APPENDIX A

Basic Statistical Concepts[1]

Descriptive statistical methods were introduced in Chapter 2, and other relevant concepts were explained as we proceeded through the book. However, there are several foundational topics that need to be understood by readers, even though they have not been discussed explicitly. The purpose of this appendix is to provide an overview of these key topics so that knowledge of the basics can be refreshed as needed.

We cover the following topics:

1. Expected values
2. The normal distribution
3. Sampling distributions
4. Point and interval estimation
5. Testing hypotheses
6. A test for normality

1 This appendix is intended to provide a brief review of key topics not discussed explicitly in the main text. A reader needing a detailed treatment of these topics should consult a standard statistics textbook, such as Anderson et al. (2014).

A.1 Expected Values

We use sample statistics based on the observed data to estimate unknown population parameters such as the mean and variance. When we select different samples, we expect the values of the statistics to differ, but the sampling process will behave in accordance with the sampling distributions described in Section A.3. Consider a variable of interest, which we denote by X. We refer to X as a *random variable*, since it will have both a set of possible values and probabilities associated with those values.

■ Example A.1

Let X denote the number of heads (H) recorded in two tosses of a fair coin. Then X may take on the values {0, 1, 2} with probabilities {0.25, 0.50, 0.25}, corresponding, respectively, to the possible outcomes TT, TH or HT, and HH. ■

As always, the probabilities are nonnegative and sum to 1.00. In general, for a random variable that takes discrete values, we may define the set of possible values for X as $\{x_0, x_1, \ldots, x_N\}$ with associated probabilities $\{p_0, p_1, \ldots, p_N\}$. In a theoretical setting, N may be either finite or infinite. We then define the population mean and variance as the *expected values*, based upon those probabilities:

EXPECTED VALUES: MEAN, VARIANCE, AND STANDARD DEVIATION

For a discrete random variable, the *mean*, or *expected value*, of X is

$$E(X) = \sum_{i=0}^{N} p_i x_i.$$

The mean is usually represented by μ (mu).

The variance of X is

$$V(X) = \sum_{i=0}^{N} p_i (x_i - \mu)^2.$$

The variance is usually represented by σ^2 (sigma squared). Its square root is the *standard deviation*:

$$SD(X) = \sigma.$$

■ Example A.2

The mean, variance, and standard deviation for the random variable defined in Example A1 are as follows:

$$E(X) = \mu = 0 \times 0.25 + 1 \times 0.50 + 2 \times 0.25 = 1.00,$$

$$V(X) = \sigma^2 = (0 - 1)^2 \times 0.25 + (1 - 1)^2 \times 0.50 + (2 - 1)^2 \times 0.25 = 0.50,$$

$$SD(X) = \sqrt{0.50} = 0.707.$$

The reader is encouraged to replicate the steps in Examples A.1 and A.2 for a single toss of the coin and for three tosses of the coin. Observe that in these cases the expected values (0.5 and 1.5, respectively) are not observable in a single set of tosses. In general, the population mean may not correspond to any of the possible outcomes. ■

Continuous Variables Our description so far assumes that the random variable is *discrete*; that is, it may take on only one of a finite (or possibly countable) set of values. However, for many theoretical purposes, it is convenient to think of the random variable as being continuous (e.g., time, height), even though such variables are always measured to a finite accuracy (nearest minute, nearest inch). In these circumstances, we define the random variable in terms of its set of possible values and a *Probability Density Function* (*PDF*).

CONTINUOUS RANDOM VARIABLE

A continuous random variable defined on the interval $[a, b]$ has probability density function (*PDF*) that equals

$$f(x) \text{ if } a \le x \le b,$$
$$0 \text{ if } x < a \text{ or } x > b.$$

The probability that the observed value of the random variable is less than or equal to some number, say x_0, is given by

$$P(x_0) = \int_a^{x_0} f(x)dx.$$

The mean, variance, and standard deviation are respectively given by

$$E(X) = \mu = \int_a^b xf(x)dx,$$

$$V(X) = \sigma^2 = \int_a^b (x - \mu)^2 f(x)dx.$$

$$SD(X) = \sigma = \sqrt{V(X)}.$$

In this book, we do not need to compute population means and variances explicitly, but it is important to understand these concepts because they are fundamental to the whole process of statistical inference.

The expected value and the variance have some simple properties that we employ in the text:

A.1.1 For two random variables X and Y, $E(aX + bY) = aE(X) + bE(Y)$.

That is, the expectation of a sum of random variables (such as the mean of a sample) equals the sum of the expectations of the variables.

A.1.2 The variance of X is given by $\text{var(X)} = E(X^2) - (E(X))^2$.

A.1.3 For two random variables X and Y, $V(aX + bY) = a^2V(X) + b^2V(Y) + 2abC(X, Y)$ where $C(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$ denotes the covariance between X and Y. This result follows from A.1.2. When the variables are independent the covariance is zero.

A.2 The Normal Distribution

The normal distribution, also known as the Gaussian distribution, lies at the heart of modern statistics. The reason for its importance is the Central Limit Theorem, which we discuss in Section A.3.

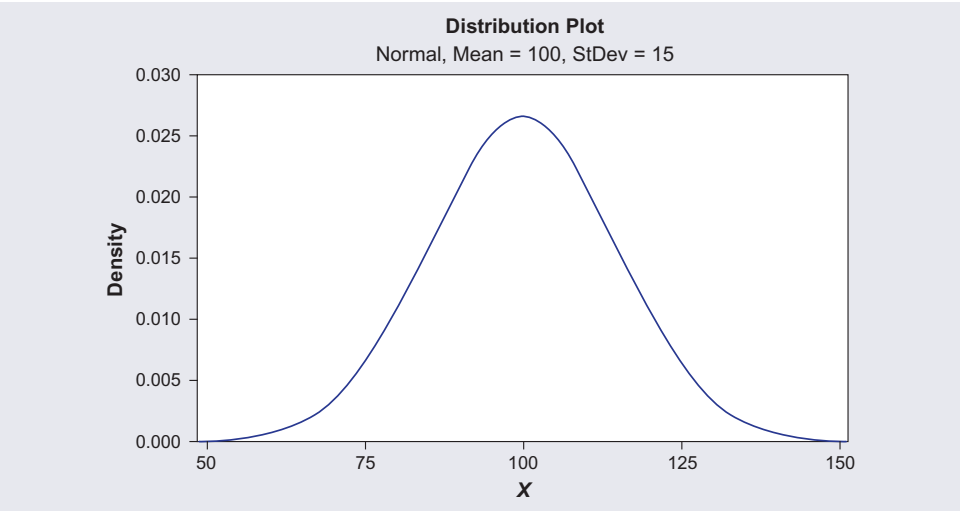
The normal distribution is the familiar "bell-shaped curve" shown in Figure A.1. As can be seen by inspection, the distribution is symmetric and unimodal, so the mean, median, and mode are all equal and are typically represented by the Greek letter μ (mu). The standard deviation is represented by σ (sigma). If we represent the random variable by X, we may make the following probability statements for any values of μ and σ:

$$P(\mu - \sigma \le X \le \mu + \sigma) = 0.68,$$

$$P(\mu - 2\sigma \le X \le \mu + 2\sigma) = 0.95,$$

$$P(\mu - 3\sigma \le X \le \mu + 3\sigma) = 0.997.$$

Figure A.1 Normal Distribution with Mean = 100 and Standard Deviation = 15



Because the probabilities are the same whatever the mean and standard deviation, we can define the standard normal distribution in terms of

$$Z = \frac{X - \mu}{\sigma}.$$

By construction, Z has mean 0 and standard deviation 1. Thus, it becomes possible to evaluate Z from X and then compute the desired probability directly in terms of Z, so that only a single numerical function (for computer usage) or a single set of tables (for manual use) is required. Either approach yields $P(Z \le z_0)$ for some value z_0. Figure A.2 illustrates the probability when $z_0 = 1.5$. If we wish to compute the value of Z for a given probability P, then the inverse function is required, or the same tables can be used, albeit with rather more effort. We will often be interested in the tail areas of the curve, as shown in Figure A.3 with 0.025 in each tail ($z_0 = 1.960$).

Figure A.2 Probability that $Z < 1.5$ for the Standard Normal Distribution

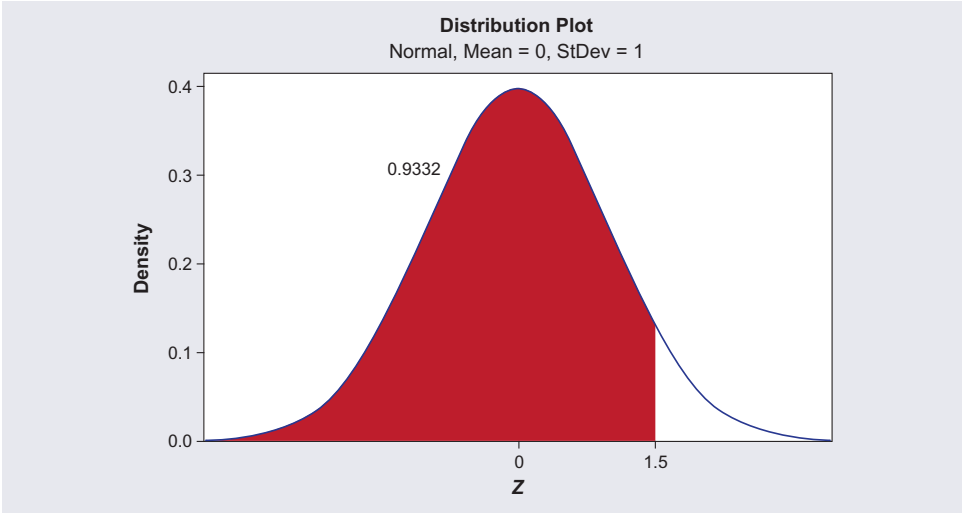


Figure A.3 Tail Area Probability, Totaling 0.05 for the Standard Normal Distribution

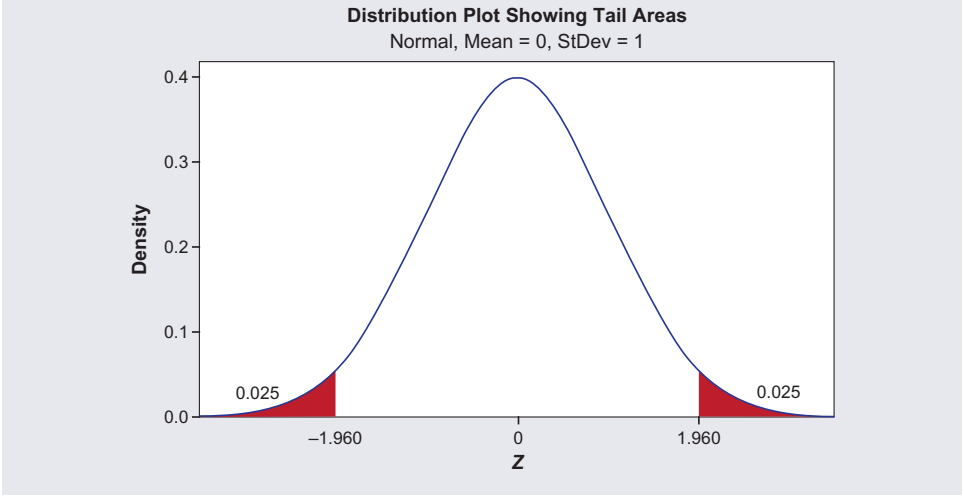


Table A.1 provides tail area probabilities that are widely used throughout the book.

Table A.1 Tail Area Probabilities in this Book

Z-Value	One-Tailed Probability	Two-Tailed Probability
1.282	0.100	0.200
1.645	0.050	0.100
1.960	0.025	0.050
2.326	0.010	0.020
2.575	0.005	0.010

Other values can be computed most easily by using the Excel function NORM.DIST (X, mean, standard deviation, 1). For example, NORM.DIST(1.5,0,1,1) delivers the probability of observing a value less than 1.5 when the distribution is normal with mean 0 and standard deviation 1. This probability equals 0.933.

The value for which $P(X < \text{Value} \mid \text{mean, standard deviation})$ equals a chosen probability P can be calculated equally easily by using NORM.INV(P, mean, standard deviation). For example, for a probability $P = 0.9$ with mean 4 and standard deviation 2, NORM.INV(0.9, 4, 2) gives a value of 6.56 (i.e., $P(X < 6.56) = 0.9$).

A.3 Sampling Distributions

The fundamental purpose of the discipline of statistics is to make inferences—that is, to make statements about a population upon the basis of properties of a sample. In surveys of human populations, the steps are *requirements* that are quite transparent. We need some kind of a sampling frame (a list of all the members of the population), and we then draw a random sample of individuals from that frame. In practice, of course, things can be difficult: Just think about the myriad problems associated with trying to draw a random sample of New York Yankee fans from the population (how defined?) of such misguided individuals. More seriously, if we were considering launching a new product, how might we establish the population of those who might choose to use it?

When we turn to model-building activities, which form the basis of this text, the issues are more abstract. Our interest focuses on the discrepancies between the observed data and the values predicted by the hypothetical model, and the framework for random sampling must be developed in the context of that model. For example, suppose we wish to examine the quarterly earnings of a major company quoted on the New York Stock Exchange. We are looking at a single company over a single time period. The concept of random sampling cannot be based upon the list type of sampling approach used in surveys. Instead, we must build a conceptual framework based upon plausible (and ultimately testable) assumptions. This idea is developed more fully in Appendix 1A in Chapter 1 of the book.

The set of basic assumptions we use throughout this book relate to the error terms:

Random error = difference between observed value and value predicted by the model.

We denote the random error by ε (Greek epsilon), the random variable of interest by Y, and the value predicted by the model by μ_Y so that

$$\varepsilon = Y - \mu_Y.$$

If we have a sample of n observations, the four basic assumptions relating to the n error terms are as follows:

1. The errors have zero means, written as $E(\varepsilon) = 0$.
2. The errors have equal variances $V(\varepsilon) = \sigma^2$.
3. The errors are independent of each other.
4. [Builds on the first three] The errors are randomly (hence independently) selected from a *normal* distribution with mean 0 and variance σ^2. This assumption is sometimes written in the compact notational form $\varepsilon \sim NID(0, \sigma^2)$, which states that the random errors are *N*ormally and *I*ndependently *D*istributed with mean zero and common variance σ^2.

The key results that follow from these assumptions can be illustrated by the case when n random variables $\{Y_i, i = 1, 2, \ldots, n\}$ have the same mean μ:

1. The mean of the n observations, $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$, has mean $E(\bar{Y}) = \mu$ and variance $V(\bar{Y}) = \frac{\sigma^2}{n}$.
2. If each Y_i is normally distributed, then $\bar{Y}$ is normally distributed.
3. [*Central Limit Theorem*, *CLT*] Even when the random variables are not normally distributed, the distribution of $\bar{Y}$ approaches the normal distribution as n increases.

The CLT would be of limited value if we did not know how large a sample size was needed for the approximation to take effect. Many numerical studies have been undertaken over the years, and the following guidelines are generally accepted:

a. The CLT can be applied if the common distribution of the Y_i is approximately symmetric and $n \geq 15$, or
b. The CLT can be applied, provided that $n \geq 30$.

A.4 Estimation

The process of statistical inference often reduces to using a sample *statistic* to estimate a population *parameter* (e.g., using the sample mean to estimate the population mean). Frequently, there will be a variety of statistics that we can use, so it is valuable to identify certain properties that the chosen statistic should satisfy.

We assume that we have available a random sample of size n and that we use the observations to create a sample statistic that we denote by U_n, which is some function of those observations (e.g., U_n might be the sample mean or the sample median). U_n is used to estimate some unknown population parameter that we denote by θ. The following properties are desirable:

a. *Unbiasedness:* U_n is an unbiased estimator for θ if $E(U_n) = \theta$. For a particular sample, the value of U_n may be above or below the (unknown) value θ, but there is no systematic bias. If U_n is biased (not unbiased), we define the bias to be $B_n = E(U_n) - \theta$. (Think of shooting arrows at a target; not many hit the bull's-eye, but your aim is unbiased if there is no tendency to shoot high or low, or to the left or right. If you shoot systematically to one corner of the target, your aim is biased.)
b. *Consistency:* U_n is a consistent estimator for θ if the probability of U_n being within some fixed small distance of θ approaches 1.0. (An analogy, albeit imperfect, would be that your aim gets better and better as you shoot more arrows.)
c. *Efficiency:* U_n is an efficient estimator for θ if we cannot find a better statistic.

A commonly used performance measure is the *Mean Square Error* (*MSE*), defined as

$$\begin{aligned} MSE(U_n) &= E[(U_n - \theta)^2] \\ &= V(U_n) + B^2(U_n). \end{aligned}$$

If U_n is unbiased, the *MSE* reduces to the variance. If U_n is consistent, the *MSE* will approach zero as n increases. Finally, if U_n is efficient, no other statistic will have a smaller *MSE* than does U_n.

We stress that there are other measures of efficiency than the *MSE*, but we make extensive use of least squares estimation in the book, so the *MSE* criterion is appropriate.

■ Example A.3 Estimation of the population mean

Suppose we have a random sample of n observations from a distribution with mean μ and variance σ^2, as in Section A.3. Then, from result 1 of that section $E(\bar{Y}) = \mu$, so $\bar{Y}$ is an unbiased estimator for μ. Because $\bar{Y}$ is unbiased and has variance $V(\bar{Y}) = \frac{\sigma^2}{n}$, the *MSE* goes to zero as n increases, so $\bar{Y}$ is a consistent estimator for μ. Finally, it may be shown that the *MSE* is minimized when $U_n = \bar{Y}$, so $\bar{Y}$ is an efficient estimator for μ. ■

A.4.1 Confidence Intervals

Given that any estimate is subject to uncertainty, we may construct a confidence interval to indicate the extent of that uncertainty. Under assumptions 1–4, the $100(1 - \alpha)$ percent confidence interval for θ takes the form

$$U_n - z_{\alpha/2}\sqrt{V(U_n)} \le \theta \le U_n + z_{\alpha/2}\sqrt{V(U_n)}.$$

In the simplest case, when the variance is known, $z_{\alpha/2}$ is determined from the normal distribution and corresponds to the cutoff value determined by the tail area probability of the normal distribution, as given in Table A.1. When the variance is unknown, we use the percentage points of Student's t-distribution.

The interpretation of the confidence interval is as follows:

> When an interval is constructed by this process, $100(1 - \alpha)$ percent of the time the interval will include the value of the unknown parameter θ.

■ Example A.4 Confidence interval for the population mean

Given a random sample of size n, as in Example A.3, we seek to set up a 95% confidence interval.

From Table A.1, $z_{\alpha/2} = 1.96$, and from result 1, $V(\bar{Y}) = \sigma^2/n$ so that the confidence interval becomes

$$\bar{Y} - 1.96\sqrt{\sigma^2/n} \le \mu \le \bar{Y} + 1.96\sqrt{\sigma^2/n}.$$

If we are now informed that $n = 12$, $\sigma = 20$, and, finally, the sample mean is 50, then the confidence interval is

$$50 - 1.96\sqrt{400/12} \le \mu \le 50 + 1.96\sqrt{400/12},$$

which reduces to $20 \le \mu \le 70$.

Under the same conditions, when the variance is unknown, we use the sample standard deviation to estimate σ and replace the z-value by the corresponding value from Student's t-distribution with $(n - 1)$ degrees of freedom. ■

A.4.2 Prediction Intervals

A *prediction interval* (*PI*) is a probability statement about a future observation or event. The interpretation of a prediction interval is as follows:

> A prediction interval will include the value of the future observation with probability $(1 - \alpha)$

Although the mode of construction is similar to that of a confidence interval, the interpretation is different because the *PI* is a probability statement. Under assumptions 1–4, the

100(1 – α) percent prediction interval for future observation Y_{n-1} about a point forecast, say F_{n-1}, takes the form

$$F_{n+1} - z_{\alpha/2}\sqrt{V(F_{n+1})} \le Y_{n+1} \le F_{n+1} + z_{\alpha/2}\sqrt{V(F_{n+1})}.$$

As before, when the variance is known, $z_{\alpha/2}$ may be determined from the normal distribution, and when the variance is unknown, we may use the percentage points of the Student's *t*-distribution. However, it must be stressed that a prediction interval relates to a single observation so that, no matter how long the series, the Central Limit Theorem cannot be used to justify the normality assumption. For this reason, some forecasters prefer to use empirically-based prediction intervals; see Section 2.8.2 of the book.

A.5 Tests of Hypotheses

A statistical test is based upon two mutually exclusive hypotheses (i.e., two conflicting views; only one can be correct). One hypothesis is designated as the *null hypothesis* (or benchmark, denoted by H_0), and the other is the *alternative hypothesis* (denoted by H_A). The null hypothesis often takes the form of assuming "no change," The two hypotheses cover all possibilities between them.

Upon the basis of the structure of the hypotheses, we identify a test statistic and formulate a decision rule that determines whether we should go with the null (do not reject H_0) or the alternative (reject H_0). Only after these preliminaries are complete do we examine the data and draw a conclusion.

■ Example A.5 Testing the coefficient of the slope in a simple linear regression equation

The simple linear regression model is defined as

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

The most common null hypothesis we wish to test is whether the slope (denoted by β_1) is zero; a zero slope implies that there is no linear relationship between the dependent variable and the explanatory variable. (See Section 7.1.1 for definitions of these terms.) That is, we formulate the hypotheses as

$$H_0 : \beta_1 = 0 \quad \text{versus} \quad H_A : \beta_1 \neq 0.$$

The test is formulated under the assumption that the null hypothesis is true, so the form of the test statistic, denoted by *t*, is

$$t = \frac{b_1 - \beta_1(\text{under } H_0)}{SE(b_1)} = \frac{b_1}{SE(b_1)}.$$

$SE(b_1)$ denotes the standard error of the sample slope. (Full details appear in Section 7.6; here we are concerned only with the conceptual framework.) ■

When the null hypothesis is true, the test statistic has a sampling distribution that can be derived under appropriate assumptions (see Section 7.5.1). We then use this sampling distribution to formulate the following decision rule:

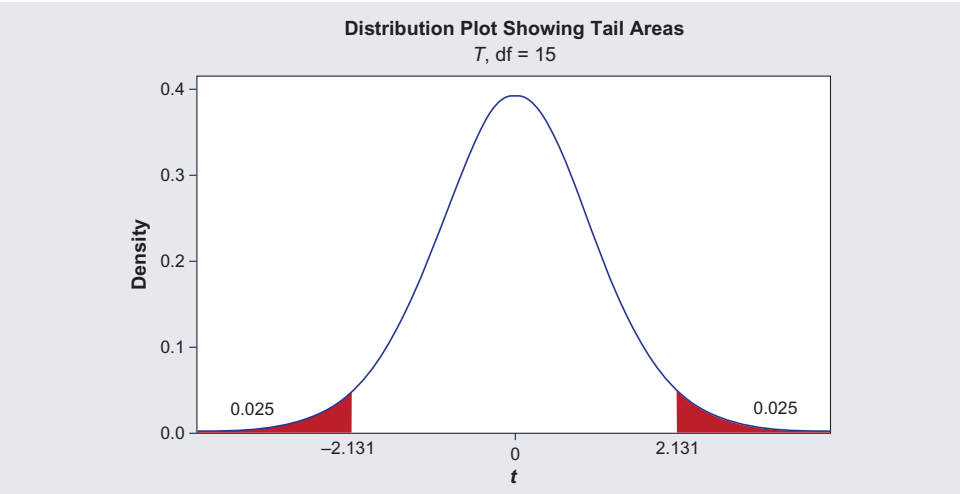
> Reject H_0 if the observed value of *t* exceeds some *critical value* (in absolute terms); do not reject H_0 otherwise. Values of *t* that exceed the critical value are said to form the *critical region*. This formulation of the rule is operationally equivalent to accepting H_0 as the working hypothesis, if the evidence we have available does not contradict the null hypothesis.

We use the absolute value because we are equally interested in both negative and positive values of the slope. The critical value is determined so that the probability of rejecting the null hypothesis when it is true is set at some preassigned level (known as the *significance level* and denoted by α). Combining these statements, we have the algebraic form

Reject H_0 if $|t_{observed}| > t_{critical}$, where $t_{critical}$ is determined so that $P(|t_{observed}| > |H_0 \text{ true}) = \alpha$.
If $|t_{observed}| < t_{critical}$, do not reject H_0.

Figure A.4 (also Figure 7.11(A) in the main text) illustrates how the critical value would be chosen in this case when the sample size is 17 (or 15 degrees of freedom) and $\alpha = 0.05$. In this example, we would reject H_0 if the observed values of t exceeded 2.131. One-sided tests are also possible when the critical region is defined in only one tail of the distribution.

Figure A.4 The *t*-Distribution with 15 Degrees of Freedom and Shaded Areas Corresponding to $\alpha = 0.05$ for a Two-Tailed Test



A.5.1 *P*-Values

As outlined, the testing procedure requires the user to specify the significance level first and then compute the critical value (historically from tables, now by computer). An equivalent procedure that is much more convenient in the computer age is the use of the *observed significance level*, denoted by P. When H_0 is true, the P-value denotes the probability of observing an absolute value of the test statistic at least as extreme as that actually observed. Suppose the observed value was –2.8; then the P-value is 0.0067, as shown in Figure A.5.

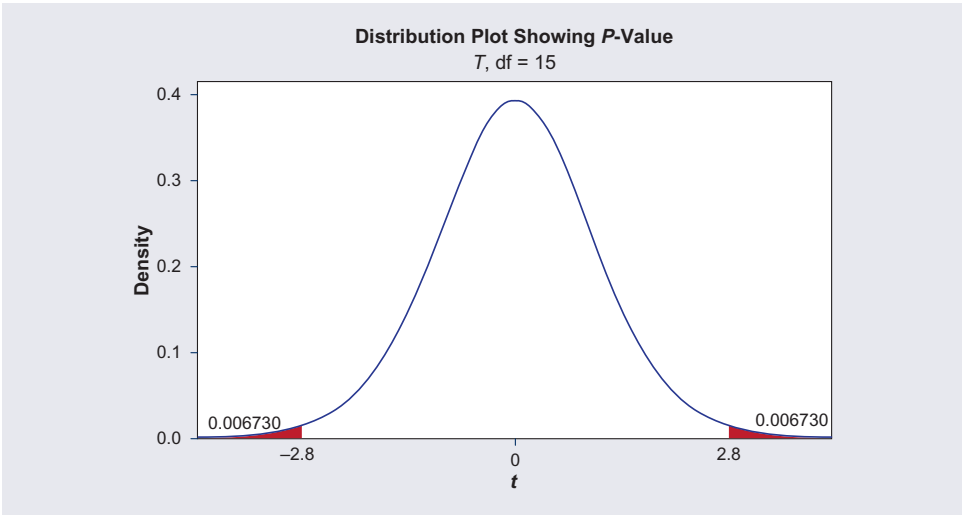
The decision rule now becomes

Reject H_0 if $P < \alpha$; otherwise do not reject H_0. (A1) (A1)

If we compare Figures A.3 and A.4, we can observe that the two rules are equivalent. We reject the null hypothesis when the observed significance level produces a smaller shaded area, exactly the requirement that the observed value be larger than the critical value. This approach has two principal benefits:

- Computer programs can calculate the P-value without requiring the user to provide a value for α.
- All the tests in this book can be formulated in such a way that the decision rule takes the form of statement (A.1).

Figure A.5 *P*-Value Corresponding to |*t*(observed)| = 2.8



A.6 Test for Nomality

A key assumption throughout the book is that the error terms in our models are normally distributed. When parameters are estimated, nonnormality is ameliorated because of the CLT. However, when we consider prediction intervals, no CLT effect exists, so we must endeavor to build a model such that the errors may be assumed to be approximately normally distributed. In order to check this assumption, we test

H_0: Errors are normally distributed, versus H_A: errors are not normally distributed.

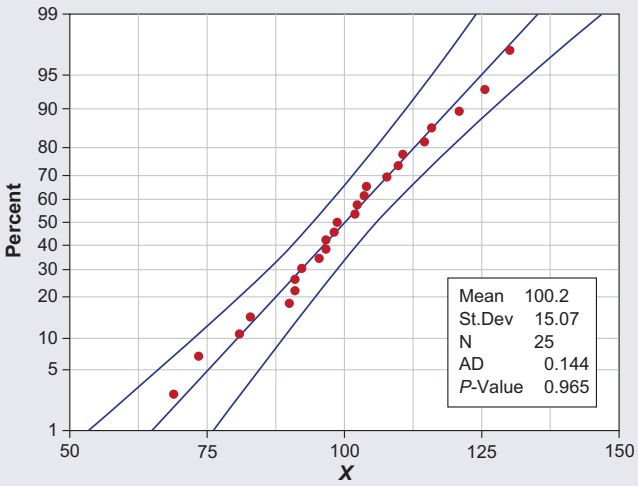
A number of tests exist for this purpose, but we focus on a test based upon the *normal probability plot* (*NPP*), which is constructed as follows:

1. Given n observations, place them in increasing order, say, $x_1 \le x_2 \le \ldots \le x_n$.
2. Calculate the coefficients $m_j = 100\,\dfrac{j - 0.375}{n + 0.25}$, $j = 1, 2, \ldots, n$.
3. Plot m_j on the vertical axis against x_j on the horizontal axis, *but*—and this is the tricky part—stretch the vertical axis so that the expected locations of the points would lie on a straight line if the observations were from a normal distribution.

Figure A.6 illustrates the nature of the plot for $n = 25$ observations from a normal distribution with mean 100 and standard deviation 15.

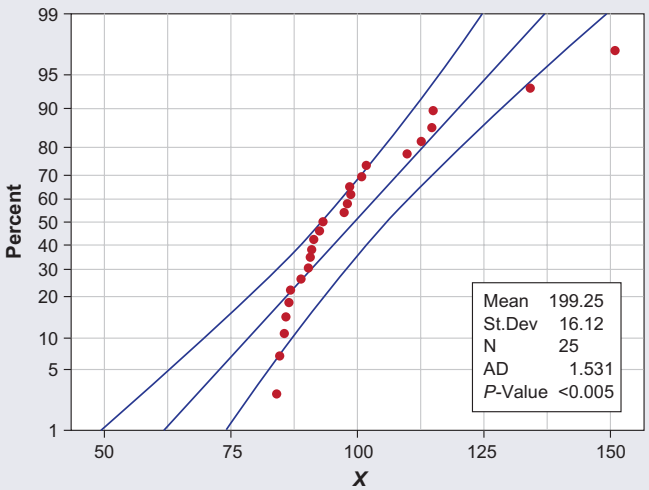
As may be observed from Figure A.6, the points are progressively higher as we move from left to right. Further, they are closely grouped in the center, but spread out at either end, consistent with the bell shape of the normal distribution. The curved upper and lower lines represent the 95 percent confidence bands, and the Anderson–Darling (AD) test statistic is a measure of how far the observations depart from the central straight line that describes the expected values.

Figure A.6 Normal Probability Plot for a Sample of Size $n = 25$
(*The outer limits are the 95 percent confidence bands*)



The *P*-value can be used to carry out the AD test. In this case, $P = 0.965$, so the data are clearly consistent with the normal assumption. Figure A.7 shows the *NPP* for data from a nonnormal distribution, with a resulting *P*-value of less than 0.005. The shape of the plot indicates that the distribution has a long upper tail (positive skewness).

Figure A.7 Normal Probability Plot for a Sample of Size $n = 25$ from a Nonnormal Distribution
(*The outer limits are the 95 percent confidence bands*)



Reference

Anderson, D. R., Sweeney, D. J., Williams, T. A., Camm, J.D., and Cochran, J.J. (2014). *Statistics for Business and Economics*, 12th ed. Mason, OH: South-Western.