varying ages; the age of each instrument and the error in its measurement are recorded.

In our descriptive analysis of these data, first we examine a frequency histogram of the measurement errors, shown in Figure 3.1(a). We observe that the errors appear to be quite symmetrically distributed around 0 (i.e., the mean and the median error are roughly 0) and that the picture approximates a bell-shaped curve. (See Section 3.3.2 for more information on data that follow this pattern.) No *outliers* (data points that are extreme in value and that may represent data errors) or other anomalies appear to be present.



(a) Histogram (b) Stem-and-leaf diagram

FIGURE 3.1 Frequency histogram and stem-and-leaf diagram of instrument error data ($n = 17$)

The frequency histogram conveys even more information if it is converted into a stem-and-leaf diagram, as in Figure 3.1(b), which shows the actual data values while maintaining the shape of the histogram. In the stem-and-leaf diagram, the top-most value has a *stem* of 2 and a *leaf* of 7, indicating that the original data value is 2.7. Beneath that is a value of 1.4 (stem 1, leaf 4); after that are two values that both share a stem of 0 and have leaves equal to 1 (i.e., both values are 0.1), followed by a value of 0.2, and so on. The last value shown in the plot is −2.8 (stem −2, leaf 8).

The second kind of useful plot is a schematic plot. Figure 3.2 presents a schematic plot of the measurement error data. A schematic plot is based entirely on the order of the values in the sample. *Quartiles* are the most important order-based statistics for the schematic plot. The *first quartile*, or 25th percentile, is the value at or below which 25% of the data values lie; the *second quartile*, or 50th percentile (or median), is the value at or below which 50% of the data values lie; the *third quartile*, or 75th percentile, is the value at or below which 75% of the data values

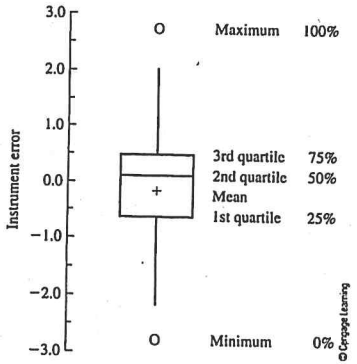


FIGURE 3.2 Schematic plot of instrument error data ($n = 17$)

lie. The *interquartile range* (IQR), calculated as the value of the third quartile minus the value of the first quartile, is a measure of the spread of a distribution, like the variance. One important difference between the IQR and the variance, however, is illustrated by the fact that, whereas doubling the largest value in the sample would, in general, increase variance dramatically, it would not change the IQR. For the error measurements, the first, second, and third quartiles are approximately −0.6, 0.1, and 0.5, respectively, with an interquartile range of 1.1.

A schematic plot is sometimes called a *box-and-whisker plot*, or simply a *boxplot*, due to its appearance. The box is outlined by three horizontal lines, which mark the values of the first quartile, the second quartile (the median), and the third quartile (see Figure 3.2). The scale is determined by the units and range of the data. The mean is indicated by a + on the backbone of the plot. If the data are symmetric, the mean and median will be close in value (i.e., the + will be marked on or close to the middle horizontal line), and the distances between the first and second quartiles and between the second and third quartiles will be similar in size. The whiskers (vertical lines) extend from the box as far as the data extend up or down, to a limit of 1.5 IQRs (in the vertical direction). An O at the end of a whisker indicates a moderate outlier. Referring to Figure 3.2, we see one positive moderate outlier and one negative moderate outlier.

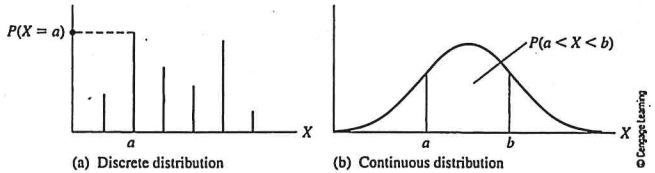
3.3 Random Variables and Distributions

The term *random variable* is used to denote a variable whose observed values may be considered outcomes of a stochastic or random experiment (e.g., the drawing of a random sample). The values of such a variable in a particular sample, then, cannot be anticipated with certainty before the sample is gathered. Thus, if we select a random sample of persons

from some community and determine the systolic blood pressure (W), cholesterol level (X), race (Y), and sex (Z) of each person, then W, X, Y, and Z are four random variables whose particular realizations (or observed values) for a given person in the sample cannot be known for sure beforehand. In this text, we shall denote random variables by capital italic letters.

The probability pattern that gives the relative frequencies associated with all the possible values of a random variable in a population is generally called the *probability distribution* of the random variable. We represent such a distribution by a table, graph, or mathematical expression that provides the probabilities corresponding to the different values or ranges of values taken by a random variable.

Discrete random variables (such as the number of deaths in a sample of patients or the number of arrivals at a clinic), whose possible values are countable, have (gappy) distributions that are graphed as a series of vertical lines; the heights of these lines represent the probabilities associated with the various possible discrete outcomes (Figure 3.3(a)). *Continuous* random variables (such as blood pressure and weight), whose possible values are uncountable, have (nongappy) distributions that are graphed as smooth curves; an *area* under such a curve represents the probability associated with a *range of values* of the continuous variable (Figure 3.3(b)). We note in passing that the probability of a continuous random variable taking one particular value is 0 because there can be no area above a single point. For discrete distributions, the sum of the probabilities for all possible values of X is equal to 1. For continuous distributions, the total area under the curve representing the distribution is equal to 1.



(a) Discrete distribution (b) Continuous distribution

FIGURE 3.3 Discrete and continuous distributions: $P(X = a)$ is read: "The probability that X takes the value a"

In the next two subsections, we will discuss two particular distributions of enormous practical importance: the binomial (a discrete distribution) and the normal (a continuous distribution).

3.3.1 The Binomial Distribution

A *binomial* random variable describes the number of occurrences of a particular event in a series of n trials, under the following four conditions:

1. The n trials are conducted identically.
2. There are two possible outcomes of each trial: "success" (i.e., the event of interest occurs) or "failure" (i.e., the event of interest does not occur), with probabilities π and $1 - \pi$, respectively.

3. The outcome of any one trial is independent of (i.e., is not affected by) the outcome of any other trial.
4. The probability of success, π, remains the same for all trials.

For example, the distribution of the number of lung cancer deaths in a random sample of $n = 400$ persons would be considered binomial only if the four conditions were all satisfied, as would the distribution of the number of persons in a random sample of $n = 70$ who favor a certain form of legislation.

The two elements of the binomial distribution that one must specify to determine the precise shape of the probability distribution and to compute binomial probabilities are the sample size n and the parameter π. The usual notation for this distribution is, therefore, $B(n, \pi)$. If X has a binomial distribution, it is customary to write

$$X \sim B(n, \pi)$$

where $\sim$ stands for "is distributed as." The probability formula for this discrete random variable X is given by the expression

$$P(X = j) = {}_nC_j \pi^j (1 - \pi)^{n-j} \quad j = 0, 1, \ldots, n$$

where ${}_nC_j = n!/[j!\,(n - j)!]$ denotes the number of combinations of n distinct objects selected j at a time.

3.3.2 The Normal Distribution

The *normal distribution*, denoted as $N(\mu, \sigma)$, where μ and σ are the two parameters, is described by the well-known bell-shaped curve (Figure 3.4). The parameters μ (the mean) and σ (the standard deviation) characterize the center and the spread, respectively, of the distribution. We generally attach a subscript to the parameters μ and σ to distinguish among variables; that is, we often write

$$X \sim N(\mu_X, \sigma_X)$$

to denote a normally distributed X.

An important property of any normal curve is its *symmetry*, which distinguishes it from some other continuous distributions that we will discuss later. This symmetry property is quite helpful when using tables to determine probabilities or percentiles of the normal distribution.

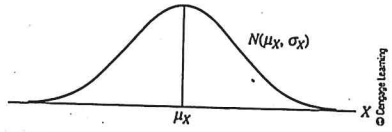


FIGURE 3.4 A normal distribution

Probability statements about a normally distributed random variable X that are of the form $P(a \le X \le b)$ require for computation the use of a single table (Table A.1 in Appendix A). This table gives the probabilities (or areas) associated with the *standard normal distribution*, which is a normal distribution with $\mu = 0$ and $\sigma = 1$. It is customary to denote a standard normal random variable by the letter Z, so we write

$$Z \sim N(0, 1)$$

To compute the probability $P(a \le X \le b)$ for an X that is $N(\mu_X, \sigma_X)$, we must transform (i.e., *standardize*) X to Z by applying the conversion formula

$$Z = \frac{X - \mu_X}{\sigma_X} \tag{3.3}$$

to each of the elements in the probability statement about X, as follows:

$$P(a \le X \le b) = P\left(\frac{a - \mu_X}{\sigma_X} \le Z \le \frac{b - \mu_X}{\sigma_X}\right)$$

We then look up the equivalent probability statement about Z in the $N(0, 1)$ tables.

For random samples, this rule also applies to the sample mean $\bar{X}$ whenever the underlying variable X is normally distributed or whenever the sample size is moderately large (by the Central Limit Theorem). But because the standard deviation of $\bar{X}$ is $\sigma_X/\sqrt{n}$, the conversion formula has the form

$$Z = \frac{\bar{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

A *percentile* is a value of a random variable X *below which* the area under the probability distribution has a certain specified value. We denote the (100p)th percentile of X by X_p and picture it as in Figure 3.5, where p is the amount of area under the curve to the left of X_p. In determining X_p for a given p, we must again use the conversion formula (3.3). Since the procedure requires that we first determine Z_p and then convert back to X_p, however, we generally rewrite the conversion formula as

$$X_p = \mu_X + \sigma_X Z_p \tag{3.4}$$

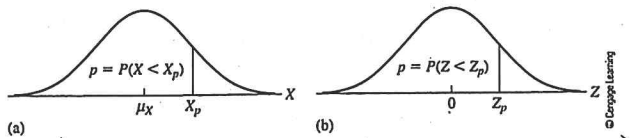


FIGURE 3.5 The (100p)th percentiles of X and Z

For example, if $\mu_X = 140$ and $\sigma_X = 40$, and we want to find $X_{0.95}$, the $N(0, 1)$ table first gives us $Z_{0.95} = 1.645$, which we convert back to $X_{0.95}$ as follows:

$$X_{0.95} = 140 + (40)Z_{0.95} = 140 + 40(1.645) = 205.8$$

Formulas (3.3) and/or (3.4) can also be used to approximate probabilities and percentiles for the binomial distribution $B(n, \pi)$ whenever n is moderately large (e.g., $n > 20$). Two conditions are usually required for this approximation to be accurate: $n\pi > 5$ and $n(1 - \pi) > 5$. Under such conditions, the mean and the standard deviation of the approximating normal distribution are

$$\mu = n\pi \quad \text{and} \quad \sigma = \sqrt{n\pi(1 - \pi)}$$

A *normal probability plot* assesses how well the sample data adhere to a normal distribution, in order to help infer whether the data are sampled from a normally distributed population. The ordered data values are plotted against corresponding percentiles from an estimated normal distribution. Plots that are linear in appearance are consistent with the assumption of normality, since the cumulative relative frequencies for a normal distribution plot as a straight line. For example, in Figure 3.6, plot (a) supports the assumption that the data constitute a random sample from a normal distribution; the other plots suggest deviations from this assumption.

The *skewness* and *kurtosis* statistics can also be helpful in assessing normality. *Skewness* indicates the degree of asymmetry of a distribution. Just as variance is the average squared deviation of observations about the mean, skewness is the average of cubed deviations about the mean. To simplify comparisons between samples and to help account for estimation in small samples, skewness is usually computed as

$$\text{sk}(X) = \left(\frac{n}{n-2}\right)\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S_X}\right)^3$$

For large n, sk(X) should be approximately equal to 0 for a random sample size of n from any symmetric probability distribution (such as a normal distribution). Positive values of sk(X) indicate that relatively more values are above the mean than below it; the sample values are thus said to be "positively skewed." A negative value for sk(X) indicates that relatively more values are below the mean than above it.

Kurtosis indicates the heaviness of the tails relative to the middle of a distribution. Because kurtosis is the average of the fourth power of the deviations about the mean, it is always nonnegative. Standardized kurtosis may be computed as

$$\text{Kur}(X) = \left[\frac{n(n+1)}{(n-2)(n-3)}\right]\left(\frac{1}{n-1}\right)\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{S_X}\right)^4$$

The term in brackets, which approaches 1 as n increases, helps to account for estimation based on a small sample. Since standardized kurtosis for a standard normal distribution is 3, this value is often subtracted from Kur(X). The resulting statistic can be as small as -3 for flat distributions with short tails; it is approximately zero for moderate to large random samples from a normal distribution, and it is positive for heavy-tailed distributions. Thus, the positive

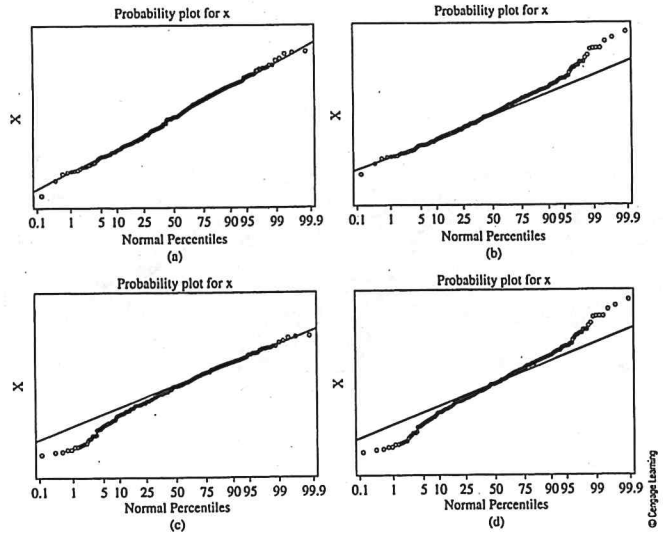


FIGURE 3.6 Normal probability plots

kurtosis value in our example (the reader is encouraged to do the required calculations) suggests a distribution with tails heavier than for a normal distribution. Skewness and kurtosis statistics are highly variable in small samples and hence are often difficult to interpret.

3.4 Sampling Distributions of t, χ^2, and F

The Student's t, chi-square (χ^2), and Fisher's F distributions are particularly important in statistical inference making.

The (*Student's*) *t distribution* (Figure 3.7(a)), which like the standard normal distribution is symmetric about 0, was originally developed to describe the behavior of the random variable

$$T = \frac{\overline{X} - \mu_X}{\frac{S_X}{\sqrt{n}}} \tag{3.5}$$

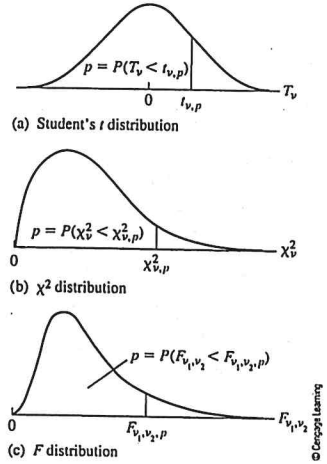


FIGURE 3.7 The t, χ^2, and F distributions

which represents an alternative to

$$Z = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}}$$

whenever the population variance σ_X^2 is unknown and is estimated by S_X^2. The denominator of (3.5), $S_X/\sqrt{n}$, is the *estimated standard error of* $\overline{X}$. When the underlying distribution of X is normal and when $\overline{X}$ and S_X^2 are calculated using a random sample from that normal distribution, then (3.5) has the *t distribution with* $n - 1$ *degrees of freedom*, where $n - 1$ is the quantity that must be specified in order to look up tabulated percentiles of this distribution. We denote all this by writing

$$T = \frac{\overline{X} - \mu_X}{\frac{S_X}{\sqrt{n}}} \sim t_{n-1}$$

It has generally been shown by statisticians that the t distribution is sometimes appropriate for describing the behavior of a random variable of the general form

$$T = \frac{\hat{\theta} - \mu_{\hat{\theta}}}{S_{\hat{\theta}}} \tag{3.6}$$

where $\hat{\theta}$ is any random variable that is normally distributed with mean $\mu_{\hat{\theta}}$ and standard deviation $\sigma_{\hat{\theta}}$, where $S_{\hat{\theta}}$ is the estimated standard error of $\hat{\theta}$, and where $\hat{\theta}$ and $S_{\hat{\theta}}$ are statistically independent. For example, when random samples are taken from two normally distributed populations with the same standard deviation (e.g., from $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$), and we consider $\hat{\theta} = \bar{X}_1 - \bar{X}_2$ in (3.6), we can write

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{3.7}$$

estimates the common variance σ^2 in the two populations. The quantity S_p^2 is called a *pooled sample variance,* since it is calculated by pooling the data from both samples in order to estimate the common variance σ^2.

The *chi-square* (or χ^2) *distribution* (Figure 3.7(b)) is a nonsymmetric distribution and describes, for example, the behavior of the nonnegative random variable

$$\frac{(n-1)S^2}{\sigma^2} \tag{3.8}$$

where S^2 is the sample variance based on a random sample of size n from a normal distribution. The variable given by (3.8) has the chi-square distribution with $n - 1$ degrees of freedom:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Because of the nonsymmetry of the chi-square distribution, both upper and lower percentage points of the distribution need to be tabulated, and such tabulations are solely a function of the degrees of freedom associated with the particular χ^2 distribution of interest. The chi-square distribution has widespread application in analyses of categorical data.

The *F distribution* (Figure 3.7(c)), which like the chi-square distribution is skewed to the right, is often appropriate for modeling the probability distribution of the ratio of independent estimators of two population variances. For example, given random samples of sizes n_1 and n_2 from $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$, respectively, so that estimates S_1^2 and S_2^2 of σ_1^2 and σ_2^2 can be calculated, it can be shown that

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \tag{3.9}$$

has the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, which are called the *numerator* and *denominator* degrees of freedom, respectively. We write this as

$$\frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F_{n_1-1,\,n_2-1}$$

The F distribution can also be related to the t distribution when the numerator degrees of freedom equal 1; that is, the square of a variable distributed as Student's t with ν degrees of freedom has the F distribution with 1 and ν degrees of freedom. In other words,

$$T^2 \sim F_{1,\nu} \quad \text{if and only if} \quad T \sim t_\nu$$

Percentiles of the t, χ^2, and F distributions may be obtained from Tables A.2, A.3, and A.4 in Appendix A. The shapes of the curves that describe these probability distributions, together with the notation we will use to denote their percentile points, are given in Figure 3.7.

3.5 Statistical Inference: Estimation

Two general categories of statistical inference—estimation and hypothesis testing—can be distinguished by their differing purposes: estimation is concerned with quantifying the specific value of an unknown population parameter; hypothesis testing is concerned with making a decision about a hypothesized value of an unknown population parameter.

In estimation, which we focus on in this section, we want to estimate an unknown parameter θ by using a random variable $\hat{\theta}$ ("theta hat," called a *point estimator* of θ). This point estimator takes the form of a formula or rule. For example,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{or} \quad S_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

tells us how to calculate a specific point estimate, given a particular set of data.

To estimate a parameter of interest (e.g., a population mean μ, a binomial proportion π, a difference between two population means $\mu_1 - \mu_2$, or a ratio of two population standard deviations σ_1/σ_2), the usual procedure is to select a random sample from the population or populations of interest, calculate the point estimate of the parameter, and then associate with this estimate a measure of its variability, which usually takes the form of a confidence interval for the parameter of interest.

As its name implies, a *confidence interval* (often abbreviated CI) consists of two random boundary points between which we have a certain specified *level of confidence* that the population parameter lies. More specifically, a 95% confidence interval for a parameter θ consists of lower and upper limits determined so that, in many repeated sets of samples of the same size, about 95% of all such intervals would be expected to contain the parameter θ. Care must be taken when interpreting such a confidence interval not to consider θ a random variable that either falls or does not fall in the calculated interval; rather, θ is a fixed (unknown) constant, and the random quantities are the lower and upper limits of the confidence interval, which vary from sample to sample.

We illustrate the procedure for computing a confidence interval with two examples using random samples from normally distributed populations—one involving estimation of a single population mean μ and one involving estimation of the difference between two population means $\mu_1 - \mu_2$. In each case, the appropriate confidence interval has the following general form:

$$\left(\begin{array}{c}\text{Point estimate of}\\ \text{the parameter}\end{array}\right) \pm \left[\left(\begin{array}{c}\text{Percentile of}\\ \text{the } t \text{ distribution}\end{array}\right)\left(\begin{array}{c}\text{Estimated standard}\\ \text{error of the estimate}\end{array}\right)\right] \qquad (3.10)$$

This general form also applies to confidence intervals for other parameters considered in the remainder of the text (e.g., those considered in multiple regression analysis).

■ **Example 3.1** Suppose that we have determined the Quantitative Graduate Record Examination (QGRE) scores for a random sample of nine student applicants to a certain graduate department in a university and that we have found $\bar{X} = 520$ and $S = 50$. If we want to estimate with 95% confidence the population mean QGRE score (μ) for all such applicants to the department, and we are willing to assume that the population of such scores from which our random sample was selected is approximately normally distributed, the confidence interval for μ is given by the general formula

$$\bar{X} \pm t_{n-1,1-\alpha/2}\left(\frac{S}{\sqrt{n}}\right) \qquad (3.11)$$

which gives the $100(1-\alpha)\%$ (small-sample) confidence interval for μ when σ is unknown. In our problem, $\alpha = 1 - .95 = .05$ and $n = 9$; therefore, by substituting the given information into (3.11), we obtain

$$520 \pm t_{8,0.975}\left(\frac{50}{\sqrt{9}}\right)$$

Since $t_{8,0.975} = 2.3060$, this formula becomes

$$520 \pm 2.3060\left(\frac{50}{\sqrt{9}}\right)$$

or

$$520 \pm 38.43$$

Our 95% confidence interval for μ is thus given by

(481.57, 558.43)

If we wanted to use this confidence interval to help determine whether 600 is a likely value for μ (i.e., if we were interested in making a decision about a specific value for μ), we would conclude that 600 is not a likely value, since it is not contained in the 95% confidence interval for μ just developed. This helps clarify the connection between estimation and hypothesis testing. ■

■ **Example 3.2** Suppose that we want to compare the change in health status of two groups of mental patients who are undergoing different forms of treatment for the same disorder. Suppose that we have a measure of change in health status based on a questionnaire given to each patient at two different times and that we are willing to assume this measure of change in health status is approximately normally distributed and has the same variance in the populations of patients from which we selected our independent random samples. The data obtained are summarized as follows:

Group 1: $n_1 = 15, \bar{X}_1 = 15.1, S_1 = 2.5$

Group 2: $n_2 = 15, \bar{X}_2 = 12.3, S_2 = 3.0$

where the underlying variable X denotes the change in health status between time 1 and time 2.

A 99% confidence interval for the true mean difference $(\mu_1 - \mu_2)$ in health status change between these two groups is given by the following formula, which assumes equal population variances (i.e., $\sigma_1^2 = \sigma_2^2$):

$$(\bar{X}_1 - \bar{X}_2) \pm t_{n_1+n_2-2,1-\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \qquad (3.12)$$

where S_p is the pooled standard deviation derived from S_p^2, the pooled sample variance given by (3.7). Here we have

$$S_p^2 = \frac{(15-1)(2.5)^2 + (15-1)(3.0)^2}{15 + 15 - 2} = 7.625$$

so

$$S_p = \sqrt{7.625} = 2.76$$

Since $\alpha = .01$, our percentile in (3.12) is given by $t_{28,0.995} = 2.7633$. So the 99% confidence interval for $\mu_1 - \mu_2$ is given by

$$(15.1 - 12.3) \pm 2.7633(2.76)\sqrt{\frac{1}{15} + \frac{1}{15}}$$

which reduces to

$$2.80 \pm 2.78$$

yielding the following 99% confidence interval for $\mu_1 - \mu_2$:

(0.02, 5.58)

Since the value 0 is not contained in this interval, we conclude that there is statistical evidence of a difference in health status change between the two groups. ■

3.6 Statistical Inference: Hypothesis Testing

Although closely related to confidence interval estimation, hypothesis testing has a slightly different orientation. When developing a confidence interval, we use our sample data to estimate what we think is a *likely* set of values for the parameter of interest. When performing a statistical test of a null hypothesis concerning a certain parameter, we use our sample data to *test* whether our estimated value for the parameter is *different enough* from the hypothesized value to support the conclusion that the null hypothesis is *unlikely* to be true.

The general procedure used in testing a statistical null hypothesis remains basically the same, regardless of the parameter being considered. This procedure (which we will illustrate by example) consists of the following seven steps:

1. Check the assumptions regarding the properties of the underlying variable(s) being measured that are needed to justify use of the testing procedure under consideration.
2. State the null hypothesis H_0 and the alternative hypothesis H_A.
3. Specify the significance level α.
4. Specify the test statistic to be used and its distribution under H_0.
5. Form the decision rule for rejecting or not rejecting H_0 (i.e., specify the rejection and nonrejection regions for the test, based on both H_A and α).
6. Compute the value of the test statistic from the observed data.
7. Draw conclusions regarding rejection or nonrejection of H_0.

■ **Example 3.3** Let us again consider the random sample of nine student applicants with mean QGRE score $\bar{X} = 520$ and standard deviation $S = 50$. The department chairperson suspects that, because of the declining reputation of the department, this year's applicants are not quite as good quantitatively as those from the previous five years for whom the average QGRE score was 600. If we assume that the population of QGRE scores from which our random sample has been selected is normally distributed, we can test the null hypothesis that the population mean score associated with this year's applicants is 600 versus the alternative hypothesis that it is less than 600. The *null hypothesis,* in mathematical terms, is H_0: $\mu = 600$, which asserts that the population mean μ for this year's applicants does not differ from what it has generally been in the past. The *alternative hypothesis* is stated as H_A: $\mu < 600$, which asserts that the QGRE scores, on average, have gotten worse.

We have thus far considered the first two steps of our testing procedure:

1. Assumptions: The variable QGRE score has a normal distribution, from which a random sample has been selected.
2. Hypotheses: H_0: $\mu = 600$; H_A: $\mu < 600$.

Our next step is to decide what error or probability we are willing to tolerate for incorrectly rejecting H_0 (i.e., making a Type I error, as discussed later in this chapter). We call this probability of making a Type I error the *significance level* α.[1]

We usually assign a value such as .1, .05, .025, or .01 to α. Suppose, for now, that we choose $\alpha = .025$. Then Step 3 is

3. Use $\alpha = .025$.

Step 4 requires us to specify the test statistic that will be used to test H_0. In this case, with H_0: $\mu = 600$, we have

4. $T = \dfrac{\bar{X} - 600}{S/\sqrt{9}} \sim t_8$ under H_0: $\mu = 600$.

Step 5 requires us to specify the decision rule that we will use to reject or not reject H_0. In determining this rule, we divide the possible values of T into two sets: the *rejection region* (or *critical region*), which consists of values of T for which we reject H_0; and the *nonrejection region,* which consists of those T-values for which we do not reject H_0. If our computed value of T falls in the rejection region, we conclude that the observed results deviate far enough from H_0 to cast considerable doubt on the validity of the null hypothesis.

In our example, we determine the critical region by choosing from t tables a point called the *critical point,* which defines the boundary between the nonrejection and rejection regions. The alternative hypothesis (H_A) informs the determination of the rejection region. Because our H_A states that the true mean is *less than* 600, an observed sample mean sufficiently less than 600 would be needed to support this alternative hypothesis. Accordingly, the test statistic T above would need to be negative, and thus all values of T in the rejection region would be negative. The value we choose is

$$-t_{8,0.975} = -2.306$$

in which case the probability that the test statistic takes a value of less than -2.306 under H_0 is exactly $\alpha = .025$, the significance level (Figure 3.8). We thus have the following decision rule:

5. Reject H_0 if $T = \dfrac{\bar{X} - 600}{S/\sqrt{9}} < -2.306$; do not reject H_0 otherwise.

[1] Two types of errors can be made when performing a statistical test. A Type II error occurs if we fail to reject H_0 when H_0 is actually false. We denote the probability of a Type II error as β and call $(1 - \beta)$ the *power* of the test. For a fixed sample size, α and β for a given test are inversely related; that is, lowering one has the effect of increasing the other. In general, the power of any statistical test can be raised by increasing the sample size. These issues are described further in Section 3.7.

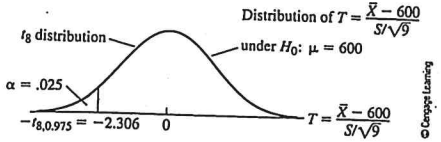


FIGURE 3.8 The critical point for Example 3.3

If H_A stated that $\mu \neq 600$ (a two-sided hypothesis), then either an extremely negative or an extremely positive value of T would support this alternative hypothesis. The rejection region would, therefore, include negative and positive values of T. Since the region would be two-tailed, α would be split between the two tails ($\alpha/2 = 0.0125$ here), yielding a decision rule of rejecting H_0 if $T < -t_{8,0.9875}$ or if $T > t_{8,0.9875}$.

Now we simply apply the decision rule to our data by computing the observed value of T. In our example, since $\bar{X} = 520$ and $S = 50$, our computed T is

6. $$T = \frac{\bar{X} - 600}{S/\sqrt{9}} = \frac{520 - 600}{50/3} = -4.8.$$

The last step is to make the decision about H_0 based on the rule given in Step 5:

7. Since $T = -4.8$, which lies below -2.306, we reject H_0 at significance level .025 and conclude that there is evidence that students currently applying to the department have QGRE scores that are, on average, *lower* than 600.

In addition to performing the procedure just described, we often want to compute a *P-value*, which quantifies *exactly how unusual the observed results would be if H_0 were true.* An equivalent way of describing the *P*-value is as follows: *The P-value gives the probability of obtaining a value of the test statistic that is at least as unfavorable to H_0 as the observed value, assuming that H_0 is true* (Figure 3.9).

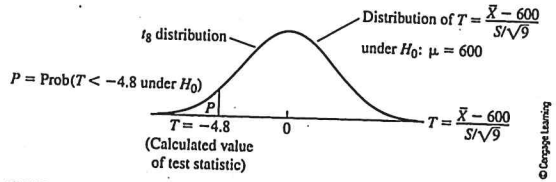


FIGURE 3.9 The *P*-value

To get an idea of the approximate size of the *P*-value in this example, our approach is to determine from the table of the distribution of T under H_0 the two percentiles that bracket the observed value of T. In this case, the two percentiles are

$$-t_{8,0.995} = -3.355 \quad \text{and} \quad -t_{8,0.9995} = -5.041$$

Since the observed value of T lies between these two values, we conclude that the area P we seek lies between the two areas corresponding to these two percentiles:

$$.0005 < P < .005$$

In interpreting this inequality, we observe that the *P*-value is *quite small*, indicating that we have observed a highly unusual result if H_0 is true. In fact, this *P*-value is so small as to lead us to reject H_0. Furthermore, the size of this *P*-value means that we would reject H_0 even for an α as small as .005.

For the general computation of a *P*-value, the appropriate *P*-value for a two-tailed test is twice that for the corresponding one-tailed test. If an investigator wants to draw conclusions about a test on the basis of the *P*-value (e.g., in lieu of specifying α a priori), the following guidelines are recommended:

1. If P is small (less than .01), reject H_0.
2. If P is large (greater than .1), do not reject H_0.
3. If $.01 < P < .1$, the significance is borderline, since we reject H_0 for $\alpha = .1$ but not for $\alpha = .01$.

Notice that, if we actually do specify α a priori, we reject H_0 when $P < \alpha$. ■

■ **Example 3.4** We now look at one more worked example about hypothesis testing—this time involving a comparison of two means, μ_1 and μ_2. Consider the following data on health status change, which were discussed earlier:

Group 1: $n_1 = 15, \bar{X}_1 = 15.1, S_1 = 2.5$
Group 2: $n_2 = 15, \bar{X}_2 = 12.3, S_2 = 3.0$
$(S_p = 2.76)$

Suppose that we want to test at significance level .01 whether the true average change in health status differs between the two groups. The steps required to perform this test are as follows:

1. Assumptions: We have independent random samples from two normally distributed populations. The population variances are assumed to be equal.
2. Hypotheses: H_0: $\mu_1 = \mu_2$; H_A: $\mu_1 \neq \mu_2$.
3. Use $\alpha = .01$.
4. $$T = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{28} \text{ under } H_0.$$

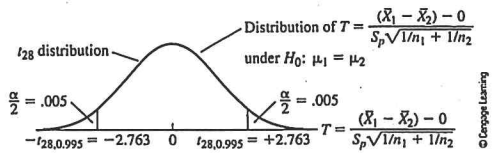


FIGURE 3.10 Critical region for the health status change example

5. Reject H_0 if $|T| \geq t_{28,0.995} = 2.763$; do not reject H_0 otherwise (Figure 3.10).

6. $$T = \frac{(\bar{X}_1 - \bar{X}_2) - 0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{15.1 - 12.3}{2.76\sqrt{\frac{1}{15} + \frac{1}{15}}} = 2.78.$$

7. Since $T = 2.78$ exceeds $t_{28,0.995} = 2.763$, we reject H_0 at $\alpha = .01$ and conclude that there is evidence the true average change in health status differs between the two groups.

The P-value for this test is given by the shaded area in Figure 3.11. For the t distribution with 28 degrees of freedom, we find that $t_{28,0.995} = 2.763$ and $t_{28,0.9995} = 3.674$. Thus, $P/2$ is given by the inequality

$$1 - .9995 < \frac{P}{2} < 1 - .995$$

so

$$.001 < P < .01$$

■

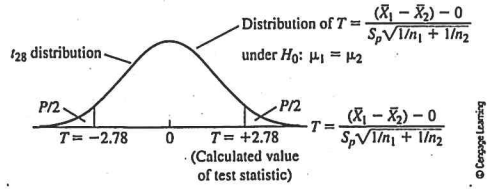


FIGURE 3.11 P-value for the health status change example

3.7 Error Rates, Power, and Sample Size

Table 3.1 summarizes the decisions that result from hypothesis testing. If the true state of nature is that the null hypothesis is true and if the decision is made that the null hypothesis is true, then a correct decision has been made. Similarly, if the true state of nature is that the alternative hypothesis is true and if the decision is made that the alternative is true, then a correct decision has been made. On the other hand, if the true state of nature is that the null hypothesis is true but the decision is made to choose the alternative, then a false positive error (commonly referred to as a *Type I error*) has been made. And if the true state of nature supports the alternative hypothesis but the decision is made that the null hypothesis is true, then a false negative error (commonly referred to as a *Type II error*) has been made.

TABLE 3.1 Outcomes of hypothesis testing

Hypothesis Chosen	True State of Nature	
	H_0	H_A
H_0	Correct decision	False negative decision (Type II error)
H_A	False positive decision (Type I error)	Correct decision

© Cengage Learning

Table 3.2 summarizes the probabilities associated with the outcomes of hypothesis testing just described. If the true state of nature corresponds to the null hypothesis but the alternative hypothesis is chosen, then a Type I error has been made, with probability denoted by the symbol α. Hence, the probability of making a correct choice of H_0 given that H_0 is true must be $1 - \alpha$. In turn, if the actual state of nature is that the alternative hypothesis is true but the null hypothesis is chosen, then a Type II error has occurred, with probability denoted by β. In turn, $1 - \beta$ is the probability of choosing the alternative hypothesis given that it is true, and this probability is often called the *power of the test*.

TABLE 3.2 Probabilities of outcomes of hypothesis testing

Hypothesis Chosen	True State of Nature	
	H_0	H_A
H_0	$1 - \alpha$	β
H_A	α	$1 - \beta$

© Cengage Learning

When we design a research study, we would like to use statistical tests for which both α and β are small (i.e., for which there is a small chance of making either a Type I or a Type II error). For a given α, we can sometimes determine the sample size required in the study to ensure that β is no larger than some desired value for a particular alternative hypothesis of interest. Such a design consideration generally involves the use of a *sample size formula* pertinent to the research question(s). This formula usually requires the researcher to make educated guesses about the values of some of the unknown parameters to be estimated in the study (see Cohen 1977; Muller and Peterson 1984; Kupper and Hafner 1989).

For example, the classical sample size formula used for a one-sided test of H_0: $\mu_1 = \mu_2$ versus H_0: $\mu_2 > \mu_1$, when a random sample of size n is selected from each of two normal populations with common variance σ^2, is as follows:

$$n \geq \frac{2(Z_{1-\alpha} + Z_{1-\beta})^2\sigma^2}{\Delta^2}$$

For chosen values of α, β, and σ^2, this formula provides the minimum sample size n required to detect a specified difference $\Delta = \mu_2 - \mu_1$ between μ_1 and μ_2 (i.e., to reject H_0: $\mu_2 - \mu_1 = 0$ in favor of H_A: $\mu_2 - \mu_1 = \Delta > 0$ with power $1 - \beta$). Thus, in addition to picking α and β, the researcher must specify the size of the population variance σ^2 and specify the difference Δ to be detected. An educated guess about the value of the unknown parameter σ^2 can sometimes be made by using information obtained from related research studies. To specify Δ intelligently, the researcher has to decide on the smallest population mean difference ($\mu_2 - \mu_1$) that is practically (as opposed to statistically) meaningful for the study.

For a fixed sample size, α and β are inversely related in the following sense, illustrated in Figure 3.12. If one tries to guard against making a Type I error by choosing a small rejection region, the nonrejection region (and hence β) will be large. Conversely, protecting against a Type II error necessitates using a large rejection region, leading to a large value for α. Increasing the sample size generally decreases the standard deviation of the test statistic (standard error) and accordingly decreases β; of course, α remains unaffected. A detailed discussion about power and sample size determination for statistical methods taught in this text is provided in Chapter 27.

It is common practice to conduct several statistical tests using the same data set. If such a data-set-specific series of tests is performed and each test is based on a size α rejection region,

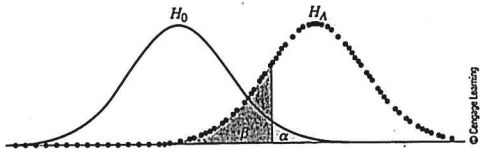


FIGURE 3.12 Distributions of a test statistic under the null (H_0) and alternative (H_A) hypotheses, displaying the relationship between α and β

the probability of making at least one Type I error will be much larger than α. This multiple-testing problem is pervasive and bothersome. One simple—but not optimal—method for addressing this problem is to employ the so-called *Bonferroni correction.* For example, if k tests are to be conducted and if the overall Type I error rate (i.e., the probability of making at least one Type I error in k tests) is to be no more than α, then a rule of thumb is to conduct *each individual* test at a Type I error rate of α/k.

This simple adjustment ensures that the overall Type I error rate will (at least approximately) be no larger than α. In many situations, however, this correction leads to such a small rejection region for each individual test that the power of each test may be too low to detect important deviations from the null hypotheses being tested. Resolving this antagonism between Type I and Type II error rates requires a conscientious study design and carefully considered error rates for planned analyses.

Problems

1. a. Give two examples of discrete random variables.
   b. Give two examples of continuous random variables.
2. Name the four levels of measurement, and give an example of a variable at each level.
3. Assume that Z is a normal random variable with mean 0 and variance 1.
   a. $P(Z \geq -1) = ?$
   b. $P(Z \leq ?) = .20$
4. a. $P(\chi^2_7 \geq ?) = .01$
   b. $P(\chi^2_{12} \leq 14) = ?$
5. a. $P(T_{13} \leq ?) = .10$
   b. $P(|T_{28}| \geq 2.05) = ?$
6. a. $P(F_{6,24} \geq ?) = .05$
   b. $P(F_{5,40} \geq 2.9) = ?$
7. What are the (a) mean, (b) median, and (c) mode of the standard normal distribution?
8. An $F_{1,\nu}$ random variable can be thought of as the square of what kind of random variable?
9. Find the (a) mean, (b) median, and (c) variance for the following set of scores:

   {0, 2, 5, 6, 3, 3, 3, 1, 4, 3}

   d. Find the set of Z scores for the data.
10. Which of the following statements about descriptive statistics is correct?
    a. *All* of the data are used to compute the median.
    b. The mean should be preferred to the median as a measure of central tendency if the data are noticeably skewed.
    c. The variance has the same units of measurement as the original observations.
    d. The variance can never be 0.
    e. The variance is like an average of squared deviations from the mean.

11. Suppose that the weight W of male patients registered at a diet clinic has the normal distribution with mean 190 and variance 100.

a. For a random sample of patients of size $n = 25$, the expression $P(\overline{W} < 180)$, in which $\overline{W}$ denotes the sample mean weight, is equivalent to saying $P(Z > ?)$. [*Note:* Z is a standard normal random variable.]

b. Find an interval (a, b) such that $P(a < \overline{W} < b) = .80$ for the same random sample in part (a).

12. The limits of a 95% confidence interval for the mean μ of a normal population with unknown variance are found by adding to and subtracting from the sample mean a certain multiple of the estimated standard error of the sample mean. If the sample size on which this confidence interval is based is 28, the *multiple* referred to in the previous sentence is the number _____ .

13. A random sample of 32 persons attending a certain diet clinic was found to have lost (over a three-week period) an average of 30 pounds, with a sample standard deviation of 11. For these data, a 99% confidence interval for the true mean weight loss by all patients attending the clinic would have the limits (?, ?).

14. From two normal populations assumed to have the same variance, independent random samples of sizes 15 and 19 were drawn. The first sample (with $n_1 = 15$) yielded mean and standard deviation 111.6 and 9.5, respectively, while the second sample ($n_2 = 19$) gave mean and standard deviation 100.9 and 11.5, respectively. The estimated standard error of the difference in sample means is _____ .

15. For the data of Problem 14, suppose that a test of H_0: $\mu_1 = \mu_2$ versus H_A: $\mu_1 > \mu_2$ yielded a computed value of the appropriate test statistic equal to 2.55.

a. What conclusions should be drawn for $\alpha = .05$?

b. What conclusions should be drawn for $\alpha = .01$?

16. Test the null hypothesis that the true population average body weight is the same for two independent diagnosis groups from one hospital versus the alternative hypothesis that these two population averages are different, using the following data:

Diagnosis group 1 data: {132, 145, 124, 122, 165, 144, 151}

Diagnosis group 2 data: {141, 139, 172, 131, 150, 125}

You may assume that the populations from which the data come are each normally distributed, with equal population variances. What conclusion should be drawn, with $\alpha = .05$?

17. Independent random samples are drawn from two normal populations, which are assumed to have the same variance. One sample (of size 5) yields mean 86.4 and standard deviation 8.0, and the other sample (of size 7) has mean 78.6 and standard deviation 10. The limits of a 99% confidence interval for the difference in population means are found by adding to and subtracting from the difference in sample means a certain multiple of the estimated standard error of this difference. This *multiple* is the number _____ .

18. If a 99% confidence interval for $\mu_1 - \mu_2$ is 4.8 to 9.2, which of the following conclusions can be drawn *based on this interval?*

a. Do not reject H_0: $\mu_1 = \mu_2$ at $\alpha = .05$ if the alternative is H_A: $\mu_1 \neq \mu_2$.
b. Reject H_0: $\mu_1 = \mu_2$ at $\alpha = .01$ if the alternative is H_A: $\mu_1 \neq \mu_2$.
c. Reject H_0: $\mu_1 = \mu_2$ at $\alpha = .01$ if the alternative is H_A: $\mu_1 < \mu_2$.
d. Do not reject H_0: $\mu_1 = \mu_2$ at $\alpha = .01$ if the alternative is H_A: $\mu_1 \neq \mu_2$.
e. Do not reject H_0: $\mu_1 = \mu_2 + 3$ at $\alpha = .01$ if the alternative is H_A: $\mu_1 \neq \mu_2 + 3$.

19. Assume that we gather data, compute a T, and reject the null hypothesis. If, in fact, the null hypothesis is true, we have made (**a**) _____ . If the null hypothesis is false, we have made (**b**) _____ . Assume instead that our data lead us to not reject the null hypothesis. If, in fact, the null hypothesis is true, we have made (**c**) _____ . If the null hypothesis is false, we have made (**d**) _____ .

20. Suppose that the critical region for a certain test of hypothesis is of the form $|T| \geq 2.5$ and that the computed value of T from the data is -2.75. Which, if any, of the following statements is correct?

a. H_0 should be rejected.
b. The significance level α is the probability that, under H_0, T is either greater than 2.75 or less than -2.75.
c. The nonrejection region is given by $-3.5 < T < 3.5$.
d. The nonrejection region consists of values of T above 3.5 or below -3.5.
e. The P-value of this test is given by the area to the right of $T = 3.5$ for the distribution of T under H_0.

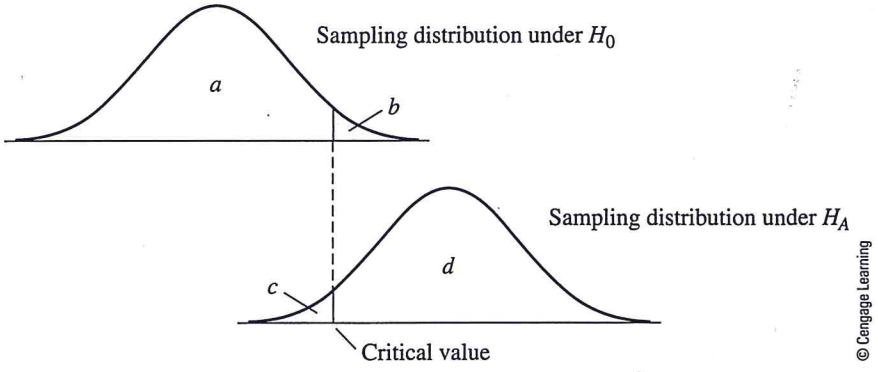
21. Suppose that $\bar{X}_1 = 125.2$ and $\bar{X}_2 = 125.4$ are the mean systolic blood pressures for two random samples of workers from different plants in the same industry. Suppose, further, that a test of H_0: $\mu_1 = \mu_2$ using these samples is rejected for $\alpha = .001$. Which of the following conclusions is most reasonable?

a. There is a meaningful difference (clinically speaking) in population means but not a statistically significant difference.
b. The difference in population means is both statistically and meaningfully significant.
c. There is a statistically significant difference but not a meaningfully significant difference in population means.
d. There is neither a statistically significant nor a meaningfully significant difference in population means.
e. The sample sizes used must have been quite small.

22. The choice of an alternative hypothesis (H_A) should depend primarily on (choose all that apply)

a. the data obtained from the study.
b. what the investigator is interested in determining.
c. the critical region.
d. the significance level.
e. the power of the test.

23. For each of the areas in the accompanying figure, labeled *a*, *b*, *c*, and *d*, select an answer from the following: α, $1 - \alpha$, β, $1 - \beta$.



24. Suppose that H_0: $\mu_1 = \mu_2$ is the null hypothesis and that $.10 < P < .25$. What is the most appropriate conclusion?

25. Suppose that H_0: $\mu_1 = \mu_2$ is the null hypothesis and that $.005 < P < .01$. Which of the following conclusions is most appropriate?

a. Do not reject H_0 because P is small.
b. Reject H_0 because P is small.
c. Do not reject H_0 because P is large.
d. Reject H_0 because P is large.
e. Do not reject H_0 at $\alpha = .01$.

References

Cohen, J. 1977. *Statistical Power Analysis for the Behavioral Sciences,* 2d ed. New York: Academic Press.

Kupper, L. L., and Hafner, K. B. 1989. "How Appropriate Are Popular Sample Size Formulas?" *The American Statistician* 43(2): 101–5.

Muller, K. E., and Peterson, B. L. 1984. "Practical Methods for Computing Power in Testing the Multivariate General Linear Hypothesis." *Computational Statistics and Data Analysis* 2: 143–58.