RANDOM PHENOMENA

FUNDAMENTALS OF PROBABILITY AND STATISTICS FOR ENGINEERS

BABATUNDE A. OGUNNAIKE

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **Informa** business

Babatunde A. Ogunnaike

Random Phenomena

Fundamentals of Probability & Statistics for Engineers

Chapter 15

Hypothesis Testing

15.1 Introduction 548
15.2 Basic Concepts 549
15.2.1 Terminology and Definitions 549
Statistical Hypothesis 550
Test Statistic, Critical Region, and Significance Level 552
Potential Errors, Risks, and Power 553
Sensitivity and Specificity 554
The p-value 555
15.2.2 General Procedure 556
15.3 Concerning Single Mean of a Normal Population 557
15.3.1 σ Known; the "z-test" 559
Using MINITAB 563
15.3.2 σ Unknown; the "t-test" 566
Using MINITAB 569
15.3.3 Confidence Intervals and Hypothesis Tests 569
15.4 Concerning Two Normal Population Means 572
15.4.1 Population Standard Deviations Known 572
15.4.2 Population Standard Deviations Unknown 574
Equal Standard Deviations 574
Using MINITAB 576
Unequal Standard Deviations 576
Confidence Intervals and Two-Sample Tests 577
An Illustrative Example: The Yield Improvement Problem 578
15.4.3 Paired Differences 581
15.5 Determining β, Power, and Sample Size 586
15.5.1 β and Power 587
15.5.2 Sample Size 589
Practical Considerations 592
15.5.3 β and Power for Lower-Tailed and Two-Sided Tests 593
15.5.4 General Power and Sample Size Considerations 594
15.6 Concerning Variances of Normal Populations 596
15.6.1 Single Variance 596
15.6.2 Two Variances 599
15.7 Concerning Proportions 602
15.7.1 Single Population Proportion 603
Large Sample Approximations 603
Exact Tests 605
15.7.2 Two Population Proportions 606
15.8 Concerning Non-Gaussian Populations 608
15.8.1 Large Sample Test for Means 608
15.8.2 Small Sample Tests 609
15.9 Likelihood Ratio Tests 611
15.9.1 General Principles 612
15.9.2 Special Cases 614
Normal Population; Known Variance 614

Normal Population; Unknown Variance 616
15.9.3 Asymptotic Distribution for Λ 618
15.10 Discussion 618
15.11 Summary and Conclusions 620
REVIEW QUESTIONS 622
EXERCISES 625
APPLICATION PROBLEMS 633

The great tragedy of science—
the slaying of a beautiful hypothesis by an ugly fact.

T. H. Huxley (1825–1895)

Since turning our attention fully to Statistics in Part IV, our focus has been on characterizing the population completely, using finite-sized samples. The discussion that began with sampling in Chapter 13, providing the mathematical foundation for characterizing the variability in random samples, and which continued with estimation in Chapter 14, providing techniques for determining values for populations parameters, concludes in this chapter with hypothesis testing. This final tier of the statistical inference edifice is concerned with making—and testing—assertive statements about the population. Such statements are often necessary to solve practical problems, or to answer questions of practical importance; and this chapter is devoted to presenting the principles, practice, and mechanics of testing the validity of hypothesized statements regarding the distribution of populations. The chapter covers extensive ground—from traditional techniques applied to traditional Gaussian problems, to non-Gaussian problems and some non-traditional techniques; it ends with a brief but frank discussion of persistent criticisms of hypothesis tests and some practical recommendations for handling such criticisms.

15.1 Introduction

We begin our discussion by returning to the first problem presented in Chapter 1 concerning yields from two chemical processes; we wish to use it to illustrate the central issues with hypothesis testing. Recall that the problem requires that we decide which process, the challenger, A, or the incumbent, B, should be chosen for commercial operation. The decision is to be based on economically driven comparisons that translate to answering the following mathematical questions about the yields Y_A and Y_B:

1. Is $Y_A \geq 74.5$ and $Y_B \geq 74.5$, consistently?
2. Is $Y_A > Y_B$?
3. If yes, is $Y_A - Y_B > 2$?

To deal with the problem systematically, inherent random variability compels us to start by characterizing the populations fully with pdfs which are then used to answer these questions. This requires that we postulate an appropriate probability model, and determine values for the unknown parameters from sample data.

Here is what we know thus far (from Chapters 1 and 12, and from various examples in Chapter 14): we have plotted histograms of the data and postulated that these are samples from Gaussian-distributed populations; we have computed sample averages, $\bar{y}_A, \bar{y}_B$, and sample standard deviations, s_A, s_B; and in various Chapter 14 examples, we have obtained point and interval estimates for the population means μ_A, μ_B, and the population standard deviations σ_A, σ_B.

But by themselves, these results are not quite sufficient to answer the questions posed above. To answer the questions, consider the following statements and the implications of being able to confirm/refute them:

1. Y_A is a random variable characterized by a normal population with mean value 75.5 and standard deviation 1.5, i.e., $Y_A \sim N(75.5, 1.5^2)$; similarly, $Y_B \sim N(72.5, 2.5^2)$; as a consequence,

2. The random variables, Y_A and Y_B, are *not* from the same distribution because $\mu_A \neq \mu_B$ *and* $\sigma_A^2 \neq \sigma_B^2$; in particular, $\mu_A > \mu_B$;

3. Furthermore, $\mu_A - \mu_B > 2$.

This is a collection of assertions about these two populations, statements which, if confirmed, will enable us answer the questions raised. For example, Statement #1 will allow us to answer Question 1 by making it possible to compute the probabilities $P(Y_A \geq 74.5)$ and $P(Y_B \geq 74.5)$; Statement #2 will allow us to answer Question 2, and Statement #3, Question 3. How practical problems are formulated as statements of this type, and how such statements are confirmed or refuted, all fall under the formal subject matter of hypothesis testing. In general, the validity of such statements (or other assumptions about the population from which the sample data were obtained) is checked by (i) proposing an appropriate "statistical hypothesis" about the problem at hand; and (ii) testing this hypothesis against the evidence contained in the data.

15.2 Basic Concepts

15.2.1 Terminology and Definitions

Before launching into a discussion of the principles and mechanics of hypothesis testing, it is important to introduce first some terminology and definitions.

Statistical Hypothesis

A statistical hypothesis is a statement (an assertion or postulate) about the distribution of one or more populations. (Theoretically, the statistical hypothesis is a statement regarding one or more postulated distributions for the random variable X—distributions for which the statement is presumed to be true. A *simple* hypothesis specifies a single distribution for X; a *composite* hypothesis specifies more than one distribution for X.)

Modern hypothesis testing involves two hypotheses:

1. The *null hypothesis*, H_0, which represents the primary, "status quo" hypothesis that we are predisposed to believe as true (a plausible explanation of the observation) unless there is evidence in the data to indicate otherwise—in which case, it will be rejected in favor of a postulated alternative.

2. The *alternative hypothesis*, H_a, the carefully defined complement to H_0 that we are willing to consider in replacement if H_0 is rejected.

For example, the portion of Statement #1 above concerning Y_A may be formulated more formally as:

$$\begin{aligned} H_0 &: \mu_A = 75.5 \\ H_a &: \mu_A \neq 75.5 \end{aligned} \tag{15.1}$$

The implication here is that we are willing to entertain the fact that the true value of μ_A, the mean value of the yield obtainable from process A, is 75.5; that any deviation of the sample data average from this value is due to purely random variability and is not significant (i.e., that this postulate explains the observed data). The alternative is that any observed difference between the sample average and 75.5 is *real* and not just due to random variability; that the alternative provides a better explanation of the data. Observe that this alternative makes no distinction between values that are less than 75.5 or greater; so long as there is evidence that the observed sample average is different from 75.5 (whether greater than or less than), H_0 is to be rejected in favor of this H_a. Under these circumstances, since the alternative admits of values of μ_A that can be less than 75.5 or greater than 75.5, it is called a *two-sided* hypothesis.

It is also possible to formulate the problem such that the alternative actually "chooses sides," for example:

$$\begin{aligned} H_0 &: \mu_A = 75.5 \\ H_a &: \mu_A < 75.5 \end{aligned} \tag{15.2}$$

In this case, when the evidence in the data does not support H_0 the only other option is that $\mu_A < 75.5$. Similarly, if the hypotheses are formulated instead

as:

$$H_0 : \mu_A = 75.5$$
$$H_a : \mu_A > 75.5 \tag{15.3}$$

the alternative, if the equality conjectured by the null hypothesis fails, is that the mean must then be greater. These are *one-sided* hypotheses, for obvious reasons.

A *test* of a statistical hypothesis is a procedure for deciding when to reject H_0. The conclusion of a hypothesis test is either a decision to *reject* H_0 in favor of H_a or else to *fail to reject* H_0. Strictly speaking, one never actually "accepts" a hypothesis; one just fails to reject it.

As one might expect, the conclusion drawn from a hypothesis test is shaped by how H_a is framed in contrast to H_0. How to formulate the H_a appropriately is best illustrated with an example.

Example 15.1: HYPOTHESES FORMULATION FOR COMPARING ENGINEERING TRAINING PROGRAMS

As part of an industrial training program for chemical engineers in their junior year, some trainees are instructed by Method A, and some by Method B. If random samples of size 10 each are taken from large groups of trainees instructed by each of these two techniques, and each trainee's score on an appropriate achievement test is shown below, formulate a null hypothesis H_0, and an appropriate alternative H_a, to use in testing the claim that Method B is more effective.

Method A	71	75	65	69	73	66	68	71	74	68
Method B	72	77	84	78	69	70	77	73	65	75

Solution:

We do return to this example later to provide a solution to the problem posed; for now, we address only the issue of formulating the hypotheses to be tested.

Let μ_A represent the true mean score for engineers trained by Method A, and μ_B, the true mean score for those trained by the other method. The status quo postulate is to presume that there is no difference between the two methods; that any observed difference is due to pure chance alone. The key now is to inquire: if there is evidence in the data that contradicts this status quo postulate, what end result are we interested in testing this evidence against? Since the claim we are interested in confirming or refuting is that Method B is more effective, then the proper formulation of the hypotheses to be tested is as follows:

$$H_0 : \mu_A = \mu_B$$
$$H_a : \mu_A < \mu_B \tag{15.4}$$

By formulating the problem in this fashion, any evidence that contradicts the null hypothesis will cause us to reject it in favor of something that is actually relevant to the problem at hand.

> Note that in this case specifying H_a as $\mu_A \neq \mu_B$ does not help us answer the question posed; by the same token, neither does specifying H_a as $\mu_A > \mu_B$ because if it is true that $\mu_A < \mu_B$, then the evidence in the data will not support the alternative—a circumstance which, by default, will manifest as a misleading lack of evidence to reject H_0.

Thus, in formulating statistical hypotheses, it is customary to state H_0 as the "no difference," *nothing-interesting-is-happening* hypothesis; the alternative, H_a, is then selected to answer the question of interest when there is evidence in the data to contradict the null hypothesis. (See Section 15.10 below for additional discussion about this and other related issues.)

A classic illustration of these principles is the US legal system in which a defendant is considered innocent until proven guilty. In this case, the null hypothesis is that this defendant is no different from any other innocent individual; after evidence has been presented to the jury by the prosecution, the verdict is handed down either that the defendant is guilty (i.e., rejecting the null hypothesis) or the defendant is not guilty (i.e., failing to reject the null hypotheses). Note that the defendant is *not* said to be "innocent"; instead, the defendant is pronounced "not guilty," which is tantamount to a decision *not* to reject the null hypothesis.

Because hypotheses are statements about populations, and, as with estimation, hypothesis tests are based on finite-sized sample data, such tests are subject to random variability and are therefore only meaningful in a probabilistic sense. This leads us to the next set of definitions and terminology.

Test Statistic, Critical Region, and Significance Level

To test a hypothesis, H_0, about a population parameter, θ, for a random variable, X, against an alternative, H_a, a random sample, $X_1, X_2, \ldots, X_n$ is acquired, from which an estimator for θ, say $U(X_1, X_2, \ldots, X_n)$, is then obtained. (Recall that U is a random variable whose specific value will vary from sample to sample.)

A *test statistic*, $Q_T(U, \theta)$, is an appropriate function of the parameter θ and its estimator, U, that will be used to determine whether or not to reject H_0. (What "appropriate" means will be clarified shortly.) A *critical region* (or rejection region), R_C, is a region representing the numerical values of the test statistic ($Q_T > q$, or $Q_T < q$, or both) that will trigger the rejection of H_0; i.e., if $Q_T \in R_C$, H_0 will be rejected. Strictly speaking, the critical region is for the random variable, X; but since the random sample from X is usually converted to a test statistic, there is a corresponding mapping of this region by $Q_T(.)$; it is therefore acceptable to refer to the critical region in terms of the test statistic.

Now, because the estimator U is a random variable, the test statistic will itself also be a random variable, with the following serious implication: *there is a non-zero probability that $Q_T \in R_C$ even when H_0 is true.* This unavoidable consequence of random variability forces us to design the hypothesis test such

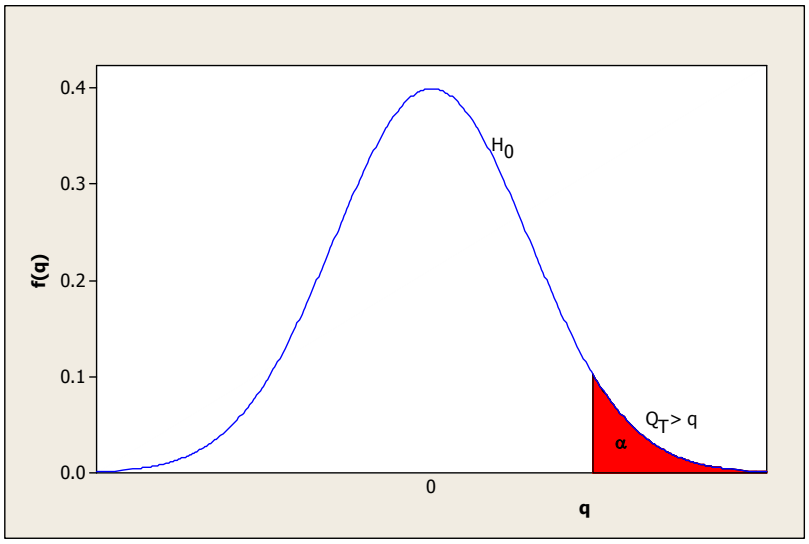

FIGURE 15.1: A distribution for the null hypothesis, H_0, in terms of the test statistic, Q_T, where the shaded rejection region, $Q_T > q$, indicates a significance level, α.

that H_0 is rejected only if it is "highly unlikely" for $Q_T \in R_C$ when H_0 is true. How unlikely is "highly unlikely"? This is quantified by specifying a value α such that

$$P(Q_T \in R_C | H_0 \text{ true}) \leq \alpha \tag{15.5}$$

with the implication that the probability of rejecting H_0 when it is in fact true, is never greater than α. This quantity, often set in advance as a small value (typically 0.1, 0.05, or 0.01), is called the *significance level* of the test. Thus, the significance level of a test is the upper bound on the probability of rejecting H_0 when it is true; it determines the boundaries of the critical region R_C.

These concepts are illustrated in Fig 15.1 and lead directly to the consideration of the potential errors to which hypothesis tests are susceptible, the associated risks, and the sensitivity of a test in leading to the correct decision.

Potential Errors, Risks, and Power

Hypothesis tests are susceptible to two types of errors:

1. *TYPE I error*: the error of rejecting H_0 when it is in fact true. This is the legal equivalent of convicting an innocent defendant.

2. *TYPE II error*: the error of failing to reject H_0 when it is false, the legal equivalent of letting a guilty defendant go scotfree.

TABLE 15.1: Hypothesis test decisions and risks

Decision → / Truth ↓	Fail to Reject H_0	Reject H_0
H_0 True	Correct Decision *Probability:* $(1-\alpha)$	Type I Error *Risk:* α
H_a True	Type II Error *Risk:* β	Correct Decision *Probability:* $(1-\beta)$

Of course a hypothesis test can also result in the correct decision in two ways: rejecting the null hypothesis when it is false, or failing to reject the null hypothesis when it is true.

From the definition of the critical region, R_C, and the significance level, the probability of committing a Type I error is α; i.e.,

$$P(Q_T \in R_C | H_0 \text{ true}) = \alpha \tag{15.6}$$

It is therefore called the α-risk. The probability of correctly refraining from rejecting H_0 when it is true will be $(1-\alpha)$.

By the same token, it is possible to compute the probability of committing a Type II error. It is customary to refer to this value as β, i.e.,

$$P(Q_T \notin R_C | H_0 \text{ false}) = \beta \tag{15.7}$$

so that the probability of committing a Type II error is called the β-risk. The probability of correctly rejecting a null hypothesis that is false is therefore $(1-\beta)$.

It is important now to note that the two correct decisions and the probabilities associated with each one are fundamentally different. Primarily because H_0 is the "status quo" hypothesis, correctly rejecting a null hypothesis, H_0, that is false is of greater interest because such an outcome indicates that the test has detected the occurrence of something significant. Thus, $(1-\beta)$, the probability of correctly rejecting the false null hypothesis when the alternative hypothesis is true, is known as the *power* of the test. It provides a measure of the sensitivity of the test. These concepts are summarized in Table 15.1 and also in Fig 15.2.

Sensitivity and Specificity

Because their results are binary decisions (reject H_0 or fail to reject it), hypothesis tests belong in the category of *binary classification tests*; and the effectiveness of such tests are characterized in terms of sensitivity and specificity. The *sensitivity* of a test is the percentage of true "positives" (in this case, H_0 deserving of rejection) that it correctly classifies as such. The *specificity* is the percentage of true "negatives" (H_0 that should *not* be rejected) that is correctly classified as such. Sensitivity therefore measures the ability to identify true positives correctly; specificity, the ability to identify true negatives correctly.

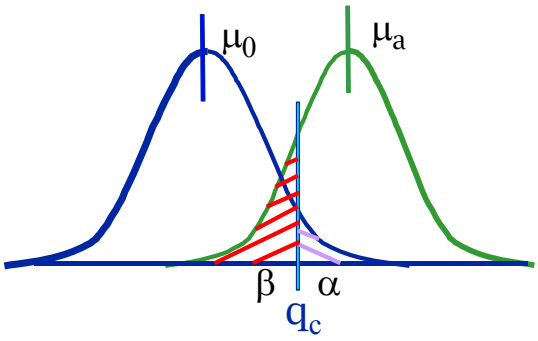

FIGURE 15.2: Overlapping distributions for the null hypothesis, H_0 (with mean μ_0), and alternative hypothesis, H_a (with mean μ_a), showing Type I and Type II error risks α, β, along with q_C the boundary of the critical region of the test statistic, Q_T.

These performance measures are related to the risks and errors discussed previously. If the percentages are expressed as probabilities, then sensitivity is $(1-\beta)$, and specificity, $(1-\alpha)$. The fraction of "false positives" (H_0 that should *not* be rejected but is) is α; the fraction of "false negatives" (H_0 that should be rejected but is not) is β. As we show later, for a fixed sample size, improving one measure can only be achieved at the expense of the other, i.e., improvements in specificity must be traded off for a commensurate loss of sensitivity, and vice versa.

The p-value

Rather than fix the significance level, α, ahead of time, suppose it is free to vary. For any given value of α, let the corresponding critical/rejection region be represented as $R_C(\alpha)$. As discussed above, H_0 is rejected whenever the test statistic, Q_T, is such that $Q_T \in R_C(\alpha)$. For example, from Fig 15.1, the region $R_C(\alpha)$ is the set of all values of Q_T that exceed the specific value q. Observe that as α decreases, the "size" of the set $R_C(\alpha)$ also decreases, and vice versa. The *smallest* value of α for which the specific value of the test statistic $Q_T(x_1, x_2, \ldots, x_n)$ (determined from the data set $x_1, x_2, \ldots, x_n$) falls in the critical region (i.e., $Q_T(x_1, x_2, \ldots, x_n) \in R_C(\alpha)$) is known as the *p-value* associated with this data set (and the resulting test statistic). Technically, therefore, the p-value is the *smallest* significance level at which H_0 will be rejected given the observed data.

This somewhat technical definition of the p-value is sometimes easier to understand as follows: given specific observations $x_1, x_2, \ldots, x_n$ and the corresponding test statistic $Q_T(x_1, x_2, \ldots, x_n)$ computed from them to yield the specific value q; the p-value associated with the observations and the corresponding test statistic is defined by the following probability statement:

$$p = P[Q_T(x_1, x_2, \cdots, x_n; \theta) \geq q | H_0] \tag{15.8}$$

In words, this is the probability of obtaining the specific test statistic value, q, or something more extreme, if the null hypothesis is true. Note that p, being a function of a statistic, is itself a statistic—a subtle point that is often easy to miss; the implication is that p is itself subject to purely random variability.

Knowing the p-value therefore allows us to carry out hypotheses tests at any significance level, without restriction to pre-specified α values. In general, a low value of p indicates that, given the evidence in the data, the null hypothesis, H_0, is highly unlikely to be true. This follows from Eq (15.8). H_0 is then rejected at the significance level, p, which is why the p-value is sometimes referred to as *the observed significance level*—observed from the sample data, as opposed to being fixed, *à-priori*, at some pre-specified value, α.

Nevertheless, in many applications (especially in scientific publications), there is an enduring traditional preference for employing fixed significance levels (usually $\alpha = 0.05$). In this case, the p-value is used to make decisions as follows: if $p < \alpha$, H_0 will be rejected at the significance level α; if $p > \alpha$, we fail to reject H_0 at the same significance level α.

15.2.2 General Procedure

The general procedure for carrying out modern hypotheses tests is as follows:

1. Define H_0, the hypothesis to be tested, and pair it with the alternative H_a, formulated appropriately to answer the question at hand;
2. Obtain sample data, and from it, the test statistic relevant to the problem at hand;
3. Make a decision about H_0 as follows: Either
   (a) Specify the significance level, α, at which the test is to be performed, and hence determine the critical region (equivalently, the critical value of the test statistic) that will trigger rejection; then
   (b) Evaluate the specific test statistic value in relation to the critical region and reject, or fail to reject, H_0 accordingly;

   or else,

   (a) Compute the p-value corresponding to the test statistic, and
   (b) Reject, or fail to reject, H_0 accordingly on this basis.

How this general procedure is applied depends on the specific problem at hand: the nature of the random variable, hence the underlying postulated population itself; what is known or unknown about the population; the particular population parameter that is the subject of the test; and the nature of the question to be answered. The remainder of this chapter is devoted to presenting the principles and mechanics of the various hypothesis tests commonly

encountered in practice, some of which are so popular that they have acquired recognizable names (for example, the z-test; t-test; χ^2-test; F-test; etc.). By taking time to provide the *principles* along with the mechanics, our objective is to supply the reader with the sort of information that should help to prevent the surprisingly common mistake of misapplying some of these tests. The chapter closes with a brief discussion of some criticisms and potential shortcomings of classical hypothesis testing.

15.3 Concerning Single Mean of a Normal Population

Let us return to the illustrative statements made earlier in this chapter regarding the yields from two competing chemical processes. In particular, let us recall the first half of the statement about the yield of process A—that $Y_A \sim N(75.5, 1.5^2)$. Suppose that we are first interested in testing the validity of this statement by inquiring whether or not the true mean of the process yield is 75.5. The starting point for this exercise is to state the null hypothesis, which in this case is:

$$\mu_A = 75.5 \tag{15.9}$$

since 75.5 is the specific postulated value for the unknown population mean μ_A. Next, we must attach an appropriate alternative hypothesis. The original statement is a categorical one that Y_A comes from the distribution $N(75.5, 1.5^2)$, with the hope of being able to use this statement to distinguish the Y_A distribution from the Y_B distribution. (How this latter task is accomplished is discussed later). Thus, the only alternative we are concerned about, should H_0 prove false, is that the true mean is not equal to 75.5; we do not care if the true mean is less than, or greater than, the postulated value. In this case, the appropriate H_a is therefore:

$$\mu_A \neq 75.5 \tag{15.10}$$

Next, we need to gather "evidence" in the form of sample data from process A. Such data, with $n = 50$, was presented in Chapter 1 (and employed in the examples of Chapter 14), from which we have obtained a sample average, $\bar{y}_A = 75.52$. And now, the question to be answered by the hypothesis test is as follows: is the observed difference between the postulated true population mean, $\mu_A = 75.5$, and the sample average computed from sample process data, $\bar{y}_A = 75.52$, due purely to random variation or does it indicate a real (and significant) difference between postulate and data? From Chapters 13 and 14, we now know that answering this question requires a sampling distribution that describes the variability intrinsic to samples. In this specific case, we know that for a sample average $\bar{X}$ obtained from a random sample of size n

from a $N(\mu, \sigma^2)$ distribution, the statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{15.11}$$

has the standard normal distribution, provided that σ is known. This immediately suggests, within the context of hypothesis testing, that the following test statistic:

$$Z = \frac{\bar{y}_A - 75.5}{1.5/\sqrt{n}} \tag{15.12}$$

may be used to test the validity of the hypothesis, for any sample average computed from any sample data set of size n. This is because we can use Z and its pdf to determine the critical/rejection region. In particular, by specifying a significance level $\alpha = 0.05$, the rejection region is determined as the values z such that:

$$R_C = \{z | z < -z_{0.025}; z > z_{0.025}\} \tag{15.13}$$

(because this is a two-sided test). From the cumulative probability characteristics of the standard normal distribution, we obtain (using computer programs such as MINITAB) $z_{0.025} = 1.96$ as the value of the standard normal variate for which $P(Z > z_{0.025}) = 0.025$, i.e.,

$$R_C = \{z | z < -1.96; z > 1.96\}; \text{ or } |z| > 1.96 \tag{15.14}$$

The implication: if the specific value computed for Z from any sample data set exceeds 1.96 in absolute value, H_0 will be rejected.

In the specific case of $\bar{y}_A = 75.52$ and $n = 50$, we obtain a specific value for this test statistic as $z = 0.094$. And now, because this value $z = 0.094$ does not lie in the critical/rejection region defined in Eq (15.14), we conclude that there is no evidence to reject H_0 in favor of the alternative. The data does not contradict the hypothesis.

Alternatively, we could compute the p-value associated with this test statistic (for example, using the cumulative probability feature of MINITAB):

$$P(z > 0.094 \text{ or } z < 0.094) = P(|z| > 0.094) = 0.925 \tag{15.15}$$

implying that if H_0 is true, the probability of observing, by pure chance alone, the sample average data actually observed, or something "more extreme," is very high at 0.925. Thus, there is no evidence in this data set to justify rejecting H_0. From a different perspective, note that this p-value is nowhere close to being *lower* than the prescribed significance level, $\alpha = 0.05$; we therefore fail to reject the null hypothesis at this significance level.

The ideas illustrated by this example can now be generalized. As with previous discussions in Chapter 14, we organize the material according to the status of the population standard deviation, σ, because whether it is known or not determines what sampling distribution—and hence test statistic—is appropriate.

15.3.1 σ Known; the "z-test"

Problem: The random variable, X, possesses a distribution, $N(\mu, \sigma^2)$, with unknown value, μ, but known σ; a random sample, $X_1, X_2, \ldots, X_n$, is drawn from this normal population from which a sample average, $\bar{X}$, can be computed; a specific value, μ_0, is hypothesized for the true population parameter; and it is desired to test whether the sample indeed came from such a population.

The Hypotheses: In testing such a hypothesis—concerning a single mean of a normal population with known standard deviation, σ—the null hypothesis is typically:

$$H_0 : \mu = \mu_0 \tag{15.16}$$

where μ_0 is the specific value postulated for the population mean (e.g., 75.5 used in the previous illustration). There are three possible alternative hypotheses:

$$H_a : \mu < \mu_0 \tag{15.17}$$

for the *lower-tailed*, one-sided (or one-tailed) alternative hypothesis; or

$$H_a : \mu > \mu_0 \tag{15.18}$$

for the *upper-tailed*, one-sided (or one-tailed) alternative hypothesis; or, finally, as illustrated above,

$$H_a : \mu \neq \mu_0 \tag{15.19}$$

for the two-sided (or two-tailed) alternative.

Assumptions: The underlying distribution in question is Gaussian, with known standard deviation, σ, implying that the sampling distribution of $\bar{X}$ is also Gaussian, with mean, μ_0, and variance, σ^2/n, if H_0 is true. Hence, the random variable $Z = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ has a standard normal distribution, $N(0, 1)$.

Test Statistic: The appropriate test statistic is therefore

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \tag{15.20}$$

The specific value obtained for a particular sample data average, $\bar{x}$, is sometimes called the "z-score" of the sample data.

Critical/Rejection Regions:

(i) For lower-tailed tests (with $H_a : \mu < \mu_0$), reject H_0 in favor of H_a if:

$$z < -z_\alpha \tag{15.21}$$

where z_α is the value of the standard normal variate, z, with a tail area probability of α; i.e., $P(z > z_\alpha) = \alpha$. By symmetry, $P(z < -z_\alpha) = P(z > z_\alpha) = \alpha$, as shown in Fig 15.3. The rationale is that if $\mu = \mu_0$ is true, then it is highly unlikely that z will be less than $-z_\alpha$ by pure chance alone; it is more likely that μ is systematically less than μ_0 if z is less than $-z_\alpha$.

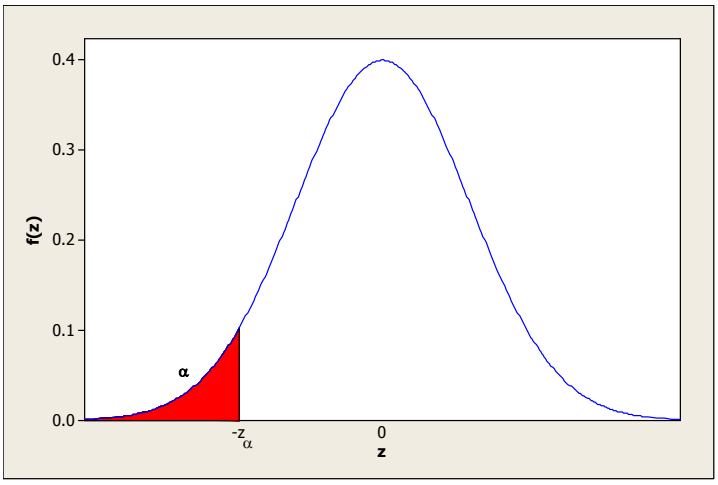

FIGURE 15.3: The standard normal variate $z = -z_\alpha$ with tail area probability α. The shaded portion is the rejection region for a lower-tailed test, $H_a : \mu < \mu_0$.

(ii) For upper-tailed tests (with $H_a : \mu > \mu_0$), reject H_0 in favor of H_a if (see Fig 15.4):

$$z > z_\alpha \tag{15.22}$$

(iii) For two-sided tests (with $H_a : \mu \neq \mu_0$), reject H_0 in favor of H_a if:

$$z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2} \tag{15.23}$$

for the same reasons as above, because if H_0 is true, then

$$P(z < -z_{\alpha/2} \text{ or } z > z_{\alpha/2}) = \frac{\alpha}{2} + \frac{\alpha}{2} = \alpha \tag{15.24}$$

as illustrated in Fig 15.5.

Tests of this type are known as "z-tests" because of the test statistic (and sampling distribution) upon which the test is based. Therefore,

> The *one-sample* z-test is a hypothesis test concerning the mean of a normal population where the population standard deviation, σ, is specified.

The key facts about the z-test for testing $H_0 : \mu = \mu_0$ are summarized in Table 15.2.

The following two examples illustrate the application of the "z-test."

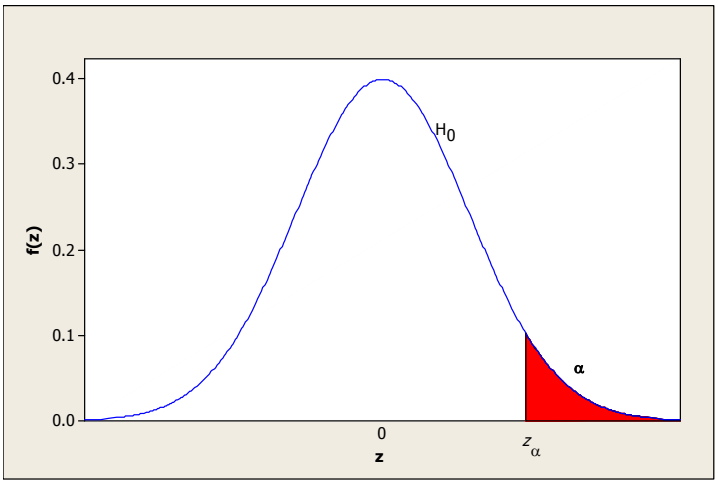

FIGURE 15.4: The standard normal variate $z = z_\alpha$ with tail area probability α. The shaded portion is the rejection region for an upper-tailed test, $H_a : \mu > \mu_0$.

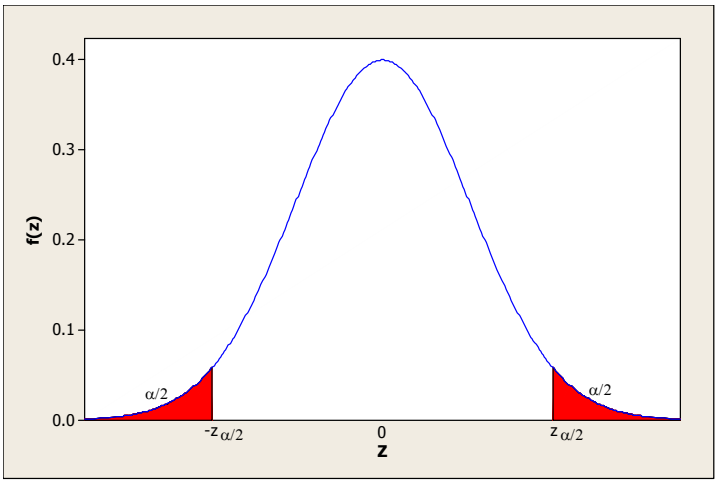

FIGURE 15.5: Symmetric standard normal variates $z = z_{\alpha/2}$ and $z = -z_{\alpha/2}$ with identical tail area probabilities $\alpha/2$. The shaded portions show the rejection regions for a two-sided test, $H_a : \mu \neq \mu_0$.

TABLE 15.2: Summary of H_0 rejection conditions for the one-sample z-test

Testing Against	For General α Reject H_0 if:	For $\alpha = 0.05$ Reject H_0 if:
$H_a : \mu < \mu_0$	$z < -z_\alpha$	$z < -1.65$
$H_a : \mu > \mu_0$	$z > z_\alpha$	$z < 1.65$
$H_a : \mu \neq \mu_0$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$	$z < -1.96$ or $z > 1.96$

Example 15.2: CHARACTERIZING YIELD FROM PROCESS B

Formulate and test (at the significance level of $\alpha = 0.05$) the hypothesis implied by the second half of the statement given at the beginning of this chapter about the mean yield of process B, i.e., that $Y_B \sim N(72.5, 2.5^2)$. Use the data given in Chapter 1 and analyzed previously in various Chapter 14 examples.

Solution:

In this case, as with the Y_A illustration used to start this section, the hypotheses to be tested are:

$$\begin{aligned} H_0 &: \mu_B = 72.5 \\ H_a &: \mu_B \neq 72.5 \end{aligned} \tag{15.25}$$

a two-sided test. From the supplied data, we obtain $\bar{y}_B = 72.47$; and since the population standard deviation, σ_B, is given as 2.5, the specific value, z, of the appropriate test statistic, Z (the "z-score"), from Eq (15.20), is:

$$z = \frac{72.47 - 72.50}{2.5/\sqrt{50}} = -0.084 \tag{15.26}$$

For this two-sided test, the critical value to the right, $z_{\alpha/2}$, for $\alpha = 0.05$, is:

$$z_{0.025} = 1.96 \tag{15.27}$$

so that the critical/rejection region, R_C, is $z > 1.96$ to the right, in conjunction with $z < -1.96$ to the left, by symmetry (recall Eq (15.14)). And now, because the specific value $z = -0.084$ does not lie in the critical/rejection region, we find no evidence to reject H_0 in favor of the alternative. We conclude therefore that Y_B is very likely well-characterized by the postulated distribution.

We could also compute the p-value associated with this test statistic

$$P(z < -0.084 \text{ or } z > 0.084) = P(|z| > 0.084) = 0.933 \tag{15.28}$$

with the following implication: if H_0 is true, the probability of observing, by pure chance alone, the actually observed sample average, $\bar{y}_B = 72.47$,

or something "more extreme" (further away from the hypothesized mean of 72.50) is 0.933. Thus, there is no evidence to support rejecting H_0. Furthermore, since this p-value is much higher than the prescribed significance level, $\alpha = 0.05$, we cannot reject the null hypothesis at this significance level.

Using MINITAB

It is instructive to walk through the typical procedure for carrying out such z-tests using computer software, in this case, MINITAB. From the MINITAB drop down menu, the sequence `Stat > Basic Statistics > 1-Sample Z` opens a dialog box that allows the user to carry out the analysis either using data already stored in MINITAB worksheet columns or from summarized data. Since we already have summarized data, upon selecting the "Summarized data" option, one enters 50 into the "Sample size:" dialog box, 72.47 into the "Mean" box, and 2.5 into the "Standard deviation" box; and upon slecting the "Perform hypothesis test" option, one enters 72.5 for the "Hypothesized mean." The "Options" button allows the user to select the confidence level (the default is 95.0) and the "Alternative" for H_a: with the 3 available options displayed as "less than," "not equal," and "greater than." The MINITAB results are displayed as follows:

One-Sample Z

```
Test of mu = 72.5 vs not = 72.5
The assumed standard deviation = 2.5
```

N	Mean	SE Mean	95% CI	Z	P
50	72.470	0.354	(71.777, 73.163)	-0.08	0.932

This output links hypothesis testing directly with estimation (as we anticipated in Chapter 14, and as we discuss further below) as follows: "SE Mean" is the standard error of the mean ($\sigma/\sqrt{n}$) from which the 95% confidence interval (shown in the MINITAB output as "95% CI") is obtained as (71.777, 73.163). Observe that the hypothesized mean, 72.5, is contained within this interval, with the implication that, since, at the 95% confidence level, the estimated average encompasses the hypothesized mean, we have no reason to reject H_0 at the significance level of 0.05. The z statistic computed by MINITAB is precisely what we had obtained in the example; the same is true of the p-value.

The results of this example (and the ones obtained earlier for Y_A) may now be used to answer the first question raised at the beginning of this chapter (and in Chapter 1) regarding whether or not Y_A and Y_B consistently exceed 74.5.

The random variable, Y_A, has now been completely characterized by the Gaussian distribution, $N(75.5, 1.5^2)$, and Y_B by $N(72.5, 2.5^2)$. From these

probability distributions, we are able to compute the following probabilities:

$$P(Y_A > 74.5) = 1 - P(Y_A < 74.5) = 0.748 \quad (15.29)$$

$$P(Y_B > 74.5) = 1 - P(Y_B < 74.5) = 0.212 \quad (15.30)$$

The sequence for calculating such cumulative probabilities with MINITAB is as follows: `Calc > Prob Dist > Normal`, which opens a dialog box for entering the desired parameters: (i) from the choices "Probability density," "Cumulative probability" and "Inverse cumulative probability," one selects the second one; "Mean" is specified as 75.5 for the Y_A distribution, "Standard deviation" is specified as 1.5; and upon entering the input constant as 74.5, MINITAB returns the following results:

Cumulative Distribution Function

```
Normal with mean = 75.5 and standard deviation = 1.5
    x  P(X<=x)
 74.5  0.252493
```

from which the required probability is obtained as $1-0.252 = 0.748$. Repeating the procedure for Y_B, with "Mean" specified as 72.5 and "Standard deviation" as 2.5 produces the result shown in Eq (15.30).

The implication of these results is that process A yields will exceed 74.5% around three-quarters of the time, whereas with the incumbent process B, exceeding yields of 74.5% will occur only one-fifths of the time. If profitability is related to yields that exceed 74.5% consistently, then process A will be roughly 3.5 times more profitable than the incumbent process B.

This next example illustrates how, in solving practical problems, "intuitive" reasoning without the objectivity of a formal hypothesis test can be misleading.

Example 15.3: CHARACTERIZING "FAST-ACTING" RAT POISON

The scientists at the ACME rat poison laboratories, who have been working non-stop to develop a new "fast-acting" formulation that will break the "thousand-second" barrier, appear to be on the verge of a breakthrough. Their target is a product that will kill rats within 1000 secs, on average, with a standard deviation of 125 secs. Experimental tests conducted in an affiliated toxicology laboratory in which pellets were made with a newly developed formulation and administered to 64 rats (selected at random from an essentially identical population). The results showed an average "acting time," $\bar{x} = 1028$ secs. The ACME scientists, anxious to declare a breakthrough, were preparing to approach management immediately to argue that the observed excess 28 secs, when compared to the stipulated standard deviation of 125 secs, is "small and insignificant." The group statistician, in an attempt to present an objective, statistically sound argument, recommended instead that a hypothesis test should first be carried out to rule out

the possibility that the mean "acting time" is still greater than 1000 secs. Assuming that the "acting time" measurements are normally distributed, carry out an appropriate hypothesis test and, at the significance level of $\alpha = 0.05$, make an informed recommendation regarding the tested rat poison's "acting time."

Solution:
For this problem, the null and alternative hypotheses are:

$$\begin{aligned} H_0 : \mu &= 1000 \\ H_a : \mu &> 1000 \end{aligned} \tag{15.31}$$

The alternative has been chosen this way because the concern is that the acting time may still be greater than 1000 secs. As a result of the normality assumption, and the fact that σ is specified as 125, the required test is the z-test, where the specific z-score is:

$$z = \frac{1028 - 1000}{125/\sqrt{64}} = 1.792 \tag{15.32}$$

The critical value, z_α, for $\alpha = 0.05$ for this upper-tailed test is:

$$z_{0.05} = 1.65 \tag{15.33}$$

obtained using MINITAB's inverse cumulative probability feature for the standard normal distribution (tail area probability 0.05), i.e.,

$$P(Z > 1.65) = 0.05 \tag{15.34}$$

Thus, the rejection region, R_C, is $z > 1.65$. And now, because $z = 1.78$ falls into the rejection region, the decision is to reject the null hypothesis at the 5% level.

Alternatively, the p-value associated with this test statistic can be obtained (also from MINITAB, using the cumulative probability feature) as:

$$P(z > 1.792) = 0.037 \tag{15.35}$$

implying that if H_0 is true, the probability of observing, by pure chance alone, the actually observed sample average, 1028 secs, or something higher, is so small that we are inclined to believe that H_0 is unlikely to be true. Observe that this p-value is lower than the specified significance level of $\alpha = 0.05$.

Thus, from these equivalent perspectives, the conclusion is that the experimental evidence *does not* support the ACME scientists premature declaration of a breakthrough; the observed excess 28 secs, in fact, appears to be significant at the $\alpha = 0.05$ significance level.

Using the procedure illustrated previously, the MINITAB results for this problem are displayed as follows:

One-Sample Z

```
Test of mu = 1000 vs > 1000
The assumed standard deviation = 125
```

N	Mean	SE Mean	95% Lower Bound	Z	P
64	1028.0	15.6	1002.3	1.79	0.037

Observe that the z- and p- values agree with what we had obtained earlier; furthermore, the additional entries, "SE Mean," for the standard error of the mean, 15.6, and the 95% lower bound on the estimate for the mean, 1002.3, link this hypothesis test to interval estimation. This connection will be explored more fully later in this section; for now, we note simply that the 95% lower bound on the estimate for the mean, 1002.3, lies entirely to the right of the hypothesized mean value of 1000. The implication is that, at the 95% confidence level, it is more likely that the true mean is *higher* than the value hypothesized; we are therefore more inclined to reject the null hypothesis in favor of the alternative, at the significance level 0.05.

15.3.2 σ Unknown; the "t-test"

When the population standard deviation, σ, is unknown, the sample standard deviation, s, will have to be substituted for it. In this case, one of two things can happen:

1. If the sample size is sufficiently large (for example, $n > 30$), s is usually considered to be a good enough approximation to σ, that the z-test can be applied, treating s as equal to σ.
2. When the sample size is small, substituting s for σ changes the test statistic and the corresponding test, as we now discuss.

For small sample sizes, when S is substituted for σ, the appropriate test statistic, becomes

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \tag{15.36}$$

which, from our discussion of sampling distributions, is known to possess a Student's t-distribution, with $\nu = n - 1$ degrees of freedom. This is the "small sample size" equivalent of Eq (15.20).

Once more, because of the test statistic, and the sampling distribution upon which the test is based, this test is known as a "t-test." Therefore,

> The *one-sample* t-test is a hypothesis test concerning the mean of a normal population when the population standard deviation, σ, is unknown, and the sample size is small.

TABLE 15.3: Summary of H_0 rejection conditions for the one-sample t-test

Testing Against	For General α Reject H_0 if:
$H_a : \mu < \mu_0$	$t < -t_\alpha(\nu)$
$H_a : \mu > \mu_0$	$t > t_\alpha(\nu)$
$H_a : \mu \neq \mu_0$	$t < -t_{\alpha/2}(\nu)$ or $t > t_{\alpha/2}(\nu)$ $(\nu = n - 1)$

The t-test is therefore the same as the z-test but with the sample standard deviation, s, used in place of the unknown σ; it uses the t-distribution (with the appropriate degrees of freedom) in place of the standard normal distribution of the z-test. The relevant facts about the t-test for testing $H_0 : \mu = \mu_0$ are summarized in Table 15.3, the equivalent of Table 15.2 shown earlier. The specific test statistic, t, is determined by introducing sample data into Eq (15.36). Unlike the z-test, even after specifying α, we are unable to determine the specific critical/rejection region because these values depend on the degrees of freedom (i.e., the sample size). The following example illustrates how to conduct a one-sample t-test.

Example 15.4: HYPOTHESES TESTING REGARDING ENGINEERING TRAINING PROGRAMS

Assume that the test results shown in Example 15.1 are random samples from normal populations. (1) At a significance level of $\alpha = 0.05$, test the hypothesis that the mean score for trainees using method A is $\mu_A = 75$, versus the alternative that it is less than 75. (2) Also, at the same significance level, test the hypothesis that the mean score for trainees using method B is $\mu_B = 75$, versus the alternative that it is not.

Solution:

(1) The first thing to note is that the population standard deviations are not specified; and since the sample size of 10 for each data set is small, the appropriate test is a one-sample t-test. The null and alternative hypotheses for the first problem are:

$$\begin{aligned} H_0 &: \mu_A = 75.0 \\ H_a &: \mu_A < 75.0 \end{aligned} \tag{15.37}$$

The sample average is obtained from the supplied data as $\bar{x}_A = 69.0$, with a sample standard deviation, $s_A = 4.85$; the specific T statistic value is thus obtained as:

$$t = \frac{69.0 - 75.0}{4.85/\sqrt{10}} = -3.91 \tag{15.38}$$

Because this is a lower-tailed, one-sided test, the critical value, $-t_{0.05}(9)$, is obtained as -1.833 (using MINITAB's inverse cumulative probability feature, for the t-distribution with 9 degrees of freedom). The rejection region, R_C, is therefore $t < -1.833$. Observe that the specific t-value for this test lies well within this rejection region; we therefore reject the null hypothesis in favor of the alternative, at the significance level 0.05.

Of course, we could also compute the p-value associated with this particular test statistic; and from the t-distribution with 9 degrees of freedom we obtain,

$$P(T(9) < -3.91) = 0.002 \tag{15.39}$$

using MINITAB's cumulative probability feature. The implication here is that the probability of observing a difference as large, or larger, between the postulated mean (75) and actual sample average (69), if H_0 is true, is so very low (0.002) that it is more likely that the alternative is true; that the sample average is more likely to have come from a distribution whose mean is less than 75. Equivalently since this p-value is less than the significance level 0.05, we reject H_0 at this significance level.

(2) The hypotheses to be tested in this case are:

$$\begin{aligned} H_0 &: \mu_B = 75.0 \\ H_a &: \mu_B \neq 75.0 \end{aligned} \tag{15.40}$$

From the supplied data, the sample average and standard deviation are obtained respectively as $\bar{x}_B = 74.0$, and $s_B = 5.40$, so that the specific value for the T statistic is:

$$t = \frac{74 - 75.0}{5.40/\sqrt{10}} = -0.59 \tag{15.41}$$

Since this is a two-tailed test, the critical values, $t_{0.025}(9)$ and its mirror image $-t_{0.025}(9)$, are obtained from MINITAB as -2.26 and 2.26 implying that the critical/rejection region, R_C, in this case is $t < -2.26$ or $t > 2.26$. But the specific value for the t-statistic (-0.59) does not lie in this region; we therefore *do not* reject H_0 at the significance level 0.05.

The associated p-value, obtained from a t-distribution with 9 degrees of freedom, is:

$$P(t(9) < -0.59 \text{ or } t(9) > 0.59) = P(|t(9)| > 0.59) = 0.572 \tag{15.42}$$

with the implication that we do not reject the null hypothesis, either on the basis of the p-value, or else at the 0.05 significance level, since $p = 0.572$ is larger than 0.05.

Thus, observe that with these two t-tests, we have established, at a significance level of 0.05, that the mean score obtained by trainees using method A is less than 75 while the mean score for trainees using method B is essentially equal to 75. We can, of course, infer from here that this means that method B must be more effective. But there are more direct methods for carrying out tests to compare two means directly, which will be considered shortly.

Using MINITAB

MINITAB can be used to carry out these t-tests directly (without having to compute, by ourselves, first the test statistic and then the critical region, etc.). After entering the data into separate columns, "Method A" and "Method B" in a MINITAB worksheet, for the first problem, the sequence `Stat > Basic Statistics > 1-Sample t` from the MINITAB drop down menu opens a dialog box where one selects the column containing the data ("Method A"); and upon selecting the "Perform hypothesis test" option, one enters the appropriate value for the "Hypothesized mean" (75) and with the "Options" button one selects the desired "Alternative" for H_a (less than) along with the default confidence level (95.0).

MINITAB provides three self-explanatory graphical options: "Histogram of data"; "Individual value plot"; and "Boxplot of data." Our discussion in Chapter 12 about graphical plots for small sample data sets recommends that, with $n = 10$ in this case, the box plot is more reasonable than the histogram for this example.

The resulting MINITAB outputs are displayed as follows:

One-Sample T: Method A

Test of mu = 75 vs < 75

Variable	N	Mean	StDev	SE Mean	95% Upper Bound	T	P
Method A	10	69.00	4.85	1.53	71.81	-3.91	0.002

The box plot along with the 95% confidence interval estimate and the hypothesized mean $H_0 = 75$ are shown in Fig 15.6. The conclusion to reject the null hypothesis in favor of the alternative is clear.

In dealing with the second problem regarding Method B, we follow the same procedure, selecting data in the "Method B" column, but this time, the "Alternative" is selected as "not equal." The MINITAB results are displayed as follows:

One-Sample T: Method B

Test of mu = 75 vs not = 75

Variable	N	Mean	StDev	SE Mean	95% CI	T	P
Method B	10	74.00	5.40	1.71	(70.14, 77.86)	-0.59	0.572

The box plot along with the 95% confidence interval for the mean and the hypothesized mean $H_0 = 75$ are shown in Fig 15.7.

15.3.3 Confidence Intervals and Hypothesis Tests

Interval estimation techniques discussed in Chapter 14 produced estimates for the parameter θ in the form of an interval, $(u_L < \theta < u_R)$, that is expected to contain the unknown parameter with probability $(1 - \alpha)$; it is therefore known as the $(1 - \alpha) \times 100\%$ confidence interval.

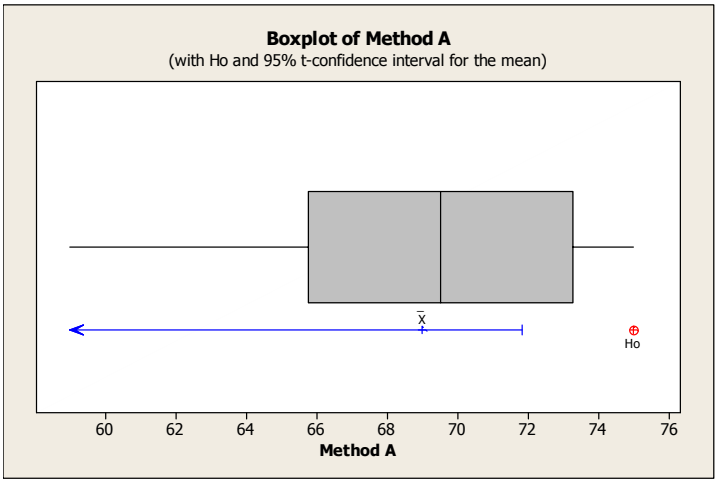

FIGURE 15.6: Box plot for Method A scores including the null hypothesis mean, $H_0 : \mu = 75$, shown along with the sample average, $\bar{x}$, and the 95% confidence interval based on the t-distribution with 9 degrees of freedom. Note how the upper bound of the 95% confidence interval lies to the left of, and does not touch, the postulated H_0 value.

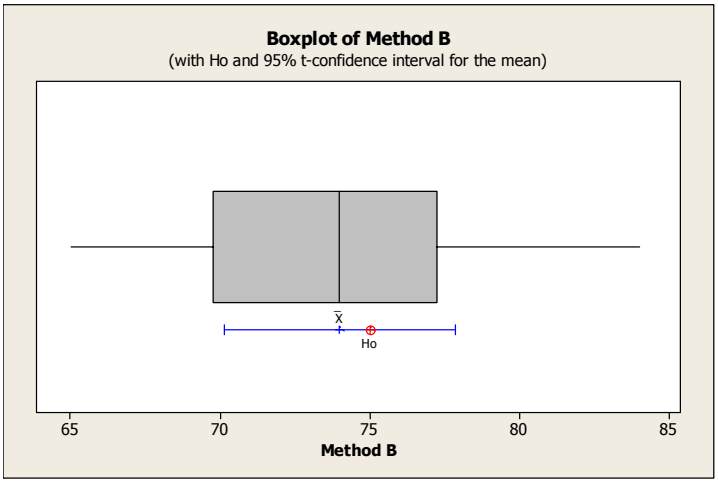

FIGURE 15.7: Box plot for Method B scores including the null hypothesis mean, $H_0, \mu = 75$, shown along with the sample average, $\bar{x}$, and the 95% confidence interval based on the t-distribution with 9 degrees of freedom. Note how the 95% confidence interval includes the postulated H_0 value.

Now, observe first from the definition of the critical/rejection region, R_C, given above, first for a two-tailed test, that at the significance level, α, R_C is precisely complementary to the $(1-\alpha)\times 100\%$ confidence interval for the estimated parameter. The implication therefore is as follows: if the postulated population parameter (say θ_0) falls *outside* the $(1-\alpha)\times 100\%$ confidence interval estimated from sample data (i.e., the postulated value is higher than the upper bound to the right, or lower than the lower bound to the left), this triggers the rejection of H_0, that $\theta = \theta_0$, at the significance level of α, in favor of the alternative H_a, that $\theta \neq \theta_0$. Conversely, if the postulated θ_0 falls within the $(1-\alpha)\times 100\%$ confidence interval, we will fail to reject H_0. This is illustrated in Example 15.2 for the mean yield of process B. The 95% confidence interval was obtained as (70.74, 74.20), which fully encompasses the hypothesized mean value of 72.5; hence we do not reject H_0 at the 0.05 significance level. Similarly, in part 2 of Example 15.4, the 95% confidence interval on the average method B score was obtained as (70.14, 77.86); and with the hypothesized mean, 75, lying entirely in this interval (as shown graphically in Fig 15.7). Once again, we find no evidence to reject H_0 at the 0.05 significance level.

For an upper-tailed test (with H_a defined as $H_a : \theta > \theta_0$), it is the *lower bound* of the $(1-\alpha)\times 100\%$ confidence interval that is now of interest. Observe that if the hypothesized value, θ_0, is to the *left* of this lower bound (i.e., it is lower than the lowest value of the $(1-\alpha)\times 100\%$ confidence interval), the implication is twofold: (i) the computed estimate falls in the rejection region; and, equivalently, (ii) value estimated from data is larger than the hypothesized value—both of which support the rejection of H_0 in favor of H_a, at the significance level of α. This is illustrated in Example 15.3 where the lower bound of the estimated "acting time" for the rat poison was obtained (from MINITAB) as 1002.3 secs, whereas the postulated mean is 1000. H_0 is therefore rejected at the 0.05 significance level in favor of H_a, that the mean value is higher. On the other hand, if the hypothesized value, θ_0, is to the *right* of this lower bound, there will be no support for rejecting H_0 at the 0.05 significance level.

The reverse is true for the lower-tailed test with $H_a : \theta < \theta_0$. The *upper bound* of the $(1-\alpha)\times 100\%$ confidence interval is of interest; and if the hypothesized value, θ_0, is to the *right* of this upper bound (i.e., it is larger than the largest value of the $(1-\alpha)\times 100\%$ confidence interval), this hypothesized value would have fallen into the rejection region. Because this indicates that the value estimated from data is smaller than the hypothesized value, the evidence supports the rejection of H_0 in favor of H_a, at the 0.05 significance level. Again, this is illustrated in part 1 of Example 15.4. The upper bound of the 95% confidence interval on the average method A score was obtained as 71.81, which is lower than the postulated average of 75, thereby triggering the rejection of H_0 in favor of H_a, at the 0.05 significance level (see Fig 15.6). Conversely, when the hypothesized value, θ_0, is to the *left* of this upper bound, we will fail to reject H_0 at the 0.05 significance level.

15.4 Concerning Two Normal Population Means

The problem of interest involves two distinct and *mutually independent* normal populations, with respective unknown means μ_1 and μ_2. In general we are interested in making inference about the difference between these two means, i.e.,

$$\mu_1 - \mu_2 = \delta \tag{15.43}$$

The typical starting point is the null hypothesis,

$$H_0 : \mu_1 - \mu_2 = \delta_0 \tag{15.44}$$

when the difference between the two population means is postulated as some value δ_0, and the hypothesis is to be tested against the usual triplet of possible alternatives:

$$\text{Lower-tailed} \quad H_a : \mu_1 - \mu_2 < \delta_0 \tag{15.45}$$

$$\text{Upper-tailed} \quad H_a : \mu_1 - \mu_2 > \delta_0 \tag{15.46}$$

$$\text{Two-tailed} \quad H_a : \mu_1 - \mu_2 \neq \delta_0 \tag{15.47}$$

In particular, specifying $\delta_0 = 0$ constitutes a test of equality of the two means; but δ_0 does not necessarily have to be zero, allowing us to test the difference against any arbitrary postulated value.

As with tests of single population means, this test will be based on the difference between two random sample means, $\bar{X}_1$ from population 1, and $\bar{X}_2$ from population 2. These tests are therefore known as "two-sample" tests; and, as usual, the specific test to be employed for any problem depends on what additional information is available about each population's standard deviation.

15.4.1 Population Standard Deviations Known

When the population standard deviations, σ_1 and σ_2 are known, we recall (from the discussion in Chapter 14 on interval estimation of the difference of two normal population means) that the test statistic:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1) \tag{15.48}$$

where n_1 and n_2 are the sizes of the samples drawn from populations 1 and 2 respectively. This fact arises from the result established in Chapter 14 for the sampling distribution of $\bar{D} = \bar{X}_1 - \bar{X}_2$ as $N(\delta, v^2)$, with δ as defined in Eq (18.10), and

$$v^2 = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{15.49}$$

TABLE 15.4: Summary of H_0 rejection conditions for the two-sample z-test

Testing Against	For General α Reject H_0 if:	For $\alpha = 0.05$ Reject H_0 if:
$H_a : \mu_1 - \mu_2 < \delta_0$	$z < -z_\alpha$	$z < -1.65$
$H_a : \mu_1 - \mu_2 > \delta_0$	$z > z_\alpha$	$z < 1.65$
$H_a : \mu_1 - \mu_2 \neq \delta_0$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$	$z < -1.96$ or $z > 1.96$

Tests based on this statistic are known as "two-sample z-tests," and as with previous tests, the specific results for testing $H_0 : \mu_1 - \mu_2 = \delta_0$ are summarized in Table 15.4.

Let us illustrate the application of this test with the following example.

Example 15.5: COMPARISON OF SPECIALTY AUXILIARY BACKUP LAB BATTERY LIFETIMES

A company that manufactures specialty batteries used as auxiliary backups for sensitive laboratory equipments in need of constant power supplies claims that its new prototype, brand A, has a longer lifetime (under constant use) than the industry-leading brand B, and at the same cost. Using accepted industry protocol, a series of tests carried out in an independent laboratory produced the following results: For brand A: sample size, $n_1 = 40$; average lifetime, $\bar{x}_1 = 647$ hrs; with a population standard deviation given as $\sigma_1 = 27$ hrs. The corresponding results for brand B are $n_2 = 40; \bar{x}_2 = 638; \sigma_2 = 31$. Determine, at the 5% level, if there is a significant difference between the observed mean lifetimes.

Solution:

Observe that in this case, $\delta_0 = 0$, i.e., the null hypothesis is that the two means are equal; the alternative is that $\mu_1 > \mu_2$, so that the hypotheses are formulated as:

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_a &: \mu_1 - \mu_2 > 0 \end{aligned} \tag{15.50}$$

The specific test statistic obtained from the experimental data is:

$$z = \frac{(647 - 638) - 0}{\sqrt{\frac{27^2}{40} + \frac{31^2}{40}}} = 1.38 \tag{15.51}$$

For this one-tailed test, the critical value, $z_{0.05}$, is 1.65; and now, since the computed z-score is not greater than 1.65, we cannot reject the null hypothesis. There is therefore insufficient evidence to support the rejection of H_0 in favor of H_a, at the 5% significance level.

Alternatively, we could compute the p-value and obtain:

$$\begin{aligned} p = P(Z > 1.38) &= 1 - P(Z < 1.38) \\ &= 1 - 0.916 = 0.084 \end{aligned} \tag{15.52}$$

Once again, since this p-value is greater than 0.05, we cannot reject H_0 in favor of H_a, at the 5% significance level. (However, observe that at the 0.1 significance level, we will reject H_0 in favor of H_a, since the p-value is less than 0.1.)

15.4.2 Population Standard Deviations Unknown

In most practical cases, it is rare that the two population standard deviations are known. Under these circumstances, we are able to identify three distinct cases requiring different approaches:

1. σ_1 and σ_2 unknown; large sample sizes n_1 and n_2;
2. Small sample sizes; σ_1 and σ_2 unknown, but equal (i.e., $\sigma_1 = \sigma_2$);
3. Small sample sizes; σ_1 and σ_2 unknown, and unequal (i.e., $\sigma_1 \neq \sigma_2$).

As usual, under the first set of conditions, the sample standard deviations, s_1 and s_2, are considered to be sufficiently good approximations to the respective unknown population parameters; they are then used in place of σ_1 and σ_2 in carrying out the two-sample z-test as outlined above. Nothing more need be said about this case. We will concentrate on the remaining two cases where the sample sizes are considered to be small.

Equal Standard Deviations

When the two population standard deviations are considered as equal, the test statistic:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}} \sim t(\nu) \tag{15.53}$$

i.e., its sampling distribution is a t-distribution with ν degrees of freedom, with

$$\nu = n_1 + n_2 - 2 \tag{15.54}$$

Here, S_p is the "pooled" sample standard deviation obtained as the positive square root of the pooled sample variance—a weighted average of the two sample variances:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \tag{15.55}$$

a reasonable estimate of the (equal) population variances based on the two sample variances.

From this test statistic and its sampling distribution, one can now carry out the "two-sample t-test," and, once more, the specific results for testing $H_0 : \mu_1 - \mu_2 = \delta_0$ against various alternatives are summarized in Table 15.5.

The following example illustrates these results.

TABLE 15.5: Summary of H_0 rejection conditions for the two-sample t-test

Testing Against	For General α Reject H_0 if:
$H_a : \mu_1 - \mu_2 < \delta_0$	$t < -t_\alpha(\nu)$
$H_a : \mu_1 - \mu_2 > \delta_0$	$t > t_\alpha(\nu)$
$H_a : \mu_1 - \mu_2 \neq \delta_0$	$t < -t_{\alpha/2}(\nu)$ or $t > t_{\alpha/2}(\nu)$ $(\nu = n_1 + n_2 - 2)$

Example 15.6: HYPOTHESES TEST COMPARING EFFECTIVENESS OF ENGINEERING TRAINING PROGRAMS
Revisit the problem in Example 15.1 and this time, at the 5% significance level, test the claim that Method B is more effective. Assume that the scores shown in Example 15.1 come from normal populations with potentially different means, but equal variances.

Solution:
In this case, because the sample size is small for each data set, the appropriate test is a two-sample t-test, with equal variance; the hypotheses to be tested are:

$$H_0 : \mu_A - \mu_B = 0$$
$$H_a : \mu_A - \mu_B < 0 \tag{15.56}$$

Care must be taken in ensuring that H_a is specified properly. Since the claim is that Method B is more effective, if the difference in the means is specified in H_0 as shown (with μ_A first), then the appropriate H_a is as we have specified. (We are perfectly at liberty to formulate H_0 differently, with μ_B first, in which case the alternative hypothesis must change to $H_a : \mu_B - \mu_A > 0$.)

From the sample data, we obtain all the quantities required for computing the test statistic: the sample means, $\bar{x}_A = 69.0, \bar{x}_B = 74.0$; the sample standard deviations, $s_A = 4.85, s_B = 5.40$; so that the estimated pooled standard deviation is obtained as:

$$s_p = 5.13$$

with $\nu = 18$. To test the observed difference ($d = 69.0 - 74.0 = -5.0$) against a hypothesized difference of $\delta_0 = 0$ (i.e., equality of the means), we obtain the t-statistic as:

$$t = -2.18$$

which is compared to the critical value for a t-distribution with 18 degrees of freedom,

$$-t_{0.05}(18) = -1.73$$

And since $t < -t_{0.05}(18)$, we reject the null hypothesis in favor of the alternative, and conclude that, at the 5% significance level, the evidence in the data supports the claim that Method B is more effective.

Note also that the associated p-value, obtained from a t distribution with 18 degrees of freedom, is:

$$P(t(18) < -2.18) = 0.021 \tag{15.57}$$

which, by virtue of being less than 0.05 recommends rejection of H_0 in favor of H_a, at the 5% significance level, as we already concluded above.

Using MINITAB

This just-concluded example illustrates the "mechanics" of how to conduct a two-sample t-test "manually"; once the mechanics are understood, however, it is recommended to use computer programs such as MINITAB.

As noted before, once the data sets have been entered into separate columns "Method A" and "Method B" in a MINITAB worksheet (as was the case in Example 15.4), the required sequence from the MINITAB drop down menu is: `Stat > Basic Statistics > 2-Sample t`, which opens a dialog box with self-explanatory options. Once the location of the relevant data are identified, the "Assume equal variance" box is selected in this case, and with the "Options" button, one selects the "Alternative" for H_a ("less than," if the hypotheses are set up as we have done above), along with the default confidence level (95.0); one enters the value for hypothesized difference, δ_0, in the "Test difference" box (0 in this case). The resulting MINITAB outputs for this problem are displayed as follows:

Two-Sample T-Test and CI: Method A, Method B

```
Two-sample T for Method A vs Method B
```

	N	Mean	StDev	SE Mean
Method A	10	69.00	4.85	1.5
Method B	10	74.00	5.40	1.7

```
Difference = mu (Method A) - mu (Method B)
Estimate for difference: -5.00
95% upper bound for difference: -1.02
T-Test of difference = 0 (vs <): T-Value = -2.18 P-Value = 0.021 DF
= 18
Both use Pooled StDev = 5.1316
```

Unequal Standard Deviations

When $\sigma_1 \neq \sigma_2$, things become a bit more complicated, and a detailed discussion lies outside the intended scope of this book. Suffice it to say that under these circumstances, the universally recommended test statistic is $\tilde{T}$

defined as:

$$\tilde{T} = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}, \tag{15.58}$$

which appears deceptively like Eq (15.53), with the very important difference that S_1 and S_2 have been reinstated individually in place of the pooled S_p. Of course, this expression is also reminiscent of the Z statistic in Eq (15.48), with S_1 and S_2 introduced in place of the population variances. However, unlike the other single variable cases where such a substitution transforms the standard normal sampling distribution to the t-distribution with the appropriate degrees of freedom, unfortunately, this time, this test statistic only has an *approximate* (not exact) t-distribution; and the degrees of freedom, ν, accompanying this approximate t-distribution is defined by:

$$\nu = \tilde{n}_{12} - 2 \tag{15.59}$$

with $\tilde{n}_{12}$ defined by the formidable-looking expression

$$\tilde{n}_{12} = \left\{ \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} \right\} \tag{15.60}$$

rounded to the nearest integer.

Under these conditions, the specific results for carrying out two-sample t-tests for testing $H_0 : \mu_1 - \mu_2 = \delta_0$ against various alternatives are summarized in Table 15.5 but with $\tilde{t}$ in place of the corresponding t-values, and using ν given above in Eqs (15.59) and (15.60) for the degrees of freedom.

Although it is possible to carry out such two-sample t-tests "manually" by computing the required quantities on our own, it is highly recommended that such tests be carried out using computer programs such as MINITAB.

Confidence Intervals and Two-Sample Tests

The relationship between confidence intervals for the difference between two normal population means and the two-sample tests discussed above perfectly mirrors the earlier discussion concerning single means of a normal population. For the two-sided test, a $(1-\alpha)\times 100\%$ confidence interval estimate for the difference between the two means that does not contain the hypothesized mean corresponds to a hypothesis test in which H_0 is rejected, at the significance level of α, in favor of the alternative that the computed difference is not equal to the hypothesized difference. Note that with a test of equality (in which case δ_0, the hypothesized difference, is 0), rejection of H_0 is tantamount to the $(1-\alpha)\times 100\%$ confidence interval for the difference not containing 0. On the contrary, an estimated $(1-\alpha)\times 100\%$ confidence interval that contains the hypothesized difference is equivalent to a two-sample test that must fail to reject H_0.

The corresponding arguments for the upper-tailed and lower-tailed tests

follow precisely as presented earlier. For an upper-tailed test, ($H_a : \delta > \delta_0$), a *lower bound* of the $(1-\alpha)\times 100\%$ confidence interval estimate of the difference, δ, that is larger than the hypothesized difference, δ_0, corresponds to a two-sample test in which H_0 is rejected in favor of H_a, at the significance level of α. Conversely, a *lower bound* of the confidence interval estimate of the difference, δ, that is smaller than the hypothesized difference, δ_0, corresponds to a test that will not reject H_0. The reverse is the case for the lower-tailed test ($H_a : \delta < \delta_0$): when the upper bound of the $(1-\alpha)\times 100\%$ confidence interval estimate of δ is smaller than δ_0, H_0 is rejected in favor of H_a. When the upper bound of the $(1-\alpha)\times 100\%$ confidence interval estimate of δ is larger than δ_0, H_0 *is not* rejected.

An Illustrative Example: The Yield Improvement Problem

The solution to the yield improvement problem first posed in Chapter 1, and revisited at the beginning of this chapter, will finally be completed in this illustrative example. In addition, the example also illustrates the use of MINITAB to carry out a two-sample t-test when population variances are not equal.

The following questions remain to be resolved: Is $Y_A > Y_B$, and if so, is $Y_A - Y_B > 2$? Having already confirmed that the random variables, Y_A and Y_B, can be characterized reasonably well with Gaussian distributions, $N(\mu_A, \sigma_A^2)$ and $N(\mu_B, \sigma_B^2)$, respectively, the supplied data may then be considered as being from normal distributions with *unequal* population variances. We will answer these two questions by carrying out appropriate two-sample t-tests.

Although the answer to the first of the two questions requires testing for the equality of μ_A and μ_B against the alternative that $\mu_A > \mu_B$, let us begin by first testing against $\mu_A \neq \mu_B$; this establishes that the two distributions means are different. Later we will test against the alternative that $\mu_A > \mu_B$, and thereby go beyond the mere existence of a difference between the population means to establish which is larger. Finally, we proceed even one step further to establish not only which one is larger, but that it is larger by a value that exceeds a certain postulated value (in this case 2).

For the first test of basic equality, the hypothesized difference is clearly $\delta_0 = 0$, so that:

$$\begin{aligned} H_0 &: \mu_A - \mu_B = 0 \\ H_a &: \mu_A \neq \mu_B = 0 \end{aligned} \tag{15.61}$$

The procedure for using MINITAB is as follows: upon entering the data into separate Y_A and Y_B columns in a MINITAB worksheet, the required sequence from the MINITAB drop down menu is: `Stat > Basic Statistics > 2-Sample t`. In the opened dialog box, one simply selects the "Samples in different columns" option, identifies the columns corresponding to each data set, but this time, the "Assume equal variance" box must not be selected. With the "Options" button one selects the "Alternative" for H_a as "not equal,"

along with the default confidence level (95.0); in the "Test difference" box, one enters the value for hypothesized difference, δ_0; 0 in this case. The resulting MINITAB outputs for this problem are displayed as follows:

Two-Sample T-Test and CI: YA, YB

```
Two-sample T for YA vs YB
```

	N	Mean	StDev	SE Mean
YA	50	75.52	1.43	0.20
YB	50	72.47	2.76	0.39

```
Difference = mu (YA) - mu (YB)
Estimate for difference: 3.047
95% CI for difference: (2.169, 3.924)
T-Test of difference = 0 (vs not =): T-Value = 6.92 P-Value = 0.000
DF = 73
```

Several points are worth noting here:

1. The most important is the p-value which is virtually zero; the implication is that at the 0.05 significance level, we must reject the null hypothesis in favor of the alternative: the two population means are in fact different, i.e., the observed difference between the population is *not* zero. Note also that the t-statistic value is 6.92, a truly extreme value for a distribution that is symmetrical about the value 0, and for which the density value, $f(t)$ essentially vanishes (i.e., $f(t) \approx 0$), for values of the t variate exceeding ± 4. The p-value is obtained as $P(|T| > 6.92)$.

2. The estimated sample difference is 3.047, with a 95% confidence interval, (2.169, 3.924); since this interval does not contain the hypothesized difference $\delta_0 = 0$, the implication is that the test will reject H_0, as indeed we have concluded in point #1 above;

3. Finally, even though there were 50 data entries each for Y_A and Y_B, the degrees of freedom associated with this test is obtained as 73. (See the expressions in Eqs (15.59) and (15.60) above.)

This first test has therefore established that the means of the Y_A and Y_B populations are different, at the 5% significance level. Next, we wish to test which of these two different means is larger. To do this, the hypotheses to be tested are:

$$\begin{aligned} H_0 &: \mu_A - \mu_B = 0 \\ H_a &: \mu_A > \mu_B = 0 \end{aligned} \tag{15.62}$$

The resulting outputs from MINITAB are identical to what is shown above for the first test, with two exceptions:

(i) the "95% CI for difference" line is replaced with 95% lower bound for difference: 2.313; and

(ii) the "T-Test of difference = 0 (vs not =)" is replaced with T-Test of difference = 0 (vs >).

The t-value, p-value, and "DF" remain the same.

Again, with a p-value that is virtually zero, the conclusion is that, at the 5% significance level, the null hypothesis must be rejected in favor of the alternative, which, this time, is specifically that μ_A is greater than μ_B. Note that the value 2.313, computed from the data as the 95% lower bound for the difference, is considerably higher than the hypothesized value of 0; i.e., the hypothesized $\delta_0 = 0$ lies well to the left of this lower bound for the difference. This is consistent with rejecting the null hypothesis in favor of the alternative, at the 5% significance level.

With the final test, we wish to sharpen the postulated difference a bit further. This time, we assert that, μ_A is not only greater than μ_B; the former is in fact greater than the latter by a value that exceeds 2. The hypotheses are set up in this case as follows:

$$\begin{aligned} H_0 &: \mu_A - \mu_B = 2 \\ H_a &: \mu_A > \mu_B = 2 \end{aligned} \tag{15.63}$$

This time, in the MINTAB options, the new hypothesized difference is indicated as 2 in the "Test difference" box. The MINITAB results are displayed as follows:

Two-Sample T-Test and CI: YA, YB

Two-sample T for YA vs YB

	N	Mean	StDev	SE Mean
YA	50	75.52	1.43	0.20
YB	50	72.47	2.76	0.39

```
Difference = mu (YA) - mu (YB)
Estimate for difference: 3.047
95% lower bound for difference: 2.313
T-Test of difference = 2 (vs >): T-Value = 2.38 P-Value = 0.010 DF
= 73
```

Note that the t-value is now 2.38 (reflecting the new hypothesized value of $\delta_0 = 2$), with the immediate consequence that the p-value is now 0.01; not surprisingly, everything else remains the same as in the first test. Thus, at the 0.05 significance level, we reject the null hypothesis in favor of the alternative. Note also that the 95% lower bound for the difference is larger than the hypothesized difference of 2.

The conclusion is therefore that, with 95% confidence (or alternatively at a significance level of 0.05), the mean yield obtainable from the challenger

TABLE 15.6: "Before" and "after" weights for patients on a supervised weight-loss program

Patient #	1	2	3	4	5	6	7	8	9	10
Before Wt (lbs)	272	319	253	325	236	233	300	260	268	276
After Wt (lbs)	263	313	251	312	227	227	290	251	262	263
Patient #	11	12	13	14	15	16	17	18	19	20
Before Wt (lbs)	215	245	248	364	301	203	197	217	210	223
After Wt (lbs)	206	235	237	350	288	195	193	216	202	214

process A is *at least* 2 points larger than that obtainable by the incumbent process B.

15.4.3 Paired Differences

A subtle but important variation on the theme of inference concerning two normal population means arises when the data naturally occur in pairs, as with the data shown in Table 15.6. This is a record of the "before" and "after" weights (in pounds) of twenty patients enrolled in a clinically-supervised 10-week weight-loss program. Several important characteristics set this problem apart from the general two-sample problem:

1. For each patient, the random variable "Weight" naturally occurs as an ordered pair of random variables (X, Y), with X as the "before" weight, and Y as the "after" weight;

2. As a result, it is highly unlikely that the two entries per patient will be totally independent, i.e., the random sample, $X_1, X_2, \ldots, X_n$, will likely *not* be independent of $Y_1, Y_2, \ldots, Y_n$;

3. In addition, the sample sizes for each random sample, $X_1, X_2, \ldots, X_n$, and $Y_1, Y_2, \ldots, Y_n$, by definition, will be identical;

4. Finally, it is quite possible that the patient-to-patient variability in each random variable X or Y (i.e., the variability *within* each group) may be much larger than the difference *between* the groups that we seek to detect.

These circumstances call for a different approach, especially in light of item #2 above, which invalidates one of the most crucial assumptions underlying the two-sample tests: independence of the random samples.

The analysis for this class of problems proceeds as follows. Let $(X_i, Y_i); i = 1, 2, \ldots, n$, be an ordered pair of random samples, where $X_1, X_2, \ldots, X_n$ is from a normal population with mean, μ_X, and variance, σ_X^2; and $Y_1, Y_2, \ldots, Y_n$, a random sample from a normal population with mean, μ_Y, and variance, σ_Y^2. Define the difference D as:

$$D_i = X_i - Y_i \tag{15.64}$$

then, $D_i, i = 1, 2, \ldots, n$, constitutes a random sample of differences with mean value,

$$\delta = \mu_X - \mu_Y \tag{15.65}$$

The quantities required for the hypothesis test are: the sample average,

$$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n} \tag{15.66}$$

(which is unbiased for δ), and the sample variance of the differences,

$$S_D^2 = \frac{\sum_{i=1}^{n}(D_i - \bar{D})^2}{n-1} \tag{15.67}$$

Under these circumstances, the null hypothesis is defined as

$$H_0 : \delta = \delta_0 \tag{15.68}$$

when δ, the difference between the paired observations, is postulated as some value δ_0. This hypothesis, as usual, is to be tested against the possible alternatives

$$\text{Lower-tailed} \quad H_a : \delta < \delta_0 \tag{15.69}$$

$$\text{Upper-tailed} \quad H_a : \delta > \delta_0 \tag{15.70}$$

$$\text{Two-tailed} \quad H_a : \delta \neq \delta_0 \tag{15.71}$$

The appropriate test statistic is

$$T = \frac{\bar{D} - \delta_0}{S_D/\sqrt{n}} \tag{15.72}$$

it possesses a $t(n-1)$ distribution. When used to carry out what is generally known as the "paired t-test," the results are similar to those obtained for earlier tests, with the specific rejection conditions summarized in Table 15.7. The next two examples illustrate the importance of distinguishing between a paired-test and a general two-sample test.

Example 15.7: WEIGHT-LOSS DATA ANALYSIS: PART 1
By treating the weight-loss patient data in Table 15.6 as "before" and "after" ordered pairs, determine at the 5% level, whether or not the weight loss program has been effective in assisting patients lose weight.

Solution:
This problem requires determining whether the mean difference between the "before" and "after" weights for the 20 patients is significantly different from zero. The null and alternative hypotheses are:

$$\begin{aligned} H_0 &: \delta = 0 \\ H_a &: \delta \neq 0 \end{aligned} \tag{15.73}$$

TABLE 15.7: Summary of H_0 rejection conditions for the paired t-test

Testing Against	For General α Reject H_0 if:
$H_a : \delta < \delta_0$	$t < -t_\alpha(\nu)$
$H_a : \delta > \delta_0$	$t > t_\alpha(\nu)$
$H_a : \delta \neq \delta_0$	$t < -t_{\alpha/2}(\nu)$ or $t > t_{\alpha/2}(\nu)$ $(\nu = n - 1)$

We can compute the twenty "before"-minus-"after" weight differences, obtain the sample average and sample standard deviation of these differences, and then compute the t-statistic from Eq (15.72) for $\delta_0 = 0$. How this t statistic compares against the critical value of $t_{0.025}(19)$ will determine whether or not to reject the null hypothesis.

We can also use MINITAB directly. After entering the data into two columns "Before WT" and "After WT", the sequence: `Stat > Basic Statistics > Paired t` opens the usual analysis dialog box: as with other hypothesis tests, data columns are identified, and with the "Options" button, the "Alternative" for H_a is selected as "not equal," along with 0 for the "Test mean" value, with the default confidence level (95.0). The resulting MINITAB outputs for this problem are displayed as follows:

Paired T-Test and CI: Before WT, After WT

`Paired T for Before WT - After WT`

	N	Mean	StDev	SE Mean
Before WT	20	258.2	45.2	10.1
After WT	20	249.9	43.3	9.7
Difference	20	8.400	3.662	0.819

```
95% CI for mean difference: (6.686, 10.114)
T-Test of mean difference = 0 (vs not = 0): T-Value = 10.26 P-Value
= 0.000
```

The mean difference (i.e., average weight-loss per patient) is 8.4 lbs, and the 95% confidence interval (6.686, 10.114), does not contain 0; also, the p-value is 0 (to three decimal places). The implication is therefore that at the significance level of 0.05, we reject the null hypothesis and conclude that the weight-loss program was effective. The average weight loss of 8.4 lbs is therefore significantly different from zero, at the 5% significance level.

A box plot of the differences between the "before" and "after" weights is shown in Fig 15.8, which displays graphically that the null hypothesis should be rejected in favor of the alternative. Note how far

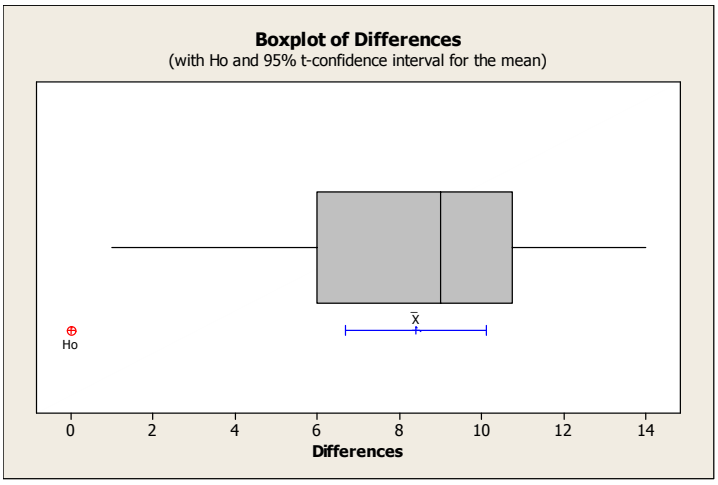

FIGURE 15.8: Box plot of differences between the "before" and "after" weights, including a 95% confidence interval for the mean difference, and the hypothesized H_0 point, $\delta_0 = 0$.

the hypothesized value of 0 is from the 95% confidence interval for the mean weight difference.

The next example illustrates the consequences of wrongly employing a two-sample t-test for this natural paired t-test problem.

Example 15.7: WEIGHT-LOSS DATA ANALYSIS: PART 2: TWO-SAMPLE T-TEST
Revisit the problem in Example 15.6 but this time treat the "before" and "after" weight data in Table 15.6 as if they were independent samples from two different normal populations; carry out a 2-sample t-test and, at the 5% level, determine whether or not the two sample means are different.

Solution:
First let us be very clear: this is *not* the right thing to do; but if a 2-sample t-test is carried out on this data set with the hypotheses as:

$$\begin{aligned} H_0 &: \mu_{before} - \mu_{after} = 0 \\ H_a &: \mu_{before} - \mu_{after} \neq 0 \end{aligned} \tag{15.74}$$

MINITAB produces the following result:

Two-Sample T-Test and CI: Before WT, After WT

Two-sample T for Before WT vs After WT

	N	Mean	StDev	SE Mean
Before WT	20	258.2	45.2	10.1
After WT	20	249.9	43.3	9.7

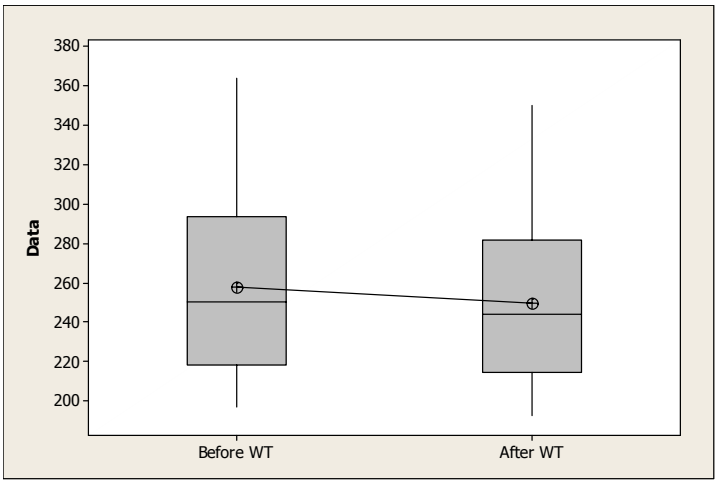

FIGURE 15.9: Box plot of the "before" and "after" weights including individual data means. Notice the wide range of each data set.

```
Difference = mu (Before WT) - mu (After WT)
Estimate for difference: 8.4
95% CI for difference: (-20.0, 36.8)
T-Test of difference = 0 (vs not =): T-Value = 0.60 P-Value =
0.552 DF = 38
Both use Pooled StDev = 44.2957
```

With a t-value of 0.6 and a p-value of 0.552, this analysis indicates that there is no evidence to support rejecting the null hypothesis at the significance level of 0.05. The estimated *difference* of the means is 8.4 (the same as the *mean* of the differences obtained in Example 15.6); but because of the large pooled standard deviation, the 95% confidence interval is $(-20.0, 36.8)$, which includes 0. As a result, the null hypothesis cannot be rejected at the 5% significance level in favor of the alternative. This, of course, will be the wrong decision (as the previous example has shown) and should serve as a warning against using the two-sample t-test improperly for paired data.

It is important to understand the sources of the failure in this last example. First, a box plot of the two data sets, shown in Fig 15.9, graphically illustrates why the two-sample t-test is entirely unable to detect the very real, and very significant, difference between the "before" and "after" weights. The variability *within* the samples is so high that it swamps out the difference *between* each pair which is actually significant. But the most important reason is illustrated in Fig 15.10, which shows a plot of "before" and "after" weights for each patient versus patient number, from where it is absolutely clear, that the two sets of weights are almost perfectly correlated. Paired data are often

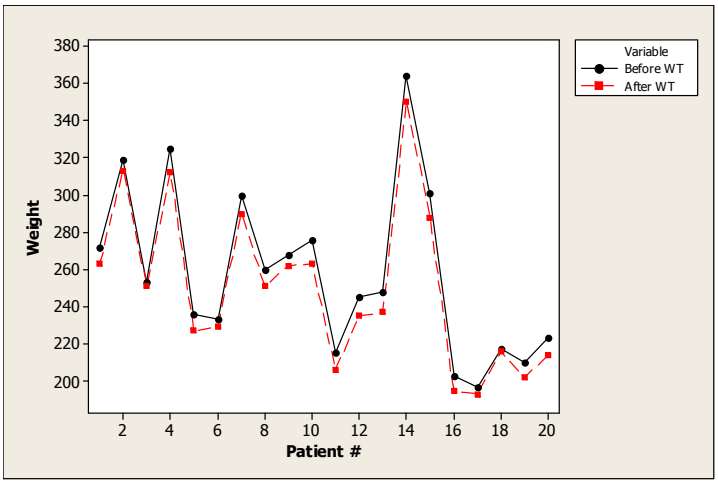

FIGURE 15.10: A plot of the "before" and "after" weights for each patient. Note how one data sequence is almost perfectly correlated with the other; in addition note the relatively large variability intrinsic in each data set compared to the difference between each point.

not independent. Observe from the data (and from this graph) that without exception, every single "before" weight is higher than the corresponding "after" weight. The issue is therefore not whether there is a weight loss; it is a question of how much. For this group of patients, however, this difference cannot be detected in the midst of the large amount of variability *within* each group ("before" or "after").

These are the primary reasons that the two-sample t-test failed miserably in identifying a differential that is quite significant. (As an exercise, the reader should obtain a scatter plot of the "before" weight versus the "after" weight to provide further graphical evidence of just how correlated the two weights are.)

15.5 Determining β, Power, and Sample Size

Determining β, the Type II error risk, and hence $(1-\beta)$, the power of any hypothesis test, depends on whether the test is one- or two-sided. The same is also true of the complementary problem: the determination of experimental sample sizes required to achieve a certain pre-specified power. We begin our discussion of such issues with the one-sided test, specifically the upper-tailed test, with the null hypothesis as in Eq (15.16) and the alternative in Eq

(15.18). The results for the lower-tailed, and the two-sided tests, which follow similarly, will be given without detailed derivations.

15.5.1 β and Power

To determine β (and hence power) for the upper-tailed test, it is not sufficient merely to state that $\mu > \mu_0$; instead, one must specify a particular value for the alternative mean, say μ_a, so that:

$$H_a : \mu = \mu_a > \mu_0 \tag{15.75}$$

is the alternative hypothesis. The Type II error risk is therefore the probability of failing to reject the null hypothesis when in truth the data came from the alternative distribution with mean μ_a (where, for the upper-tailed test, $\mu_a > \mu_0$).

The difference between this alternative and the postulated null hypothesis distribution mean,

$$\delta^* = \mu_a - \mu_0 \tag{15.76}$$

is the margin by which the null hypothesis is falsified in comparison to the alternative. As one might expect, the magnitude of δ^* will be a factor in how easy or difficult it is for the test to detect, amidst all the variability in the data, a difference between H_0 and H_a, and therefore correctly reject H_0 when it is false. (Equivalently, the magnitude of δ^* will also factor into the risk of incorrectly failing to reject H_0 in favor of a true H_a.)

As shown earlier, if H_0 is true, then the distribution of the sample mean, $\bar{X}$, is $N(\mu_0, \sigma^2/n)$, so that the test statistic, Z, in Eq (15.20), possesses a standard normal distribution; i.e.,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{15.77}$$

However, if H_a is true, then in fact the more appropriate distribution for $\bar{X}$ is $N(\mu_a, \sigma^2/n)$. And now, because $E(\bar{X}) = \mu_a$ under these circumstances, not μ_0 as postulated, the most important implication is that the distributional characteristics of the computed Z statistic, instead of following the standard normal distribution, will be:

$$Z \sim N\left(\frac{\delta^*}{\sigma/\sqrt{n}}, 1\right) \tag{15.78}$$

i.e., the standard normal distribution shifted to the right (for this upper-tailed test) by a factor of $(\delta^*\sqrt{n})/\sigma$. Thus, as a result of a true differential, δ^*, between alternative and null hypothesized means, the standardized alternative distribution will show a "z-shift"

$$z_{shift} = \frac{\delta^*\sqrt{n}}{\sigma} \tag{15.79}$$

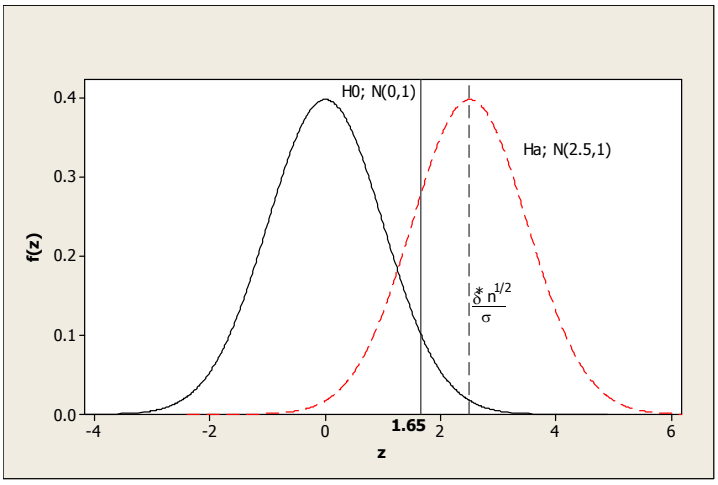

FIGURE 15.11: Null and alternative hypotheses distributions for upper-tailed test based on $n = 25$ observations, with population standard deviation $\sigma = 4$, where the true alternative mean, μ_a, exceeds the hypothesized one by $\delta^* = 2.0$. The figure shows a "z-shift" of $(\delta^*\sqrt{n})/\sigma = 2.5$; and with reference to H_0, the critical value $z_{0.05} = 1.65$. The area under the H_0 curve to the *right* of the point $z = 1.65$ is $\alpha = 0.05$, the significance level; the area under the dashed H_a curve to the *left* of the point $z = 1.65$ is β.

For example, for a test based on 25 observations, with population standard deviation $\sigma = 4$ where the true alternative mean, μ_a, exceeds the hypothesized one by $\delta^* = 2.0$, the mean value of the standardized alternative distribution, following Eq (15.78), will be 2.5, and the two distributions will be as shown in Fig 15.11, with the alternative hypothesis distribution shown with the dashed line.

In terms of the standard normal variate, z, under H_0, the shifted variate under the alternative hypothesis, H_a, is:

$$\zeta = z - \frac{\delta^*\sqrt{n}}{\sigma} \tag{15.80}$$

And now, to compute β, we recall that, by definition,

$$\beta = P(z < z_\alpha | H_a) \tag{15.81}$$

which, by virtue of the "z-shift" translates to:

$$\beta = P\left(z < z_\alpha - \frac{\delta^*\sqrt{n}}{\sigma}\right) \tag{15.82}$$

from where we obtain the expression for the power of the test as:

$$(1 - \beta) = 1 - P\left(z < z_\alpha - \frac{\delta^*\sqrt{n}}{\sigma}\right) \tag{15.83}$$

Thus, for the illustrative example test given above, based on 25 observations, with $\sigma = 4$ and $\mu_a - \mu_0 = \delta^* = 2.0$, the β-risk and power are obtained as

$$\begin{aligned} \beta &= P(z < 1.65 - 2.5) = 0.198 \\ \text{Power} &= (1 - \beta) = 0.802 \end{aligned} \tag{15.84}$$

as shown in Fig 15.12.

15.5.2 Sample Size

In the same way in which z_α was defined earlier, let z_β be the standard normal variate such that

$$P(z > z_\beta) = \beta \tag{15.85}$$

so that, by symmetry,

$$P(z < -z_\beta) = \beta \tag{15.86}$$

Then, from Eqs (15.82) and (15.86) we obtain:

$$-z_\beta = z_\alpha - \frac{\delta^* \sqrt{n}}{\sigma} \tag{15.87}$$

which rearranges to give the important expression,

$$z_\alpha + z_\beta = \frac{\delta^* \sqrt{n}}{\sigma} \tag{15.88}$$

which relates the α- and β-risk variates to the three hypothesis test characteristics: δ^*, the hypothesized mean shift to be detected by the test (the "signal"); σ, the population standard deviation, a measure of the variability inherent in the data (the "noise"); and finally, n, the number of experimental observations to be used to carry out the hypothesis test (the "sample size"). (Note that these three terms comprise what we earlier referred to as the "z-shift," the precise amount by which the standardized H_a distribution has been shifted away from the H_0 distribution; see Fig 15.11.)

This relationship, fundamental to power and sample size analyses, can also be derived in terms of the unscaled critical value, x_C, which marks the boundary of the rejection region for the unscaled sample mean.

Observe that by definition of the significance level, α, the critical value, and the Z statistic,

$$z_\alpha = \frac{x_C - \mu_0}{\sigma / \sqrt{n}} \tag{15.89}$$

so that:

$$x_C = z_\alpha \frac{\sigma}{\sqrt{n}} + \mu_0 \tag{15.90}$$

By definition of β, under H_a,

$$\beta = P\left(z < \frac{x_C - \mu_a}{\sigma / \sqrt{n}} \right) \tag{15.91}$$

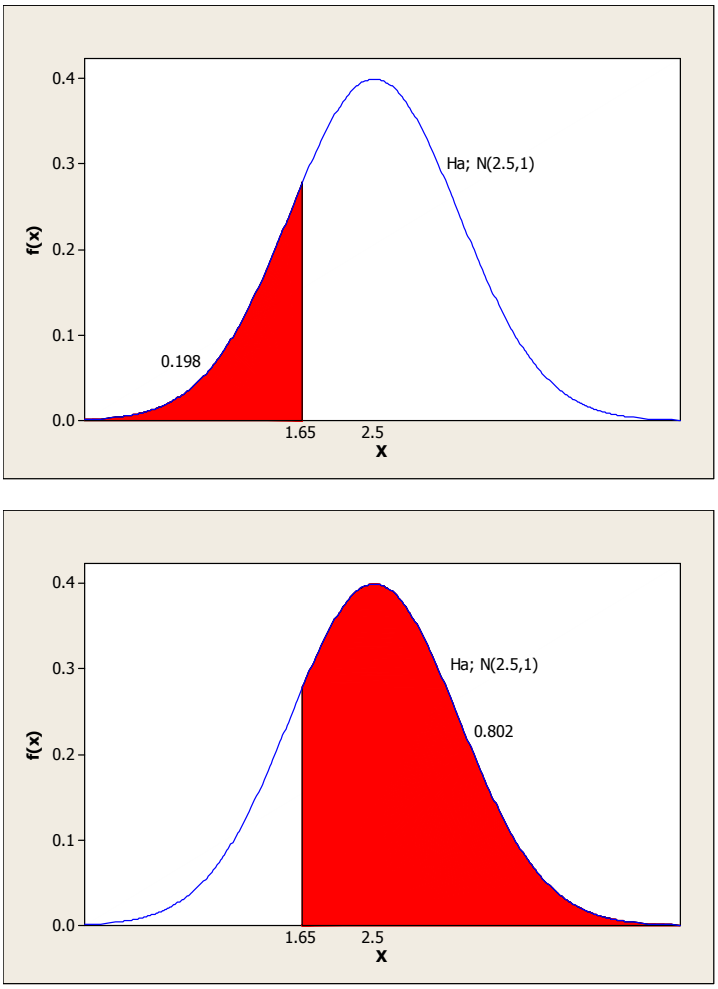

FIGURE 15.12: β and power values for hypothesis test of Fig 15.11 with $H_a \sim N(2.5, 1)$. Top: β; Bottom: Power $= (1 - \beta)$.

and from the definition of the z_β variate in Eq (15.86), we obtain:

$$-z_\beta = \frac{x_C - \mu_a}{\sigma/\sqrt{n}} \tag{15.92}$$

and upon substituting Eq (15.90) in for x_C, and recalling that $\mu_a - \mu_0 = \delta^*$, Eq (15.92) immediately reduces to

$$\begin{aligned} -z_\beta &= z_\alpha - \frac{\delta^*\sqrt{n}}{\sigma}, \text{or} \\ z_\alpha + z_\beta &= \frac{\delta^*\sqrt{n}}{\sigma} \end{aligned} \tag{15.93}$$

as obtained earlier from the standardized distributions.

Several important characteristics of hypothesis tests are embedded in this important expression that are worth drawing out explicitly; but first, a general statement regarding z-variates and risks. Observe that any tail area, τ, *decreases* as $|z_\tau|$ *increases*; similarly, tail area, τ, increases as z_τ decreases; similarly, tail area, τ, increases as $|z_\tau|$ decreases. We may thus note the following about Eq (15.93):

1. The equation shows that for any particular hypothesis test with fixed characteristics δ^*, σ, and n, there is a conservation of the sum of the α- and β-risk variates; if z_α increases, z_β must decrease by a commensurate amount, and vice versa.

2. Consequently, if, in order to reduce the α-risk, z_α is increased, z_β will decrease commensurately to maintain the left-hand side sum constant, with the result that the β-risk must automatically increase. The reverse is also true: increasing z_β for the purpose of reducing the β-risk will result in z_α decreasing to match the increase in z_β, so that the α risk will then increase. Therefore, *for a fixed set of test characteristics, the associated Type I and Type II risks are such that a reduction in one risk will result in an increase in the other in mutual fashion.*

3. The only way to reduce either risk *simultaneously* (which will require increasing the total sum of the risk variates) is by increasing the "z-shift." This is achievable most directly by increasing n, the sample size, since neither σ, the population standard deviation, nor δ^*, the hypothesized mean shift to be detected by the test, is usually under the direct control of the experimenter.

This last point leads directly to the issue of determining how many experimental samples are required to attain a certain power, given basic test characteristics. This question is answered by solving Eq (15.88) explicitly for n to obtain:

$$n = \left[\frac{(z_\alpha + z_\beta)\sigma}{\delta^*}\right]^2 \tag{15.94}$$

Thus, by specifying the desired α- and β-risks along with the test characteristics, δ^*, the hypothesized mean shift to be detected by the test, and σ, the population standard deviation, one can use Eq (15.94) to determine the sample size required to achieve the desired risk levels. In particular, it is customary to specify the risks as $\alpha = 0.05$ and $\beta = 0.10$, in which case, $z_\alpha = z_{0.05} = 1.645$; and $z_\beta = z_{0.10} = 1.28$. Eq (15.94) then reduces to:

$$n = \left(\frac{2.925\sigma}{\delta^*}\right)^2 \tag{15.95}$$

from which, given δ^* and σ, one can determine n.

Example 15.8: SAMPLE SIZE REQUIRED TO IMPROVE POWER OF HYPOTHESIS TEST

The upper-tailed hypothesis test illustrated in Fig 15.11 was shown in Eq (15.84) to have a power of 0.802 (equivalent to a β-risk of 0.182). It is based on a sample size of $n = 25$ observations, population standard deviation $\sigma = 4$, and where the true alternative mean μ_a exceeds the hypothesized one by $\delta^* = 2.0$. Determine the sample size required to improve the power from 0.802 to the customary 0.9.

Solution:

Upon substituting $\sigma = 4; \delta^* = 2$ into Eq (15.95), we immediately obtain $n = 34.2$, which should be rounded up to the nearest integer to yield 35. This is the required sample size, an increase of 10 additional observations. To compute the actual power obtained with $n = 35$ (since it is technically different from the precise, but impractical, $n = 34.2$ obtained from Eq (15.95)), we introduce $n = 35$ in Eq (15.94) and obtain the corresponding z_β as 1.308; from here we may obtain β from MINITAB's cumulative probability feature as $\beta = 0.095$, and hence

$$\text{Power} = 1 - \beta = 0.905 \tag{15.96}$$

is the actual power.

Practical Considerations

In practice, prior to performing the actual hypothesis test, no one knows whether or not H_a is true compared to H_0; it is even less likely that one will know the precise amount by which μ_a will exceed the postulated μ_0 should H_a turn out to be true. The implication therefore is that δ^* is never known in an objective fashion *à-priori*. In determining the power of a hypothesis test, therefore, δ^* is treated not as "known" but as a *design parameter*: the minimum difference we would like to detect, if such a difference exists. Thus, δ^* is to be considered properly as the magnitude of the smallest difference we wish to detect with the hypothesis test.

In a somewhat related vein, the population standard deviation, σ, is rarely known *à priori* in many practical cases. Under these circumstances, it has

TABLE 15.8: Sample size n required to achieve a power of 0.9 for various values of signal-to-noise ratio, ρ_{SN}

ρ_{SN}	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.2	1.5
n	95.06	53.47	34.22	23.77	17.46	13.37	10.56	8.56	5.94	3.80
n^+	96	53	35	24	18	14	11	9	6	4

often been recommended to use educated guesses, or results from prior experiments under similar circumstances, to provide pragmatic surrogates for σ. We strongly recommend an alternative approach: casting the problem in terms of the "signal-to-noise" ratio (SNR):

$$\rho_{SN} = \frac{\delta^*}{\sigma} \tag{15.97}$$

a ratio of the magnitude of the "signal" (difference in the means) to be detected and the intrinsic "noise" (population standard deviation) in the midst of which the signal is to be detected. In this case, the general Eq (15.94), and the more specific Eq (15.95) become:

$$\begin{aligned} n &= \left[\frac{(z_\alpha + z_\beta)}{\rho_{SN}}\right]^2 \\ n &= \left(\frac{2.925}{\rho_{SN}}\right)^2 \end{aligned} \tag{15.98}$$

Without necessarily knowing either δ^* or σ independently, the experimenter then makes a sample-size decision by designing for a test to handle a "design" SNR.

Example 15.9: SAMPLE SIZE TABLE FOR VARIOUS SIGNAL-TO-NOISE RATIOS: POWER OF 0.9
Obtain a table of sample sizes required to achieve a power of 0.9 for various signal-to-noise ratios from 0.3 to 1.5.

Solution:
Table 15.8 is generated from Eq (15.98) for the indicated values of the signal-to-noise ratio, where n^+ is the value of the computed n rounded up to the nearest integer. As expected, as the signal-to-noise ratio improves, the sample size required to achieve a power of 0.9 reduces; fewer data points are required to detect signals that are large relative to the standard deviation. Note in particular that for the example considered in Fig 15.11 and Example 15.8, $\rho_{SN} = 2/4 = 0.5$; from Table 15.8, the required sample size, 35, is precisely as obtained in Example 15.8.

15.5.3 β and Power for Lower-Tailed and Two-Sided Tests

For the sake of clarity, the preceding discussion was specifically restricted to the upper-tailed test. Now that we have presented and illustrated the es-

sential concepts, it is relatively straightforward to extend them to other types of tests without having to repeat the details.

First, because the sampling distribution for the test statistic employed for these hypothesis tests is symmetric, it is easy to see that with the lower-tailed alternative

$$H_a : \mu = \mu_a < \mu_0 \tag{15.99}$$

this time, with

$$\delta^* = \mu_0 - \mu_a \tag{15.100}$$

the β risk is obtained as:

$$\beta = P\left(z > z_\alpha + \frac{\delta^*\sqrt{n}}{\sigma}\right) \tag{15.101}$$

the equivalent of Eq (15.82), from where the power is obtained as $(1-\beta)$. Again, because of symmetry, it is easy to see that the expression for determining sample size is precisely the same as derived earlier for the upper tailed test; i.e.,

$$n = \left[\frac{(z_\alpha + z_\beta)\sigma}{\delta^*}\right]^2$$

All other results therefore follow.

For the two-tailed test, things are somewhat different, of course, but the same principles apply. The β risk is determined from:

$$\beta = P\left(z < z_{\alpha/2} - \frac{\delta^*\sqrt{n}}{\sigma}\right) - P\left(z < -z_{\alpha/2} - \frac{\delta^*\sqrt{n}}{\sigma}\right) \tag{15.102}$$

because of the two-sided rejection region. Unfortunately, as a result of the additional term in this equation, there is no closed-form solution for n that is the equivalent of Eq (15.94). When $P\left(z < -z_{\alpha/2} - \frac{\delta^*\sqrt{n}}{\sigma}\right) \ll \beta$, the approximation,

$$n \approx \left[\frac{(z_{\alpha/2} + z_\beta)\sigma}{\delta^*}\right]^2 \tag{15.103}$$

is usually good enough. Of course, given the test characteristics, computer programs can solve for n precisely in Eq (15.102) without the need to resort to the approximation shown here.

15.5.4 General Power and Sample Size Considerations

For general power and sample size considerations, it is typical to start by specifying α and σ; as a result, in either Eq (15.94) for one-tailed tests, or Eq (15.103) for the two-sided test, this leaves 3 parameters to be determined: δ^*, n, and z_β. By specifying any two, a value for the third unspecified parameter that is consistent with the given information can be computed from these equations.

In MINITAB the sequence required for carrying out this procedure is: `Stat > Power and Sample Size` which produces a drop down menu containing a collection of hypothesis tests (and experimental designs—see later). Upon selecting the hypothesis test of interest, a dialog box opens, with the instruction to "`Specify values for any two of the following,`" with three appropriately labeled spaces for "Sample size(s)," "Difference(s)," and "Power value(s)." The "Options" button is used to specify the alternative hypothesis and the α-risk value. The value of the unspecified third parameter is then computed by MINITAB.

The following example illustrates this procedure.

Example 15.10: POWER AND SAMPLE SIZE DETERMINATION USING MINITAB

Use MINITAB to compute power and sample size for an upper-tailed, one sample z-test, with $\sigma = 4$, designed to detect a difference of 2, at the significance level of $\alpha = 0.05$: (1) if $n = 25$, determine the resulting power; (2) when the power is desired to be 0.9, determine the required sample size. (3) With a sample size of $n = 35$, determine the minimum difference that can be detected with a power of 0.9.

Solution:

(1) Upon entering the given parameters into the appropriate boxes in the MINITAB dialog box, and upon choosing the appropriate alternative hypothesis, the MINITAB result is shown below:

Power and Sample Size

```
1-Sample Z Test
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 4
```

Difference	Size	Sample Power
2	25	0.803765

This computed power value is what we had obtained earlier.

(2) When the power is specified and the sample size removed, the MINITAB result is:

Power and Sample Size

```
1-Sample Z Test
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 4
```

Difference	Sample Size	Target Power	Actual Power
2	35	0.9	0.905440

This is exactly the same sample size value and the same actual power value we had obtained earlier.

(3) With n specified as 35 and the difference unspecified, the MINITAB result is:

Power and Sample Size

```
1-Sample Z Test
Testing mean = null (versus > null)
Calculating power for mean = null + difference
Alpha = 0.05 Assumed standard deviation = 4
```

Difference	Sample Size	Target Power	Difference
2	35	0.9	1.97861

The implication is that any difference greater than 1.98 can be detected at the desired power. A difference of 2.0 is therefore detectable at a power that is at least 0.9.

These results are all consistent with what we had obtained earlier.

15.6 Concerning Variances of Normal Populations

The discussions up until now have focused exclusively on hypothesis tests concerning the means of normal populations. But if we recall, for example, the earlier statements made regarding, say, the yield of process A, that $Y_A \sim N(75.5, 1.5^2)$, we see that in this statement is a companion assertion about the associated variance. To confirm or refute this statement *completely* requires testing the validity of the assertion about the variance also.

There are two classes of tests concerning variances of normal population: the first concerns testing the variance obtained from a sample against a postulated population variance (as is the case here with Y_A); the second concerns testing two (independent) normal populations for equality of their variances. We shall now deal with each case.

15.6.1 Single Variance

When the variance of a sample is to be tested against a postulated value, σ_0^2, the null hypothesis is:

$$H_0 : \sigma^2 = \sigma_0^2 \tag{15.104}$$

Under the assumption that the sample in question came from a normal population, then the test statistic:

$$C^2 = \frac{(n-1)S^2}{\sigma_0^2} \tag{15.105}$$

TABLE 15.9: Summary of H_0 rejection conditions for the χ^2-test

Testing Against	For General α Reject H_0 if:
$H_a : \sigma^2 < \sigma_0^2$	$c^2 < \chi^2_{1-\alpha}(n-1)$
$H_a : \sigma^2 > \sigma_0^2$	$c^2 > \chi^2_{\alpha}(n-1)$
$H_a : \sigma^2 \neq \sigma_0^2$	$c^2 < \chi^2_{1-\alpha/2}(n-1)$ or $c^2 > \chi^2_{\alpha/2}(n-1)$

has a $\chi^2(n-1)$ distribution, if H_0 is true. As a result, this test is known as a "Chi-squared" test; and the rejection criteria for the usual triplet of alternatives is shown in Table 15.9. The reader should note the lack of symmetry in the boundaries of these rejection regions when compared with the symmetric boundaries for the corresponding z- and t-tests. This, of course, is a consequence of the asymmetry of the $\chi^2(n-1)$ distribution. For example, for one-sided tests based on 10 samples from a normal distribution, the null hypothesis distributions for C^2 is shown in Fig 15.13.

The next example is used to illustrate a two-sided test.

Example 15.11: VARIANCE OF "PROCESS A" YIELD
Formulate and test an appropriate hypothesis, at the significance level of 0.05, regarding the variance of the yield obtainable from process A implied by the assertion that the sample data presented in Chapter 1 for Y_A is from a normal population with the distribution $N(75.5, 1.5^2)$.

Solution:
The hypothesis to be tested is that $\sigma_A^2 = 1.5^2$ against the alternative that it is not; i.e.,

$$\begin{aligned} H_0 : &\quad \sigma_A^2 = 1.5^2 \\ H_a : &\quad \sigma_A^2 \neq 1.5^2 \end{aligned} \tag{15.106}$$

The sample variance computed from the supplied data is $s_A^2 = 2.05$, so that the specific value for the χ^2 test statistic is:

$$c^2 = \frac{49 \times 2.05}{2.25} = 44.63 \tag{15.107}$$

The rejection region for this two-sided test, with $\alpha = 0.05$, is shown in Fig 15.14, for a $\chi^2(49)$ distribution. The boundaries of the rejection region are obtained from the usual cumulative probabilities; the left boundary is obtained by finding $\chi^2_{1-\alpha/2}$ such that

$$\begin{aligned} P(c^2 > \chi^2_{1-\alpha/2}(49)) &= 0.975 \\ \text{or } P(c^2 < \chi^2_{1-\alpha/2}(49)) &= 0.025 \\ \text{i.e., } \chi^2_{1-\alpha/2} &= 31.6 \end{aligned} \tag{15.108}$$

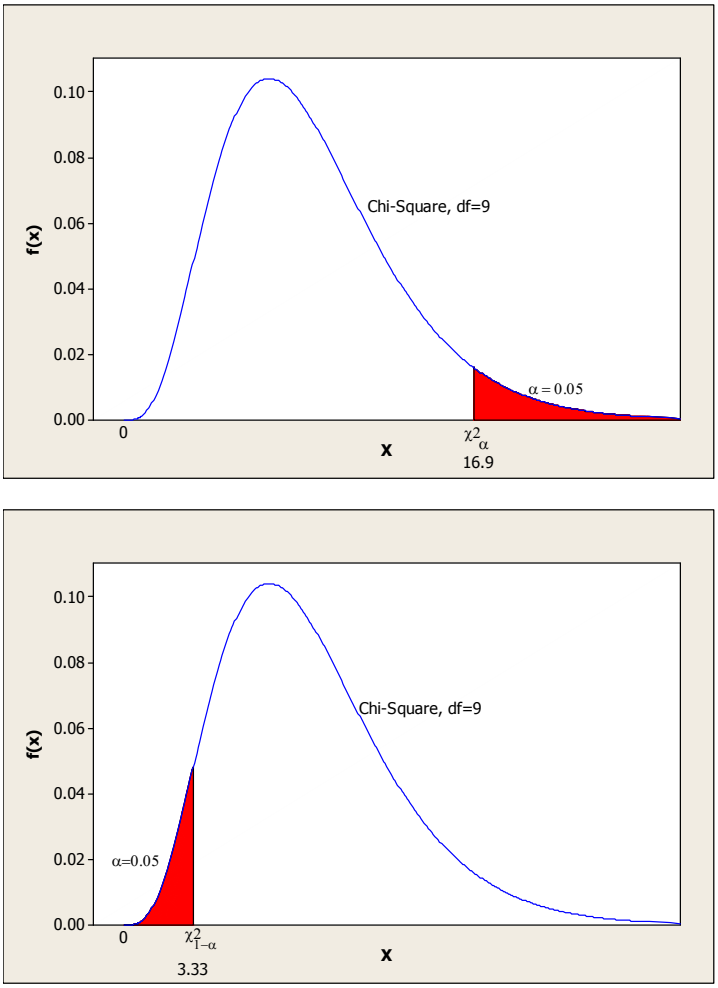

FIGURE 15.13: Rejection regions for one-sided tests of a single variance of a normal population, at a significance level of $\alpha = 0.05$, based on $n = 10$ samples. The distribution is $\chi^2(9)$; Top: for $H_a : \sigma^2 > \sigma_0^2$, indicating rejection of H_0 if $c^2 > \chi^2_\alpha(9) = 16.9$; Bottom: for $H_a : \sigma^2 < \sigma_0^2$, indicating rejection of H_0 if $c^2 < \chi^2_{1-\alpha}(9) = 3.33$.

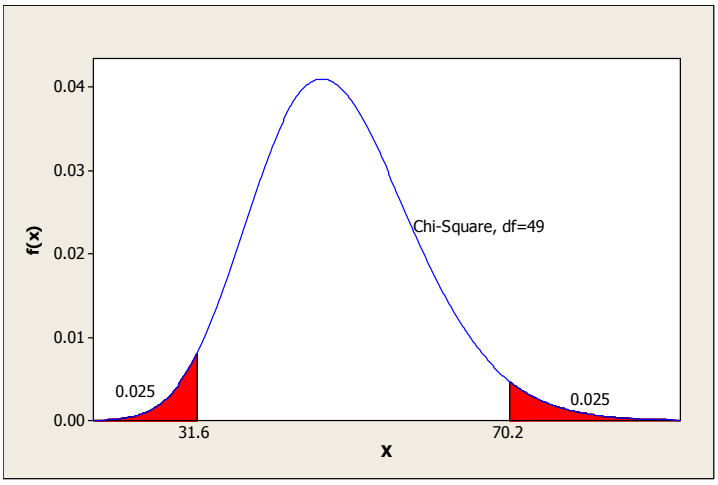

FIGURE 15.14: Rejection regions for the two-sided tests concerning the variance of the process A yield data $H_0 : \sigma_A^2 = 1.5^2$, based on $n = 50$ samples, at a significance level of $\alpha = 0.05$. The distribution is $\chi^2(49)$, with the rejection region shaded; because the test statistic, $c^2 = 44.63$, falls outside of the rejection region, we do not reject H_0.

and the right boundary from:

$$\begin{aligned} P(c^2 > \chi^2_{\alpha/2}(49)) &= 0.025 \\ \text{or } P(c^2 < \chi^2_{\alpha/2}(49)) &= 0.975 \\ \text{i.e., } \chi^2_{\alpha/2} &= 70.2 \end{aligned} \tag{15.109}$$

Since the value for c^2 above does not fall into this rejection region, we do not reject the null hypothesis.

As before, MINITAB could be used directly to carry out this test. The self-explanatory procedure follows along the same lines as those discussed extensively above.

The conclusion: at the 5% significance level, we cannot reject the null hypothesis concerning σ_A^2.

15.6.2 Two Variances

When two variances from mutually independent normal populations are to be compared, the null hypothesis is:

$$H_0 : \sigma_1^2 = \sigma_2^2 \tag{15.110}$$

If the samples (sizes n_1 and n_2 respectively) come from independent normal distributions, then the test statistic:

$$F = \frac{S_1^2}{S_2^2} \tag{15.111}$$

TABLE 15.10: Summary of H_0 rejection conditions for the F-test

Testing Against	For General α Reject H_0 if:
$H_a : \sigma_1^2 < \sigma_2^2$	$f < F_{1-\alpha}(\nu_1, \nu_2)$
$H_a : \sigma_1^2 > \sigma_2^2$	$f > F_{\alpha}(\nu_1, \nu_2)$
$H_a : \sigma_1^2 \neq \sigma_2^2$	$f < F_{1-\alpha/2}(\nu_1, \nu_2)$ or $f > F_{\alpha/2}(\nu_1, \nu_2)$

has an $F(\nu_1, \nu_2)$ distribution, where $\nu_1 = (n_1 - 1)$ and $\nu_2 = (n_2 - 1)$, if H_0 is true. Such tests are therefore known as "F-tests." As with other tests, the rejection regions are determined from the F-distribution with appropriate degrees-of-freedom pairs on the basis of the desired significance level, α. These are shown in Table 15.10.

It is often helpful in carrying out F-tests to recall the following property of the F-distribution:

$$F_{1-\alpha}(\nu_1, \nu_2) = \frac{1}{F_{\alpha}(\nu_2, \nu_1)} \tag{15.112}$$

an easy enough relationship to prove directly from the definition of the F-statistic in Eq (15.111). This relationship makes it possible to reduce the number of entries in old-fashioned F-tables. As we have repeatedly advocated in this chapter, however, it is most advisable to use computer programs for carrying out such tests.

Example 15.12: COMPARING VARIANCES OF YIELDS FROM PROCESSES A AND B

From the data supplied in Chapter 1 on the yields obtained from the two chemical processes A and B, test a hypothesis on the potential equality of these variances, at the 5% significance level.

Solution:

The hypothesis to be tested is that $\sigma_A^2 = \sigma_B^2$ against the alternative that it is not; i.e.,

$$\begin{aligned} H_0 &: \sigma_A^2 = \sigma_B^2 \\ H_a &: \sigma_A^2 \neq \sigma_B^2 \end{aligned} \tag{15.113}$$

From the supplied data, we obtain $s_A^2 = 2.05$, and $s_B^2 = 7.62$, so that the specific value for the F-test statistic is obtained as:

$$f = \frac{2.05}{7.62} = 0.27 \tag{15.114}$$

The rejection region for this two-sided F-test, with $\alpha = 0.05$, is shown in

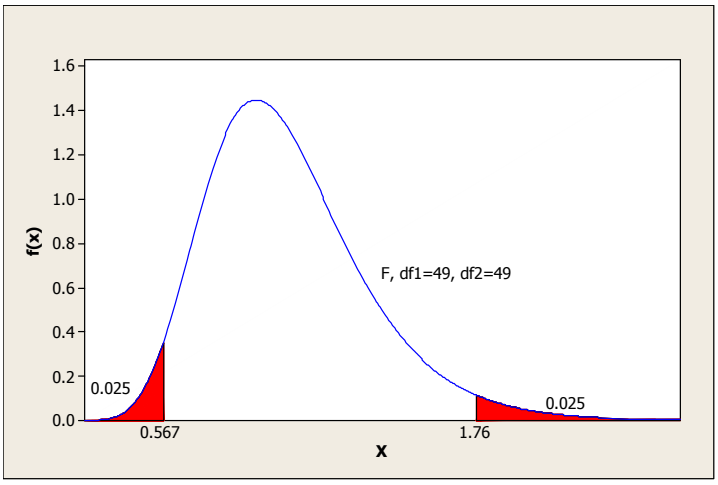

FIGURE 15.15: Rejection regions for the two-sided tests of the equality of the variances of the process A and process B yield data, i.e., $H_0 : \sigma_A^2 = \sigma_B^2$, at a significance level of $\alpha = 0.05$, based on $n = 50$ samples each. The distribution is $F(49, 49)$, with the rejection region shaded; since the test statistic, $f = 0.27$, falls within the rejection region to the left, we reject H_0 in favor of H_a.

Fig 15.15, for an $F(49, 49)$ distribution, with boundaries at $f = 0.567$ to the left and 1.76 to the right, obtained as usual from cumulative probabilities. (Note that the value of f at one boundary is the reciprocal of the value at the other boundary.) Since the specific test value, 0.27, falls in the left side of the rejection region, we must therefore reject the null hypothesis in favor of the alternative that these two variances are unequal.

The self-explanatory procedure for carrying out the test in MINITAB generates results that include a p-value of 0.000, in agreement with the conclusion above to reject the null hypothesis at the 5% significance level.

The F-test is particularly useful for ascertaining whether or not the assumption of equality of variances is valid *before* performing a two-sample t-test. If the null hypothesis regarding the equality assumption is rejected, then one must not use the "equal variance" option of the test. If one is unable to reject the null hypothesis, one may proceed to use the "equal variance" option. As discussed in subsequent chapters, the F-test is also at the heart of ANOVA (**AN**alysis **O**f **VA**riance), a methodology that is central to much of statistical design of experiments and the systematic analysis of the resulting data statistical tests involving several means, and even regression analysis.

Finally, we note that the F-test is quite sensitive to the normality assumption: if this assumption is invalid, the test results will be unreliable. Note that the assumption of normality is not about the mean of the data but about

the raw data set itself. One must therefore be careful to ensure that this normality assumption is reasonable before carrying out an F-test. If the data is from non-normal distributions, most computer programs provide alternatives (based on non-parametric methods).

15.7 Concerning Proportions

As noted at the beginning of this chapter, a statistical hypothesis, in the most fundamental sense, is an assertion or statement about one or more populations; and the hypothesis test provides an objective means of ascertaining the truth or falsity of such a statement. So far, our discussions have centered essentially around normal populations because a vast majority of practical problems are of this form, or can be safely approximated as such. However, not all problems of practical importance involve sampling from normal populations; and the next section will broach this topic from a more general perspective. For now, we want to consider first a particular important class of problems involving sampling from a non-Gaussian population: hypotheses concerning proportions.

The general theoretical characteristics of problems of this kind were studied extensively in Chapter 8. Out of a total number of n samples examined for a particular attribute, X is the total number of (discrete) observations sharing the attribute in question; X/n is therefore the observed sample proportion sharing the attribute. Theoretically, the random variable, X, is known to follow the binomial distribution, characterized by the parameter p, the theoretical population proportion sharing the attribute (also known as the "probability of success"). Statements about such proportions are therefore statistical hypotheses concerning samples from binomial populations. Market/opinion surveys (such as the example used to open Chapter 14) where the proportion preferring a certain brand is of interest, and manufacturing processes where the concern is the proportion of defective products, provide the prototypical examples of problems of this nature. Hypotheses about the probability of successful embryo implantation in in-vitro fertilization (discussed in Chapter 7), or any other such binomial process probability, also fall into this category.

We deal first with hypotheses concerning single population proportions, and then hypotheses concerning two proportions. The underlying principles remain the same as with other tests: find the appropriate test statistic and its sampling distribution, and, given a specific significance level, use these to make probabilistic statements that will allow the determination of the appropriate rejection region.

15.7.1 Single Population Proportion

The problem of interest involves testing a hypothesis concerning a single binomial population proportion, p, given a sample of n items from which one observes X "successes" (the same as the detection of the attribute in question); the null hypothesis is:

$$H_0 : p = p_0 \tag{15.115}$$

with p_0 as the specific value postulated for the population proportion. The usual three possible alternative hypotheses are:

$$H_a : p < p_0 \tag{15.116}$$

$$H_a : p > p_0 \tag{15.117}$$

$$H_a : p \neq p_0 \tag{15.118}$$

To determine an appropriate test statistic and its sampling distribution, we need to recall several characteristics of the binomial random variable from Chapter 8. First, the estimator, Π, defined as:

$$\Pi = \frac{X}{n} \tag{15.119}$$

the mean number of successes, is unbiased for the binomial population parameter; the mean of the sampling distribution for Π is therefore p. Next, the variance of Π is σ_X^2/n^2, where

$$\sigma_X^2 = npq = np(1-p) \tag{15.120}$$

is the variance of the binomial random variable, X. Hence,

$$\sigma_\Pi^2 = \frac{p(1-p)}{n} \tag{15.121}$$

Large Sample Approximations

From the Central Limit Theorem we know that, in the limit as $n \to \infty$, the sampling distribution of the mean of any population (including the binomial) tends to the normal distribution. The implication is that the statistic, Z, defined as:

$$Z = \frac{\frac{X}{n} - p}{\sqrt{p(1-p)/n}} \tag{15.122}$$

has an approximate standard normal, $N(0,1)$, distribution for large n.

The test statistic for carrying out the hypothesis test in Eq (15.115) versus any of the three alternatives is therefore:

$$Z = \frac{\Pi - p_0}{\sqrt{p_0(1-p_0)/n}} \tag{15.123}$$

TABLE 15.11: Summary of H_0 rejection conditions for the single-proportion z-test

Testing Against	For General α Reject H_0 if:	For $\alpha = 0.05$ Reject H_0 if:
$H_a : p < p_0$	$z < -z_\alpha$	$z < -1.65$
$H_a : p > p_0$	$z > z_\alpha$	$z < 1.65$
$H_a : p \neq p_0$	$z < -z_{\alpha/2}$ or $z > z_{\alpha/2}$	$z < -1.96$ or $z > 1.96$

a test statistic with precisely the same properties as those used for the standard z-test. The rejection conditions are identical to those shown in Table 15.2, which, when modified appropriately for the one-proportion test, is as shown in Table 15.11.

Since this test is predicated upon the sample being "sufficiently large," it is important to ensure that this is indeed the case. A generally agreed upon objective criterion for ascertaining the validity of this approximation is that the interval

$$I_0 = p_0 \pm 3\sqrt{[p_0(1-p_0)]/n} \tag{15.124}$$

does not include 0 or 1. The next example illustrates these concepts.

Example 15.13: EXAM TYPE PREFERENCE OF UNDERGRADUATE CHEMICAL ENGINEERING STUDENTS

In the opening sections of Chapter 14, we reported the result of an opinion poll of 100 undergraduate chemical engineering students in the United States: 75 of the students prefer "closed-book" exams to "opened-book" ones. At the 5% significance level, test the hypothesis that the true proportion preferring "closed-book" exams is in fact 0.8, against the alternative that it is not.

Solution:

If the sample size is confirmed to be large enough, then this is a single proportion test which employs the z-statistic. The interval $p_0 \pm 3\sqrt{[p_0(1-p_0)]/n}$ in this case is 0.8 ± 0.12, or (0.68, 0.92), which does not include 0 or 1; the sample size is therefore considered to be sufficiently large.

The hypothesis to be tested is therefore the two-sided

$$\begin{aligned} H_0 : p &= 0.8 \\ H_a : p &\neq 0.8; \end{aligned} \tag{15.125}$$

the z-statistic in this case is:

$$z = \frac{0.75 - 0.8}{\sqrt{(0.8 \times 0.2)/100}} = -1.25 \tag{15.126}$$

Since this value does not lie in the two-sided rejection region for $\alpha = 0.05$, we do not reject the null hypothesis.

MINITAB could be used to tackle this example problem directly. The self-explanatory sequence (when one chooses the `"use test and interval based on normal distribution"` option) produces the following result:

Test and CI for One Proportion

`Test of p = 0.8 vs p not = 0.8`

Sample	X	N	Sample p	95% CI	Z-Value	P-Value
1	75	100	0.750000	(0.665131, 0.834869)	-1.25	0.211

`Using the normal approximation.`

As with similar tests discussed earlier, we see here that the 95% confidence interval for the parameter, p, contains the postulated $p_0 = 0.8$; the associated p-value for the test (an unfortunate and unavoidable notational clumsiness that we trust will not confuse the reader unduly[1]) is 0.211, so that we do not reject H_0 at the 5% significance level.

Exact Tests

Even though it is customary to invoke the normal approximation in dealing with tests for single proportions, this is in fact not necessary. The reason is quite simple: if $X \sim Bi(n, p)$, then $\Pi = X/n$ has a $Bi(n, p/n)$ distribution. This fact can be used to compute the probability that $\Pi = p_0$, or any other value—providing the means for determining the boundaries of the various rejection regions (given desired tail area probabilities), just as with the standard normal distribution, or any other standardized test distribution. Computer programs such as MINITAB provide options for obtaining exact p-values for the single proportion test that are based on exact binomial distributions.

When MINITAB is used to carry out the test in Example 15.13 above, this time without invoking the normal approximation option, the result is as follows:

Test and CI for One Proportion

`Test of p = 0.8 vs p not = 0.8`

Sample	X	N	Sample p	95% CI	Exact P-Value
1	75	100	0.750000	(0.653448, 0.831220)	0.260

The 95% confidence interval, which is now based on a binomial distribution, not a normal approximation, is now slightly different; the p-value is also now slightly different, but the conclusion remains the same.

[1]The latter p of the "p-value" should not be confused with the binomial "probability of success" parameter.

15.7.2 Two Population Proportions

In comparing two population proportions, p_1 and p_2, as with the 2-sample tests of means from normal populations, the null hypothesis is:

$$H_0 : \Pi_1 - \Pi_2 = \delta_0 \tag{15.127}$$

where $\Pi_1 = X_1/n_1$ and $\Pi_2 = X_2/n_2$ are, respectively, the random proportions of "successes" obtained from population 1 and population 2, based on samples of respective sizes n_1 and n_2. For example, Π_1 could be the fraction of defective chips in a sample of n_1 chips manufactured at one facility whose true proportion of defectives is p_1, while Π_2 is the defective fraction contained in a sample from a different facility. The difference between the two population proportions is postulated as some value δ_0 that need not be zero.

As usual, the hypothesis is to be tested against the possible alternatives:

$$\text{Lower-tailed} \quad H_a : \Pi_1 - \Pi_2 < \delta_0 \tag{15.128}$$

$$\text{Upper-tailed} \quad H_a : \Pi_1 - \Pi_2 > \delta_0 \tag{15.129}$$

$$\text{Two-tailed} \quad H_a : \Pi_1 - \Pi_2 \neq \delta_0 \tag{15.130}$$

As before, $\delta_0 = 0$ constitutes a test of equality of the two proportions.

To obtain an appropriate test statistic and its sampling distribution, we begin by defining:

$$D_\Pi = \Pi_1 - \Pi_2 \tag{15.131}$$

We know in general that

$$E(D_\Pi) = \mu_{D_\Pi} = p_1 - p_2 \tag{15.132}$$

$$\sigma_{D_\Pi} = \sqrt{\left(\frac{p_1 q_1}{n_1}\right) + \left(\frac{p_2 q_2}{n_2}\right)} \tag{15.133}$$

But now, if the sample sizes n_1 and n_2 are large, then it can be shown that

$$D_\Pi \sim N(\mu_{D_\Pi}, \sigma^2_{D_\Pi}) \tag{15.134}$$

again allowing us to invoke the normal approximation (for large sample sizes). This immediately implies that the following is an appropriate test statistic to use for this two-proportion test:

$$Z = \frac{(\Pi_1 - \Pi_2) - \delta_0}{\sqrt{\left(\frac{p_1 q_1}{n_1}\right) + \left(\frac{p_2 q_2}{n_2}\right)}} \sim N(0, 1) \tag{15.135}$$

Since population values, p_1 and p_2, are seldom available in practice, it is customary to substitute sample estimates,

$$\hat{p}_1 = \frac{x_1}{n_1}; \text{ and } \hat{p}_2 = \frac{x_2}{n_2} \tag{15.136}$$

Finally, since this test statistic possesses a standard normal distribution, the rejection regions are precisely the same as those in Table 15.4.

In the special case when $\delta_0 = 0$, which is equivalent to a test of equality of the proportions, the most important consequence is that if the null hypothesis is true, then $p_1 = p_2 = p$, which is then estimated by the "pooled" proportion:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \tag{15.137}$$

As a result, the standard deviation of the difference in proportions, σ_{D_Π}, becomes:

$$\sigma_{D_\Pi} = \sqrt{\left(\frac{p_1 q_1}{n_1}\right) + \left(\frac{p_2 q_2}{n_2}\right)} \approx \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \tag{15.138}$$

so that the test statistic in Eq (15.135) is modified to

$$Z = \frac{(\Pi_1 - \Pi_2)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1) \tag{15.139}$$

The rejection regions are the same as in the general case.

Example 15.14: REGIONAL PREFERENCE FOR PEPSI
To confirm persistent rumors that the preference for PEPSI on engineering college campuses is higher in the Northeast of the United States than on comparable campuses in the Southeast, a survey was carried out on 125 engineering students chosen at random on the MIT campus in Cambridge, MA, and the same number of engineering students selected at random at Georgia Tech in Atlanta, GA. Each student was asked to indicate a preference for PEPSI versus other soft drinks, with the following results: 44 of the 125 at MIT indicate preference for PEPSI versus 26 at GA Tech. At the 5% level, determine whether the Northeast proportion, $\hat{p}_1 = 0.352$, is essentially the same as the Southeast proportion, $\hat{p}_2 = 0.208$, against the alternative that they are different.

Solution:
The hypotheses to be tested are:

$$\begin{aligned} H_0 : \Pi_1 - \Pi_2 &= 0 \\ H_a : \Pi_1 - \Pi_2 &\neq 0 \end{aligned} \tag{15.140}$$

and from the given data, the test statistic computed from Eq (15.139) is $z = 2.54$. Since this number is greater than 1.96, and therefore lies in the rejection region of the two-sided test, we reject the null hypothesis in favor of the alternative. Using MINITAB to carry out this test, selecting the `"use pooled estimate of p for test,"` produces the following result:

Test and CI for Two Proportions

Sample	X	N	Sample p
1	44	125	0.352000
2	26	125	0.208000

```
Difference = p (1) - p (2)
Estimate for difference: 0.144
95% CI for difference: (0.0341256, 0.253874)
Test for difference = 0 (vs not = 0): Z = 2.54 P-Value = 0.011
```

Note that the 95% confidence interval around the estimated difference of 0.144 does not include zero; the p-value associated with the test is 0.011 which is less than 0.05; hence, we reject the null hypothesis at the 5% significance level.

As an exercise, the reader should extend this example by testing $\delta_0 = 0.02$ against the alternative that the difference is greater than 0.02.

15.8 Concerning Non-Gaussian Populations

The discussion in the previous section has opened up the issue of testing hypotheses about non-Gaussian populations, and has provided a strategy for handling such problems in general. The central issue is finding an appropriate test statistic and its sampling distribution, as was done for the binomial distribution. This cause is advanced greatly by the relationship between interval estimates and hypothesis tests (discussed earlier in Section 15.3.3) and by the discussion at the end of Chapter 14 on interval estimates for non-Gaussian distributions.

15.8.1 Large Sample Test for Means

First, if the statistical hypothesis is about the mean of a non-Gaussian population, so long as the sample size, n, used to compute the sample average, $\bar{X}$, is reasonably large (e.g., $n > 30$ or so), then, regardless of the underlying distribution, we known that the statistic $Z = (\bar{X} - \mu)/\sigma_{\bar{X}}$ possesses an approximate standard normal distribution—an approximation that improves as $n \to \infty$. Thus, hypotheses about the means of non-Gaussian populations that are based on large sample sizes are essentially the same as z-tests.

Example 15.15: HYPOTHESIS TEST ON MEAN OF INCLUSIONS DATA

If the data in Table 1.2 is considered a random sample of 60 observations of the number of *inclusions* found on glass sheets produced in the manufacturing process discussed in Chapter 1, test at the 5% significance level, the hypothesis that this data came from a Poisson population with

mean $\lambda = 1$, against the alternative that λ is not 1.

Solution:
The hypotheses to be tested are:

$$\begin{aligned} H_0 : \lambda &= 1 \\ H_a : \lambda &\neq 1 \end{aligned} \tag{15.141}$$

While the data is from a Poisson population, the sample size is large; hence, the test statistic:

$$Z = \frac{\bar{X} - \lambda_0}{\sigma/\sqrt{60}} \tag{15.142}$$

where σ is the standard deviation of the raw data (so that $\sigma/\sqrt{60}$ is the standard deviation of the sample average), essentially has a standard normal distribution.

From the supplied data, we obtain the sample average $\hat{\lambda} = \bar{x} = 1.02$, with the sample standard deviation, $s = 1.1$, which, because of the large sample, will be considered to be a reasonable approximation of σ. The test statistic is therefore obtained as $z = 0.141$. Since this value is not in the two-sided rejection region $|z| > 1.96$ for $\alpha = 0.05$, we do not reject the null hypothesis. We therefore conclude that there is no evidence to contradict the statement that $X \sim \mathcal{P}(1)$, i.e., the inclusions data is from a Poisson population with mean number of inclusions = 1.

It is now important to recall the results in Example 14.13 where the 95% confidence interval estimate for the mean of the inclusions data was obtained as:

$$\lambda = 1.02 \pm 1.96(1.1/\sqrt{60}) = 1.02 \pm 0.28 \tag{15.143}$$

i.e., $0.74 < \lambda < 1.30$. Note that this interval contains the hypothesized value $\lambda = 1.0$, indicating that we cannot reject the null hypothesis.

We can now use this result to answer the following question raised in Chapter 1 as a result of the potentially "disturbing" data obtained from the quality control lab apparently indicating too many glass sheets with too many inclusions: *if the process was designed to produce glass sheets with a mean number of inclusions $\lambda^* = 1$ per m^2, is there evidence in this sample data that the process has changed, that the number of observed "inclusions" is significantly different from what one can reasonably expect from the process when operating as designed?*

From the results of this example, the answer is, No: at the 5% significance level, there no evidence that the process has deviated from its design target.

15.8.2 Small Sample Tests

When the sample size on which the sample average is based is small, or when we are dealing with aspects of the population other than the mean (say the variance), we are left with only one option: go back to "first principles,"

derive the sampling distribution for the appropriate statistic and use it to carry out the required test. One can use the sampling distribution to determine $\alpha\times100\%$ rejection regions, or the complementary region, the $(1-\alpha)\times100\%$ confidence interval estimates for the appropriate parameter.

For tests involving single parameters, it makes no difference which of these two approaches we choose; for tests involving two parameters, however, it is more straightforward to compute confidence intervals for the parameters in question and then use these for the hypothesis test. The reason is that for tests involving two parameters, confidence intervals can be computed directly from the individual sampling distributions; on the other hand, computing rejection regions for the difference between these two parameters technically requires an additional step of deriving yet another sampling distribution for the *difference*. And the sampling distribution of the difference between two random variables may not always be easy to derive. Having discussed earlier in this chapter the equivalence between confidence intervals and hypothesis tests, we now note that for non-Gaussian problems, one might as well just base the hypotheses tests on $(1-\alpha)\times100\%$ confidence intervals and avoid the additional hassle of having to derive distributions for differences. Let us illustrate this concept with a problem involving the exponential random variable discussed in Chapter 14.

In Example 14.3, we presented a problem involving an exponential random variable, the waiting time (in days) until the occurrence of a recordable safety incident in a certain company's manufacturing site. The safety performance data for the first and second years were presented, from which point estimates of the unknown population parameter, β, were determined from the sample averages, $\bar{x}_1 = 30.1$ days, for Year 1 and $\bar{x}_2 = 32.9$ days for Year 2; the sample size in each case is $n = 10$, which is considered small.

To test the two-sided hypothesis that these two safety performance parameters (Year 1 versus Year 2) are the same, versus the alternative that they are significantly different (at the 5% significance level), we proceed as follows: we first obtain the sampling distribution for $\bar{X}_1$ and $\bar{X}_2$ given that $X \sim E(\beta)$; we then use these to obtain 95% confidence interval estimates for the population means β_i for Year i; if these intervals overlap, then at the 5% significance level, we cannot reject the null hypothesis that these means are the same; if the intervals do not overlap, we reject the null hypothesis.

Much of this, of course, was already accomplished in Example 14.14: we showed that $\bar{X}_i$ has a gamma distribution, more specifically, $\bar{X}_i/\beta_i \sim \gamma(n, 1/n)$, from where we obtain 95% confidence intervals estimates for β_i from sample data. In particular, for $n = 10$, we obtained from the Gamma(10,0.1) distribution that:

$$P\left(0.48 < \frac{\bar{X}}{\beta} < 1.71\right) = 0.95 \tag{15.144}$$

which, upon introducing $\bar{x}_1 = 30.1$, and $\bar{x}_2 = 32.9$, produces, upon careful rearrangement, the 95% confidence interval estimates for the Year 1 and Year

2 parameters respectively as:

$$17.6 < \beta_1 < 62.71 \tag{15.145}$$

$$19.24 < \beta_2 < 68.54 \tag{15.146}$$

These intervals may now be used to answer a wide array of questions regarding hypotheses concerning two parameters, even questions concerning a single parameter. For instance,

1. For the two-parameter null hypothesis, $H_0 : \beta_1 = \beta_2$, versus $H_a : \beta_1 \neq \beta_2$, because the 95% confidence intervals overlap considerably, we find no evidence to reject H_0 at the 5% significance level.

2. In addition, the single parameter null hypothesis, $H_0 : \beta_1 = 40$, versus $H_a : \beta_1 \neq 40$, cannot be rejected at the 5% significance level because the postulated value is contained in the 95% confidence interval for β_1; on the contrary, the null hypothesis $H_0 : \beta_1 = 15$, versus $H_a : \beta_1 \neq 15$ will be rejected at the 5% significance level because the hypothesized value falls outside of the 95% confidence interval (i.e., falls in the rejection region).

3. Similarly, the null hypothesis $H_0 : \beta_2 = 40$, versus $H_a : \beta_2 \neq 40$, cannot be rejected at the 5% significance level because the postulated value is contained in the 95% confidence interval for β_2; on the other hand, the null hypothesis $H_0 : \beta_2 = 17$, versus $H_a : \beta_2 \neq 17$ will be rejected at the 5% significance level because the hypothesized value falls outside of the 95% confidence interval (i.e., it falls in the rejection region).

The principles illustrated here can be applied to any non-Gaussian population provided the sampling distribution of the statistic in question can be determined.

Another technique for dealing with populations characterized by any general pdf (Gaussian or not), and based on the maximum likelihood principle discussed in Chapter 14 for estimating unknown population parameters, is discussed next in its own separate section.

15.9 Likelihood Ratio Tests

In its broadest sense, a likelihood ratio (LR) test is a technique for assessing how well a simpler, "restricted" version of a probability model compares to its more complex, unrestricted version in explaining observed data. Within the context of this current chapter, however, the discussion here will be limited to testing hypotheses about the parameters, $\boldsymbol{\theta}$, of a population characterized by the pdf $f(x, \boldsymbol{\theta})$. Even though based on fundamentally different premises, some

of the most popular tests considered above (the z- and t-tests, for example) are equivalent to LR tests under recognizable conditions.

15.9.1 General Principles

Let X be a random variable with the pdf $f(x, \boldsymbol{\theta})$, where the population parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$; i.e., $\boldsymbol{\Theta}$ represents the set of possible values that the parameter vector can take. Given a random sample, $X_1, X_2, \ldots, X_n$, estimation theory, as discussed in Chapter 14, is concerned with using such sample information to determine reasonable estimates for $\boldsymbol{\theta}$. In particular, we recall that the maximum likelihood (ML) principle requires choosing the estimate, $\hat{\boldsymbol{\theta}}_{ML}$, as the value of $\boldsymbol{\theta}$ that maximizes the likelihood function:

$$L(\boldsymbol{\theta}) = f_1(x_1, \boldsymbol{\theta}) f_2(x_2, \boldsymbol{\theta}) \cdots f_n(x_n, \boldsymbol{\theta}) \tag{15.147}$$

the joint pdf of the random sample, treated as a function of the unknown population parameter.

The same random sample and the same ML principle can be used to test the null hypotheses

$$H_0 : \boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \tag{15.148}$$

stated in a more general fashion in which $\boldsymbol{\theta}$ is restricted to a certain range of values, $\boldsymbol{\Theta}_0$ (a subset of $\boldsymbol{\Theta}$), over which H_0 is hypothesized to be valid. For example, to test a hypothesis about the mean of X by postulating that $X \sim N(75, 1.5^2)$, in this current context, $\boldsymbol{\Theta}$, the full set of possible parameter vales, is defined as follows:

$$\boldsymbol{\Theta} = \{(\theta_1, \theta_2) : -\infty < \theta_1 = \mu < \infty; \theta_2 = \sigma^2 = 1.5^2\} \tag{15.149}$$

since the variance is given and the only unknown parameter is the mean; $\boldsymbol{\Theta}_0$, the restricted parameter set range over which H_0 is conjectured to be valid, is defined as:

$$\boldsymbol{\Theta}_0 = \{(\theta_1, \theta_2) : \theta_1 = \mu_0 = 75; \theta_2 = \sigma^2 = 1.5^2\} \tag{15.150}$$

The null hypothesis in Eq (15.148) is to be tested against the alternative:

$$H_a : \boldsymbol{\theta} \in \boldsymbol{\Theta}_a \tag{15.151}$$

again stated in a general fashion in which the parameter set, $\boldsymbol{\Theta}_a$, is (a) disjoint from $\boldsymbol{\Theta}_0$, and (b) also complementary to it, in the sense that

$$\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_a \tag{15.152}$$

For example, the two-sided alternative to the hypothesis above regarding $X \sim N(75, 1.5^2)$ translates to:

$$\boldsymbol{\Theta}_a = \{(\theta_1, \theta_2) : \theta_1 = \mu_0 \neq 75; \theta_2 = \sigma^2 = 1.5^2\} \tag{15.153}$$

Note that the union of this set with $\boldsymbol{\Theta}_0$ in Eq (15.150) is the full parameter set range, $\boldsymbol{\Theta}$ in Eq (15.149).

Now, define the largest likelihood under H_0 as

$$L^*(\boldsymbol{\Theta}_0) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0} L(\boldsymbol{\theta}) \tag{15.154}$$

and the unrestricted maximum likelihood value as:

$$L^*(\boldsymbol{\Theta}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_a} L(\boldsymbol{\theta}) \tag{15.155}$$

Then the ratio:

$$\Lambda = \frac{L^*(\boldsymbol{\Theta}_0)}{L^*(\boldsymbol{\Theta})} \tag{15.156}$$

is known as the *likelihood ratio*; it possesses some characteristics that make it attractive for carrying out general hypothesis tests. But first, we note that by definition, $L^*(\boldsymbol{\Theta})$ is the maximum value achieved by the likelihood function when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_{ML}$. Also, Λ is a random variable (it depends on the random sample, $X_1, X_2, \ldots, X_n$); this is why it is sometimes called the *likelihood ratio test statistic*. When specific data values, $x_1, x_2, \ldots, x_n$, are introduced into Eq (15.156), the result is a specific value, λ, for the likelihood ratio such that $0 \leq \lambda \leq 1$, for the following reasons:

1. $\lambda \geq 0$. This is because each likelihood function contributing to the ratio is a pdf (joint pdfs, but pdfs nonetheless), and each legitimate pdf is such that $f(x, \boldsymbol{\theta}) > 0$;

2. $\lambda \leq 1$. This is because $\boldsymbol{\Theta}_0 \subset \boldsymbol{\Theta}$; consequently, since $L^*(\boldsymbol{\Theta})$ is the largest achievable value of the likelihood function in the entire unrestricted set $\boldsymbol{\Theta}$, the largest likelihood value achieved in the subset $\boldsymbol{\Theta}_0$, $L^*(\boldsymbol{\Theta}_0)$, will be less than, or at best equal to, $L^*(\boldsymbol{\Theta})$.

Thus, Λ is a random variable defined on the unit interval (0,1) whose pdf, $f(\lambda|\boldsymbol{\theta}_0)$ (determined by $f(x, \boldsymbol{\theta})$), can be used, in principle, to test H_0 in Eq (15.148) versus H_a in Eq (15.151). It should not come as a surprise that, in general, the form of $f(\lambda|\boldsymbol{\theta}_0)$ can be quite complicated. However there are certain general principles regarding the use of Λ for hypothesis testing:

1. If a specific sample $x_1, x_2, \ldots, x_n$, generates a value of λ close to zero, the implication is that the observation is highly unlikely to have occurred had H_0 been true relative to the alternative;

2. Conversely, if λ is close to 1, then the likelihood of the observed data, $x_1, x_2, \ldots, x_n$, occurring if H_0 is true is just about as high as the unrestricted likelihood that $\boldsymbol{\theta}$ can take any value in the entire unrestricted parameter space $\boldsymbol{\Theta}$;

3. Thus, small values of λ provide evidence *against* the validity of H_0; larger values provide evidence in support.

How "small" λ has to be to trigger rejection of H_0 is formally determined in the usual fashion: using the distribution for Λ, the pdf $f(\lambda|\boldsymbol{\theta}_0)$, obtain a critical value, λ_c, such that $P(\Lambda < \lambda_c) = \alpha$, i.e.,

$$P(\Lambda < \lambda_c) = \int_0^{\lambda_c} f(\lambda|\boldsymbol{\theta}_0) = \alpha \tag{15.157}$$

Any value of λ less than this critical value will trigger rejection of H_0.

Likelihood ratio tests are very general; they can be used even for cases involving structurally different H_0 and H_a probability distributions, or for random variables that are correlated. While the form of the pdf for Λ that is appropriate for each case may be quite complicated, in general, it is always possible to perform the required computations numerically using computer programs. Nevertheless, there are many special cases for which closed-form analytical expressions can be derived directly either for $f(\lambda|\boldsymbol{\theta}_0)$, the pdf of Λ itself, or else for the pdf of a monotonic function of Λ. See Pottmann *et al.*, (2005),[2] for an application of the likelihood ratio test to an industrial sensor data analysis problem.

15.9.2 Special Cases

Normal Population; Known Variance

Consider first the case where a random variable $X \sim N(\mu, \sigma^2)$, has known variance, but an unknown mean; and let $X_1, X_2, \ldots, X_n$ be a random sample from this population. From a specific sample data set, $x_1, x_2, \ldots, x_n$, we wish to test $H_0 : \mu = \mu_0$ against the alternative, $H_a : \mu \neq \mu_0$.

Observe that in this case, with $\boldsymbol{\theta} = (\theta_1, \theta_2) = (\mu, \sigma^2)$, the parameter spaces of interest are:

$$\boldsymbol{\Theta}_0 = \{(\theta_1, \theta_2) : \theta_1 = \mu_0; \theta_2 = \sigma^2\} \tag{15.158}$$

and

$$\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_a = \{(\theta_1, \theta_2) : -\infty < \theta_1 = \mu < \infty; \theta_2 = \sigma^2\} \tag{15.159}$$

Since $f(x, \boldsymbol{\theta})$ is Gaussian, the likelihood function, given the data, is

$$\begin{aligned} L(\mu, \sigma) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{\frac{-(x_i - \mu)^2}{2\sigma^2}\right\} \\ &= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left\{\frac{-\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2}\right\} \end{aligned} \tag{15.160}$$

This function is maximized (when σ^2 is known) by the maximum likelihood

[2] Pottmann, M., B. A. Ogunnaike, and J. S. Schwaber. (2005). "Development and Implementation of a High-Performance Sensor System for an Industrial Polymer Reactor," *Ind. Eng. Chem. Res.*, 44, 2606–2620.

estimator for μ, the sample average, $\bar{X}$; thus, the unrestricted maximum value, $L^*(\boldsymbol{\Theta})$, is obtained by introducing $\bar{X}$ for μ in Eq (15.160); i.e.,

$$L^*(\boldsymbol{\Theta}) = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left\{\frac{-\sum_{i=1}^{n}(x_i - \bar{X})^2}{2\sigma^2}\right\} \tag{15.161}$$

On the other hand, the likelihood function, restricted to $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$ (i.e., $\mu = \mu_0$) is obtained by introducing μ_0 for μ in Eq (15.160). Because, in terms of μ, this function is now a constant, its maximum (in terms of μ) is given by:

$$L^*(\boldsymbol{\Theta}_0) = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left\{\frac{-\sum_{i=1}^{n}(x_i - \mu_0)^2}{2\sigma^2}\right\} \tag{15.162}$$

From here, the likelihood ratio statistic is obtained as:

$$\Lambda = \frac{\exp\left\{\frac{-\sum_{i=1}^{n}(x_i-\mu_0)^2}{2\sigma^2}\right\}}{\exp\left\{\frac{-\sum_{i=1}^{n}(x_i-\bar{X})^2}{2\sigma^2}\right\}} \tag{15.163}$$

Upon rewriting $(x_i - \mu_0)^2$ as $[(x_i - \bar{X}) - (\bar{X} - \mu_0)]^2$ so that:

$$\sum_{i=1}^{n}(x_i - \mu_0)^2 = \sum_{i=1}^{n}(x_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2 \tag{15.164}$$

and upon further simplification, the result is:

$$\Lambda = \exp\left\{\frac{-n(\bar{X} - \mu_0)^2}{2\sigma^2}\right\} \tag{15.165}$$

To proceed from here, we need the pdf for the random variable, Λ; but rather than confront this challenge directly, we observe that:

$$-2\ln\Lambda = \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right)^2 = Z^2 \tag{15.166}$$

where Z, of course, is the familiar z-test statistic

$$Z = \left(\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}\right) \tag{15.167}$$

with a standard normal distribution, $N(0, 1)$. Thus the random variable, $\Psi = -2\ln\Lambda$, therefore has a $\chi^2(1)$ distribution. From here it is now a straightforward exercise to obtain the rejection region in terms of not Λ, but $\Psi = -2\ln\Lambda$ (or Z^2). For a significance level of $\alpha = 0.05$, we obtain from tail area probabilities of the $\chi^2(1)$ distribution that

$$P(Z^2 \geq 3.84) = 0.05 \tag{15.168}$$

so that the null hypothesis is rejected when:

$$\frac{n(\bar{X}-\mu_0)^2}{\sigma^2} > 3.84 \tag{15.169}$$

Upon taking square roots, being careful to retain both positive as well as negative values, we obtain the familiar rejection conditions for the z-test:

$$\begin{aligned} \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} &< -1.96 \text{ or} \\ \frac{\sqrt{n}(\bar{X}-\mu_0)}{\sigma} &> 1.96 \end{aligned} \tag{15.170}$$

The LR test under these conditions is therefore exactly the same as the z-test.

Normal Population; Unknown Variance

When the population variance is unknown for the test discussed above, some things change slightly. First, the parameter spaces become:

$$\boldsymbol{\Theta}_0 = \{(\theta_1, \theta_2) : \theta_1 = \mu_0; \theta_2 = \sigma^2 > 0\} \tag{15.171}$$

along with,

$$\boldsymbol{\Theta} = \boldsymbol{\Theta}_0 \cup \boldsymbol{\Theta}_a = \{(\theta_1, \theta_2) : -\infty < \theta_1 = \mu < \infty; \theta_2 = \sigma^2 > 0\} \tag{15.172}$$

The likelihood function remains the same:

$$L(\mu, \sigma) = \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sigma^n} \exp\left\{\frac{-\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2}\right\}$$

but this time both parameters are unknown, even though the hypothesis test is on μ alone. As a result, the function is maximized by the maximum likelihood estimators for both μ, and σ^2. As obtained in Chapter 14, these are the sample average, $\bar{X}$, and,

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(x_1-\bar{X})^2}{n}$$

respectively.

The unrestricted maximum value, $L^*(\boldsymbol{\Theta})$, in this case is obtained by introducing these ML estimators for the respective unknown parameters in Eq (15.160) and rearranging to obtain:

$$L^*(\boldsymbol{\Theta}) = \left\{\frac{n}{2\pi\sum_{i=1}^{n}(x_i-\bar{X})^2}\right\}^{n/2} e^{-n/2} \tag{15.173}$$

When the parameters are restricted to $\boldsymbol{\theta} \in \boldsymbol{\Theta}_0$, this time, the likelihood function is maximized, after substituting $\mu = \mu_0$, by the MLE for σ^2, so that the largest likelihood value is obtained as:

$$L^*(\boldsymbol{\Theta}_0) = \left\{\frac{n}{2\pi\sum_{i=1}^{n}(x_i-\mu_0)^2}\right\}^{n/2} e^{-n/2} \tag{15.174}$$

Thus, the likelihood ratio statistic becomes:

$$\Lambda = \left\{ \frac{\sum_{i=1}^{n}(x_i - \bar{X})^2}{\sum_{i=1}^{n}(x_i - \mu_0)^2} \right\}^{n/2} \tag{15.175}$$

And upon employing the sum-of-squares identity in Eq (15.164), and simplifying, we obtain:

$$\Lambda = \left\{ \frac{1}{1 + \frac{n(\bar{X}-\mu_0)^2}{\sum_{i=1}^{n}(x_i - \bar{X})^2}} \right\}^{n/2} \tag{15.176}$$

If we now introduce the sample variance $S^2 = \sum_{i=1}^{n}(x_i - \bar{X})^2/(n-1)$, this expression is easily rearranged to obtain:

$$\Lambda = \left\{ \frac{1}{1 + \frac{1}{n-1}\frac{n(\bar{X}-\mu_0)^2}{S^2}} \right\}^{n/2} \tag{15.177}$$

As before, to proceed from here, we need to obtain the pdf for the random variable, Λ. However, once again, we recognize a familiar statistic embedded in Eq (15.177), i.e.,

$$T^2 = \left(\frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right)^2 \tag{15.178}$$

where T has the student's t-distribution with $\nu = n - 1$ degrees of freedom. The implication therefore is that:

$$\Lambda^{2/n} = \frac{1}{1 + T^2/\nu} \tag{15.179}$$

From here we observe that because $\Lambda^{2/n}$ (and hence Λ) is a strictly monotonically decreasing function of T^2 in Eq (15.179), then the rejection region $\lambda < \lambda_c$ for which say $P(\Lambda < \lambda_c) = \alpha$, is exactly equivalent to a rejection region $T^2 > t_c^2$, for which,

$$P(T^2 > t_c^2) = \alpha \tag{15.180}$$

Once more, upon taking square roots, retaining both positive as well as negative values, we obtain the familiar rejection conditions for the t-test:

$$\begin{aligned} \frac{(\bar{X} - \mu_0)}{S/\sqrt{n}} &< -t_{\alpha/2}(\nu) \text{ or} \\ \frac{(\bar{X} - \mu_0)}{S/\sqrt{n}} &> t_{\alpha/2}(\nu) \end{aligned} \tag{15.181}$$

which, of course, is the one-sample, two-sided t-test for a normal population with unknown variance.

Similar results can be obtained for tests concerning the variance of a single

normal population (yielding the χ^2-test) or concerning two variances from independent normal populations, yielding the F-test.

The point, however, is that having shown that the LR tests in these well-known special cases reduce to tests with which we are already familiar, we have the confidence that in the more complicated cases, where the population pdfs are non-Gaussian and closed-form expressions for Λ cannot be obtained as easily, the results (mostly determined numerically) can be trusted.

15.9.3 Asymptotic Distribution for Λ

As noted repeatedly above, it is often impossible to obtain closed-form pdfs for the likelihood ratio test statistic, Λ, or for appropriate functions thereof. Nevertheless, for large samples, there exists an asymptotic distribution:

Asymptotic Distribution Result for LR Test Statistic: The distribution of the random variable $\Psi = -2\ln\Lambda$ tends asymptotically to a $\chi^2(\nu)$ distribution with ν degrees of freedom, with $\nu = \mathcal{N}_p(\boldsymbol{\Theta}) - \mathcal{N}_p(\boldsymbol{\Theta}_0)$ where $\mathcal{N}_p()$ is the number of independent parameters in the parameter space in question, i.e., the number of parameters in $\boldsymbol{\Theta}$ exceeds those in $\boldsymbol{\Theta}_0$ by ν.

Observe, for example, that the distribution of $\Psi = -2\ln\Lambda$ in the first special case (Gaussian distribution with known variance) is exactly $\chi^2(1)$: $\boldsymbol{\Theta}$ contains one unknown parameter, μ, while $\boldsymbol{\Theta}_0$ contains no unknown parameter since $\mu = \mu_0$.

This asymptotic result is exactly equivalent to the large sample approximation to the sampling distribution of means of arbitrary populations. Note that in the second special case (Gaussian distribution with *unknown* variance), $\boldsymbol{\Theta}$ contains two unknown parameter, μ and σ^2, while $\boldsymbol{\Theta}_0$ contains only one unknown parameter, σ^2. The asymptotic distribution of $\Psi = -2\ln\Lambda$ will then also be $\chi^2(1)$, in precisely the same sense in which $t(\nu) \rightarrow N(0,1)$.

15.10 Discussion

This chapter should not end without bringing to the reader's attention some of the criticisms of certain aspects of hypothesis testing. The primary issues have to do not so much with the mathematical foundations of the methodology as with the implementation and interpretation of the results in practice. Of several controversial issues, the following are three we wish to highlight:

1. *Point null hypothesis and statistical-versus-practical significance:* When the null hypothesis about a population parameter is that $\theta = \theta_0$, where θ_0 is a point on the real line, such a literal mathematical statement, can almost always be proven false with computations carried *to a sufficient number of decimal places*. For example, if $\theta_0 = 75.5$, a large enough sample that generates $\bar{x} = 75.52$ (a routine possibility even when the population parameter is indeed 75.5) will lead to the rejection of H_0, to two decimal places. However, in actual practice (engineering or science), is the distinction between two real numbers 75.5 and 75.52 truly of importance? That is, is the statement $75.5 \neq 75.52$, which is true in the strictest, literal mathematical sense, meaningful in practice? Sometimes yes, sometime no; but the point is that such null hypotheses can almost always be falsified, raising the question: what then does rejecting H_0 really mean?

2. *Borderline p-values and variability:* Even when the p-value is used to determine whether or not to reject H_0, it is still customary to relate the computed p-value to some value of α, typically 0.05. But what happens for $p = 0.06$, or $p = 0.04$? Furthermore, an important fact that often goes unnoticed is that were we to repeat the experiment in question, the new data set will almost always lead to results that are "different" from those obtained earlier; and consequently the new p-value will also be different from that obtained earlier. One cannot therefore rule out the possibility of a "borderline" p-value "switching sides" purely as a result of intrinsic variability in the data.

3. *Probabilistic interpretations:* From a more technical perspective, if δ represents the observed discrepancy between the observed postulated population parameter and the value determined from data (a realization of the random variable, Δ), the p-value (or else the actual significance level of the test) is defined as $P(\Delta \geq \delta | H_0)$; i.e., the probability of observing the computed difference or something more extreme if the null hypothesis is true. In fact, the probability we should be interested in is the reverse: $P(H_0 | \Delta \geq \delta)$, i.e., the probability that the null hypothesis is true given the evidence in the data, which truly measures how much the observed data supports the proposed statement of H_0. These two conditional probabilities are generally not the same.

In light of these issues (and others we have not discussed here), how should one approach hypothesis testing in practice? First, statistical significance should not be the only factor in drawing conclusions from experimental results—the nature of the problem at hand should be taken into consideration as well. The yield from process A may in fact not be precisely 75.5% (after all, the probability that a random variable will take on a precise value on the real line is exactly zero), but 75.52% is sufficiently close that the difference is of no practical consequence. Secondly, one should be careful in basing the entire

decision about experimental results on a *single* hypothesis test, especially with p-values at the border of the traditional $\alpha = 0.05$. A single statistical hypothesis test of data obtained in a single study is just that: it can hardly be considered as having definitively "confirmed" something. Thirdly, decisions based on confidence intervals around the estimated population parameters tend to be less confusing and are more likely to provide the desired solution more directly.

Finally, the reader should be aware of the existence of other recently proposed alternatives to conventional hypothesis testing, e.g., Jones and Tukey (2000),[3] or Killeen (2005).[4] These techniques are designed to ameliorate some of the problems discussed above, but any discussions on them, even of the most cursory type, lie outside of the intended scope of this chapter. Although not yet as popular as the classical techniques discussed here, they are worth exploring by the curious reader.

In the meantime, the key results of this chapter may be found in Table 15.12 at the very end of the chapter.

15.11 Summary and Conclusions

If the heart of statistics is inference—drawing conclusions about populations from information in a sample—then this chapter *and* Chapter 14 jointly constitute the heart of Part IV of this book. Following the procedures discussed in Chapter 14 for determining population parameters from sample data, we have focused primarily in this chapter on procedures by which one makes and tests the validity of assertive statements about these population parameters. Thus, with some perspective, we may now observe the following: in order to characterize a population fully using the information contained in a finite sample drawn from it, (a) the results of Chapter 13 enable us to characterize the variability in the sample, so that (b) the unknown parameters may be estimated with a prescribed degree of confidence using the techniques in Chapter 14; and (c) what these estimated parameters tell us about the true population characteristics is then framed in the form of hypotheses that are subsequently tested using the techniques presented in this chapter. Specifically, the null hypothesis, H_0, is stated as the status quo characteristic; this is then tested against an appropriate alternative that we are willing to entertain should there be sufficient evidence in the sample data against the validity of the null hypothesis—each null hypothesis and the specific competing alternative having been jointly designed to answer the specific question of interest.

[3] Jones, L. V., and J. W. Tukey. (2000), "A Sensible Formulation of the Significance Test," *Psych. Methods*, 5 (4), 411–414

[4] P.R. Killeen (2005), "An Alternative to Null-Hypothesis Significance Tests," *Psychol Sci*, 16(5), 345–353.

This has been a long chapter, and perhaps justifiably so, considering the sheer number of topics covered; but since hypotheses tests can be classified into a relatively small number of categories, the key results can be summarized briefly as we have done in Table 15.12 (found at the very end of the chapter). There are tests for population *means* (for single populations or two populations; with population variance known, or not known; with large samples or small); there are also tests for (normal) population *variances* (single variances or two); and then there are tests for *proportions* (one or two). In each case, once the appropriate test statistic is determined, with slight variations depending on specific circumstances, the principles are all the same. With fixed significance levels, α, the H_0 rejection regions are determined and are used straightforwardly to reach conclusions about each test. Alternatively, the p-value (also known as the *observed significance level*) is easily computed and used to reach conclusions. It bears restating that in carrying out the required computations not only in this chapter but in the book as a whole, we have consistently advocated the use of computer programs such as MINITAB. These programs are so widely available now that there is practically no need to make reference any longer to old-fashioned statistical tables. As a result, we have left out all but the most cursory references to any statistical tables, and instead included specific illustrations of how to use MINITAB (as an example software package).

The discussions of power and sample size considerations is important, both as a pre-experimentation design tool and as a post-analysis tool for ascertaining just how much stock one can realistically put in the result of a just-concluded test. Sadly, such considerations are usually given short-shrift by most students; this should *not* be the case. It is also easy to develop the mistaken notion that statistical inference is *only* concerned with Gaussian populations. Once more, as in Chapter 14, it is true that the *general* results we have presented have been limited to normal populations. This is due to the stubborn individuality of non-Gaussian distributions and the remarkable versatility of the Gaussian distribution both in representing truly Gaussian populations (of course), but also as a reasonable approximation to the sampling distribution of the means of most non-Gaussian populations. Nevertheless, the discussion in Section 15.8 and the overview of likelihood ratio tests in Section 15.9 should serve to remind the reader that there is statistical inference life beyond samples from normal populations. A few of the exercises and application problems at the end of the chapter also buttress this point.

There is a sense in which the completion of this chapter can justifiably be considered as a pivotal point in the journey that began with the illustrative examples of Chapter 1. These problems, posed long ago in that introductory chapter, have now been fully solved in this chapter; and, in a very real sense, many practical problems can now be solved using only the techniques discussed up until this point. But this chapter is actually a convenient launching point for the rest of the discussion in this book, not a stopping point. For example, we have only discussed how to compare at most two population means; when the

problem calls for the *simultaneous* comparison of more than two population means, the appropriate technique, ANOVA, is yet to be discussed. Although based on the F-test, to which we were introduced in this chapter, there is much more to the approach, as we shall see later, particularly in Chapter 19. Furthermore, ANOVA is only a part—albeit a foundational part—of Chapter 19, a chapter devoted to the design of experiments, the third pillar of statistics, which is concerned with ensuring that the samples used for statistical inference are as information rich as possible.

Immediately following this chapter, Chapter 16 (Regression Analysis) deals with estimation of a different kind, when the population parameters of interest are not constant as they have been thus far, but functions of another variable; naturally, much of the results of Chapter 14 and this current chapter are employed in dealing with such problems. Chapter 17 (Probability Model Validation) builds directly on the hypothesis testing results of this chapter in presenting techniques for explicitly validating postulated probability models; Chapter 18 (Nonparametric Methods) presents "distribution free" versions of many of the hypothesis tests discussed in this current chapter—a useful set of tools to have when one is unsure about the validity of the probability distributional assumptions (mostly the normality assumption) upon which classical tests are based. Even the remaining chapters beyond Chapter 19 (on case studies and special topics) all draw heavily from this chapter. A good grasp of the material in this chapter will therefore facilitate comprehension of the upcoming discussions in the remainder of the book.

REVIEW QUESTIONS

1. What is a statistical hypothesis?

2. What differentiates a simple hypothesis from a composite one?

3. What is H_0, the null hypothesis, and what is H_a, the alternative hypothesis?

4. What is the difference between a two-sided and a one-sided hypothesis?

5. What is a test of a statistical hypothesis?

6. How is the US legal system illustrative of hypothesis testing?

7. What is a test statistic?

8. What is a critical/rejection region?

9. What is the definition of the significance level of a hypothesis test?

10. What are the types of errors to which hypothesis tests are susceptible, and what are their legal counterparts?

11. What is the α-risk, and what is the β-risk?

12. What is the power of a hypothesis test, and how is it related to the β-risk?

13. What is the sensitivity of a test as opposed to the specificity of a test?

14. How are the performance measures, sensitivity and specificity, related to the α-risk and the β-risk?

15. What is the p-value, and why is it referred to as the *observed significance level*?

16. What is the general procedure for carrying out hypothesis testing?

17. What test statistic is used for hypotheses concerning the single mean of a normal population when the variance is *known*?

18. What is a z-test?

19. What is an "upper-tailed" test as opposed to a "lower-tailed" test?

20. What is the "one-sample" z-test?

21. What test statistic is used for hypotheses concerning the single mean of a normal population when the variance is *unknown*?

22. What is the "one-sample" t-test, and what differentiates it from the "one-sample" z-test?

23. How are confidence intervals related to hypothesis tests?

24. What test statistic is used for hypotheses concerning two normal population means when the variances are *known*?

25. What test statistic is used for hypotheses concerning two normal population means when the variances are *unknown* but equal?

26. What test statistic is used for hypotheses concerning two normal population means when the variances are *unknown* but unequal?

27. Is the distribution of the t-statistic used for the two-sample t-test with unknown and unequal variances an exact t-distribution?

28. What is a paired t-test, and what are the important characteristics that set the

problem apart from the general two-sample t-test?

29. In determining power and sample size, what is the "z-shift"?

30. In determining power and sample size, what are the three hypothesis test characteristic parameters making up the "z-shift"? What is the equation relating them to the α- and β-risks?

31. How can the α-risk be reduced without simultaneously increasing the β-risk?

32. What are some practical considerations discussed in this chapter regarding the determination of the power of a hypothesis test and sample size?

33. For general power and sample size determination problems, it is typical to specify which two problem characteristics, leaving which three parameters to be determined?

34. What is the test concerning the single variance of a normal population variance called?

35. What test statistic is used for hypotheses concerning the single variance of a normal population?

36. What test statistic is used for hypotheses concerning two variances from mutually independent normal populations?

37. What is the F-test?

38. The F-test is quite sensitive to which assumption?

39. What test statistic is used in the large sample approximation test concerning a single population proportion?

40. What is the objective criterion for ascertaining the validity of the large sample assumption in tests concerning a single population proportion?

41. What is involved in exact tests concerning a single population proportion?

42. What test statistic is used for hypotheses concerning two population proportions?

43. What is the central issue in testing hypotheses about non-Gaussian populations?

44. How does sample size influence how hypotheses about non-Gaussian populations are tested?

45. What options are available when testing hypotheses about non-Gaussian populations with small samples?

46. What are likelihood ratio tests?

47. What is the likelihood ratio test statistic?

48. Why is the likelihood ratio parameter λ such that $0 < \lambda < 1$? What does a value close to zero indicate? And what does a value close to 1 indicate?

49. Under what condition does the likelihood ratio test become identical to the familiar z-test?

50. Under what condition does the likelihood ratio test become identical to the familiar t-test?

51. What is the asymptotic distribution result for the likelihood ratio statistic?

52. What are some criticisms of hypothesis testing highlighted in this chapter?

53. In light of some of the criticisms discussed in this chapter, what recommendations have been proposed for approaching hypothesis testing in practice?

EXERCISES

Section 15.2

15.1 The target "mooney viscosity" of the elastomer produced in a commercial process is 44.0; if the average "mooney viscosity" of product samples acquired from the process hourly and analyzed in the quality control laboratory exceeds or falls below this target, the process is deemed "out of control" and in need of corrective control action. Formulate the decision-making about the process performance as a hypothesis test, stating the null and the alternative hypotheses.

15.2 A manufacturer of energy-saving light bulbs wants to establish that the lifetime of its new brand exceeds the specification of 1000 hours. State the appropriate null and alternative hypotheses.

15.3 A pharmaceutical company wishes to show that its newly developed acne medication reduces teenage acne by an average of 55% in the first week of usage. What are the null and alternative hypotheses?

15.4 The owner of a fleet of taxi cabs wants to determine if there is a difference in the lifetime of two different brands of car batteries used in the fleet of cabs. State the appropriate null and alternative hypotheses.

15.5 The safety coordinator of a manufacturing facility wishes to demonstrate that the mean time (in days) between safety incidents has deteriorated from the tradi-

TABLE 15.12: Summary of selected hypothesis tests and their characteristics

Population Parameter, θ (Null Hypothesis, H_0)	**Point Estimator**, $\hat{\Theta}$	**Test Statistic**	**Test**	H_0 **Rejection Condition**
μ; $(H_0 : \mu = \mu_0)$	$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$	$Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$	z-test	Table 15.2
Small sample $n < 30$	(S for unknown σ)	$T = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$	t-test	Table 15.3
$\delta = \mu_1 - \mu_2$; $(H_0 : \delta = \delta_0)$	$\bar{D} = \bar{X}_1 - \bar{X}_2$	$Z = \frac{\bar{D}-\delta_0}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}}$	2-sample z-test	Table 15.4
Small sample $n < 30$	$\left(S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}\right)$	$T = \frac{\bar{D}-\delta_0}{\sqrt{s_p^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}$	2-sample t-test	Table 15.5
$\delta = \mu_1 - \mu_2$; $(H_0 : \delta = \delta_0)$ (Paired)	$\bar{D} = \frac{\sum_{i=1}^{n} D_i}{n}$ $(D_i = X_{1_i} - X_{2_i})$ $\left(S_D^2 = \frac{\sum_{i=1}^{n}(D_i-\bar{D})^2}{n-1}\right)$	$T = \frac{\bar{D}-\delta_0}{S_D/\sqrt{n}}$	Paired t-test	Table 15.7
σ^2; $(H_0 : \sigma^2 = \sigma_0^2)$	$S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$	$C^2 = \frac{(n-1)S^2}{\sigma_0^2}$	Chi-squared-test	Table 15.9
σ_1^2/σ_2^2; $(H_0 : \sigma_1^2 = \sigma_2^2)$	S_1^2/S_2^2	$F = \frac{S_1^2}{S_2^2}$	F-test	Table 15.10

Chapter 16

Regression Analysis

16.1 Introductory Concepts 644
16.1.1 Dependent and Independent Variables 646
16.1.2 The Principle of Least Squares 647
16.2 Simple Linear Regression 648
16.2.1 One-Parameter Model 648
16.2.2 Two-Parameter Model 649
Primary Model Assumption 650
Ordinary Least Squares (OLS) Estimates 650
Maximum Likelihood Estimates 653
Actual Regression Line and Residuals 653
16.2.3 Properties of OLS Estimators 656
16.2.4 Confidence Intervals 657
Slope and Intercept Parameters 657
Regression Line 659
16.2.5 Hypothesis Testing 660
16.2.6 Prediction and Prediction Intervals 664
16.2.7 Coefficient of Determination and the F-Test 666
Orthogonal Decomposition of Variability 666
R^2, The Coefficient of Determination 668
F-Test for Significance of Regression 669
16.2.8 Relation to the Correlation Coefficient 672
16.2.9 Mean-Centered Model 673
16.2.10 Residual Analysis 673
16.3 "Intrinsically" Linear Regression 678
16.3.1 Linearity in Regression Models 678
16.3.2 Variable Transformations 681
16.4 Multiple Linear Regression 682
16.4.1 General Least Squares 683
16.4.2 Matrix Methods 684
Properties of the Estimates 685
Residuals Analysis 687
16.4.3 Some Important Special Cases 690
Weighted Least Squares 690
Constrained Least Squares 692
Ridge Regression 692
16.4.4 Recursive Least Squares 693
Problem Formulation 693
Recursive Least-Squares Estimation 694
16.5 Polynomial Regression 696
16.5.1 General Considerations 696
16.5.2 Orthogonal Polynomial Regression 700
An Example: Gram Polynomials 700
Application in Regression 704
16.6 Summary and Conclusions 706
REVIEW QUESTIONS 707

EXERCISES 709
APPLICATION PROBLEMS 715

The mathematical facts worthy of being studied are those which, by their analogy with other facts are capable of leading us to the knowledge of a mathematical law just as experimental facts lead us to the knowledge of a physical law.

Henri Poicaré (1854–1912)

It is often the case in many practical problems that the variability observed in a random variable, Y, consists of more than just the purely randomly varying phenomena that have occupied our attention up till now. For this new class of problems, an underlying functional relationship exists between Y and an independent variable, x (deliberately written in the lower case for reasons that will soon become clear), with a purely random component superimposed on this otherwise deterministic component. This chapter is devoted to dealing with problems of this kind. The values observed for the random variable Y depend on the values of the (deterministic) variable, x, and, were it not for the presence of the purely random component, Y would have been perfectly predictable given x. Regression analysis is concerned with obtaining, from data, the best estimate of the relationship between Y and x.

Although apparently different from what we have dealt with up until now, we will see that regression analysis in fact builds directly upon many of the results obtained thus far, especially estimation and hypothesis testing.

16.1 Introductory Concepts

Consider the data in Table 16.1 showing the boiling point (in $^{\circ}C$) of 8 hydrocarbons in a homologous series, along with n, the number of carbon atoms in each molecule. A scatter plot of boiling point versus n is shown in Fig 16.1, where we notice right away that as the number of carbon atoms in this homologous series increases, so does the boiling point of the hydrocarbon compound. In fact, the implied relationship between these two variables appears to be so strong that one is immediately inclined to conclude that it must be possible to predict the boiling point of compounds in this series on the basis of the number of carbon atoms. There is therefore no doubt that there is some sort of a functional relationship between n and boiling point. If determined "correctly," such a relationship will provide, among other things, a simple way to capture the extensive data on such "physical properties" of compounds in this particular homologous series.

TABLE 16.1: Boiling points of a series of hydrocarbons

Hydrocarbon Compound	n, Number of Carbon Atoms	Boiling Point $^{\circ}C$
Methane	1	-162
Ethane	2	-88
Propane	3	-42
n-Butane	4	1
n-Pentane	5	36
n-Hexane	6	69
n-Heptane	7	98
n-Octane	8	126

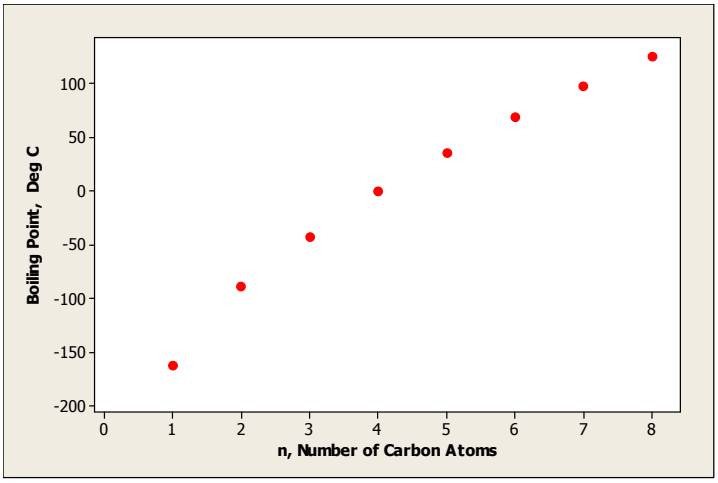

FIGURE 16.1: Boiling point of hydrocarbons in Table 16.1 as a function of the number of carbon atoms in the compound.

16.1.1 Dependent and Independent Variables

Many cases such as the one illustrated above arise in science and engineering where the value taken by one variable appears to depend on the value taken by another. Not surprisingly, it is customary to refer the variable whose value depends on the value of another as the *dependent* variable, while the other variable is known as the *independent* variable. It is often desired to capture the relationship between these two variables in some mathematical form. However, because of measurement errors and other sources of variability, this exercise requires the use of probabilistic and statistical techniques. Under these circumstances, the independent variable is considered as a fixed, deterministic quantity that is not subject to random variability. This is perfectly exemplified in n, the number of carbon atoms in the hydrocarbon compounds of Table 16.1; it is a known quantity not subject to random variability. The dependent variable, on the other hand, is the random variable, subject to a wide variety of potential sources of random variability, including, but not limited to measurement uncertainties. The dependent variable is therefore represented as the random variable, Y, while the independent variable is represented as the deterministic variable, x, represented in the lower case to underscore its deterministic nature.

The variability observed in the random variable, Y, is typically considered to consist of two distinct components, i.e., for each observation, $Y_i, i = 1, 2, \ldots, n$:

$$Y_i = g(x_i; \boldsymbol{\theta}) + \epsilon_i \tag{16.1}$$

where $g(x_i; \boldsymbol{\theta})$ is the deterministic component, a functional relationship, with $\boldsymbol{\theta}$ as a set of unknown parameters, and ϵ_i is the random component. The deterministic mathematical relationship between these two variables is a "model" of how the independent x (also known as the "predictor") affects the the predictable part of the dependent Y, sometimes known as the "response."

In some cases, the functional form of $g(x_i)$ is known from fundamental scientific principles. For example, if Y is the distance (in cm) traveled in time t_i seconds by a particle launched with an initial velocity, u (cm/sec), and traveling at a constant acceleration a (cm/sec^2), then we know that

$$g(t_i; u, a) = ut_i + \frac{1}{2}at_i^2 \tag{16.2}$$

with $\boldsymbol{\theta} = (u, a)$ as the parameters.

In most cases, however, there is no such fundamental scientific principle to suggest an appropriate form for $g(x_i; \boldsymbol{\theta})$; simple forms (typically polynomials) are postulated and validated with data, as we show subsequently. The result in this case is known as an "empirical" model because it is strictly dependent on data and not on some known fundamental scientific principle.

Regression analysis is primarily concerned with the following tasks:

- Obtaining the "best estimates" $\hat{\boldsymbol{\theta}}$ for the model parameters, $\boldsymbol{\theta}$;

- Characterizing the random sequence ϵ_i; and,
- Making inference about the parameter estimates, $\hat{\boldsymbol{\theta}}$.

The classical treatment is based on "least squares estimation" which we will discuss briefly now, before using it in the context of regression.

16.1.2 The Principle of Least Squares

Consider the case where the random sample, $Y_1, Y_2, \ldots, Y_n$, is drawn from a population characterized by a single, constant parameter, θ, the population mean. The random variable Y may then be written as:

$$Y_i = \theta + \epsilon_i \tag{16.3}$$

where the observed random variability is due to random component ϵ_i. Furthermore, let the variance of Y be σ^2. Then from Eq (16.3), we obtain:

$$E[Y_i] = \theta + E[\epsilon_i] \tag{16.4}$$

and since, by definition, $E[Y_i] = \theta$, this implies that $E[\epsilon_i] = 0$. Furthermore,

$$Var(Y_i) = Var(\epsilon_i) = \sigma^2 \tag{16.5}$$

since θ is a constant. Thus, from the fact that Y has a distribution (unspecified) with mean θ and variance σ^2 implies that in Eq (16.3), the random "error" term, ϵ_i has zero mean and variance σ^2.

To estimate θ from the given random sample, it seems reasonable to choose a value that is "as close as possible" to all the observed data. This concept may be represented mathematically as:

$$\min_{\theta} S(\theta) = \sum_{i=1}^{n} (Y_i - \theta)^2 \tag{16.6}$$

The usual calculus approach to this optimization problem leads to:

$$\left.\frac{\partial S}{\partial \theta}\right|_{\theta=\hat{\theta}} = -2\sum_{i=1}^{n}(Y_i - \hat{\theta}) = 0 \tag{16.7}$$

which, when solved, produces the result:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{16.8}$$

A second derivative with respect to θ yields

$$\frac{\partial^2 S}{\partial \theta^2} = 2n > 0 \tag{16.9}$$

so that indeed $S(\theta)$ achieves a minimum for $\theta = \hat{\theta}$ in Eq (16.8).

The quantity $\hat{\theta}$ in Eq (16.8) is referred to as a least-squares estimator for θ in Eq (16.3), for the obvious reason that the value produced by this estimator achieves the minimum for the sum-of-squared deviation implied in Eq (16.6). It should not be lost on the reader that this estimator is also precisely the same as the familiar sample average.

The problems we have dealt with up until now may be represented in the form shown in Eq (16.3). In that context, the probability models we developed earlier may now be interpreted as models for ϵ_i, the random variation around the constant random variable mean. This allows us to put the upcoming discussion on the regression problem in context of the earlier discussions.

Finally, we note that the principle of least-squares also affords us the flexibility to treat each observation, Y_i, differently in how it contributes to the estimation of θ. This is done by applying appropriate weights W_i to Eq (16.3) to obtain:

$$W_i Y_i = W_i\theta + W_i\epsilon_i \tag{16.10}$$

Consequently, for example, more reliable observations can be assigned larger weights than less reliable ones. Upon using the same calculus techniques, the least-squares estimate in this case can be shown to be:

$$\hat{\theta}_\omega = \frac{\sum_{i=1}^{n} W_i^2 Y_i}{\sum_{i=1}^{n} W_i^2} = \sum_{i=1}^{n} \omega_i Y_i \tag{16.11}$$

(see Exercise 16.2) where

$$\omega_i = \frac{W_i^2}{\sum_{i=1}^{n} W_i^2} \tag{16.12}$$

Note that $0 < \omega_i < 1$. The result in Eq (16.11) is therefore an appropriately weighted average—a generalization of Eq (16.8) where $\omega_i = 1/n$. This variation on the least-squares approach is known appropriately as "weighted least-squares"; we shall encounter it later in this chapter.

16.2 Simple Linear Regression

16.2.1 One-Parameter Model

As a direct extension of Eq (16.3), let the relationship between the random variable Y and the independent (deterministic) variable, x, be:

$$Y = \theta x + \epsilon \tag{16.13}$$

where the random error, ϵ, has zero mean and constant variance, σ^2. Then, $E(Y|x)$, the conditional expectation of Y given a specific value for x is:

$$\mu_{Y|x} = E(Y|x) = \theta x \tag{16.14}$$

recognizable as the equation of a straight line with slope θ and zero intercept. It is also known as the "one-parameter" regression model, a classic example of which is the famous Ohm's law in physics: the relationship between the voltage, V, across a resistor with unknown resistance, R, and the current I flowing through the resistive element, i.e.,

$$V = IR \tag{16.15}$$

From data $y_i; i = 1, 2, \ldots, n$, actual values of the random variable, Y_i, observed for corresponding values of x_i, the problem at hand is to obtain an estimate of the characterizing parameter θ. Using the method of least-squares outlined above requires minimizing the sum-of-squares function:

$$S(\theta) = \sum_{i=1}^{n} (y_i - \theta x_i)^2 \tag{16.16}$$

from where $\partial S/\partial\theta = 0$ yields:

$$-2\sum_{i=1}^{n} x_i(y_i - \theta x_i) = 0 \tag{16.17}$$

which is solved for θ to obtain:

$$\hat{\theta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{16.18}$$

This is the expression for the slope of the "best" (i.e., least-squares) straight line (with zero intercept) through the points (x_i, y_i).

16.2.2 Two-Parameter Model

More general is the two-parameter model,

$$Y = \theta_0 + \theta_1 x + \epsilon \tag{16.19}$$

indicating a functional relationship, $g(x; \boldsymbol{\theta})$, that is a straight line with slope θ_1 and potentially non-zero intercept θ_0 as the parameters, i.e.,

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \end{bmatrix} \tag{16.20}$$

along with $E(\epsilon) = 0$; $Var(\epsilon) = \sigma^2$. In this case, the conditional expectation of Y given a specific value for x is given by:

$$\mu_{Y|x} = E(Y|x) = \theta_0 + \theta_1 x \tag{16.21}$$

In this particular case, regression analysis is primarily concerned with obtaining the best estimates for $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1)$; characterizing the random sequence ϵ_i; and, making inference about the parameter estimates, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1)$.

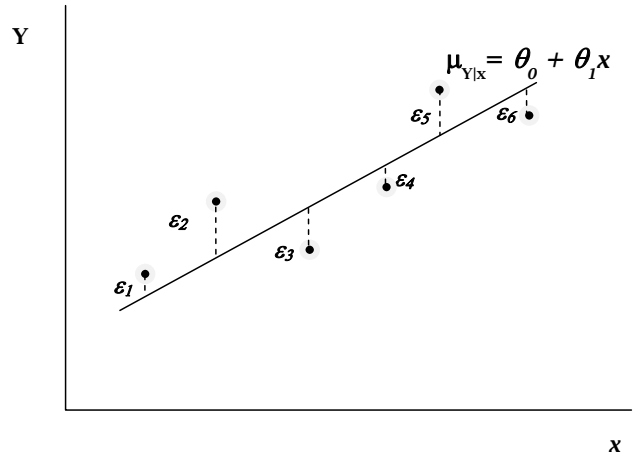

FIGURE 16.2: The true regression line and the zero mean random error ϵ_i.

Primary Model Assumption

In this case, the true but unknown regression line is represented by Eq (16.21), with data scattered around it. The fact that $E(\epsilon) = 0$, indicates that the data scatters "evenly" around the true line; more precisely, the data varies randomly around a mean value that is the function of x defined by the true but unknown regression line in Eq (16.21). This is illustrated in Fig 16.2.

It is typical to assume that each ϵ_i, the random component of the model, is mutually independent of the others and follows a Gaussian distribution with zero mean and variance σ^2, i.e., $\epsilon_i \sim N(0, \sigma^2)$. The implication in this particular case is therefore that each data point, (x_i, y_i), comes from a Gaussian distribution whose mean is dependent on the value of x, and falls on the true regression line, as illustrated in Fig 16.3. Equivalently, the true regression line passes through the mean of the series of Gaussian distributions having the same variance. The two main assumptions underlying regression analysis may now be summarized as follows:

1. ϵ_i forms an independent random sequence, with zero mean and variance σ^2 that is constant for all x;
2. $\epsilon_i \sim N(0, \sigma^2)$ so that $Y_i \sim (\theta_0 + \theta_1 x, \sigma^2)$

Ordinary Least Squares (OLS) Estimates

Obtaining the least-squares estimates of the intercept, θ_0, and slope, θ_1, from data (x_i, y_i) involves minimizing the sum-of-squares function,

$$S(\theta_0, \theta_1) = \sum_{i=1}^{n} [y_i - (\theta_1 x_i + \theta_0)]^2 \tag{16.22}$$

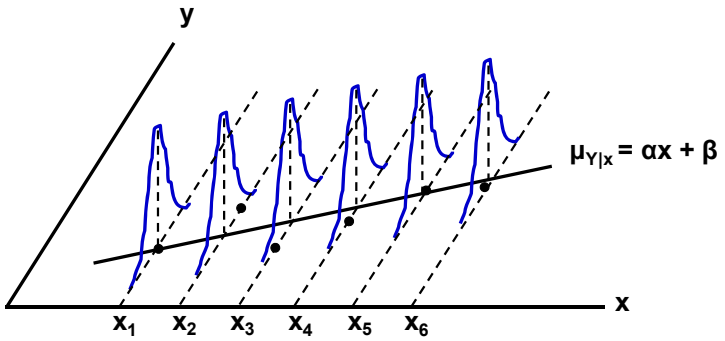

FIGURE 16.3: The Gaussian assumption regarding variability around the true regression line giving rise to $\epsilon \sim N(0, \sigma^2)$. The 6 points represent the data at $x_1, x_2, \ldots, x_6$; the solid straight line is the true regression line which passes through the mean of the sequence of the indicated Gaussian distributions.

where the usual first derivatives of the calculus approach yield:

$$\frac{\partial S}{\partial \theta_0} = 2\sum_{i=1}^{n}[y_i - (\theta_1 x_i + \theta_0)] = 0 \tag{16.23}$$

$$\frac{\partial S}{\partial \theta_1} = -2\sum_{i=1}^{n} x_i[y_i - (\theta_1 x_i + \theta_0)] = 0 \tag{16.24}$$

These expressions rearrange to give:

$$\theta_1 \sum_{i=1}^{n} x_i + \theta_0 n = \sum_{i=1}^{n} y_i \tag{16.25}$$

$$\theta_1 \sum_{i=1}^{n} x_i^2 + \theta_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i \tag{16.26}$$

collectively known as the "normal equations," to be solved simultaneously to produce the least squares estimates, $\hat{\theta}_0$, and $\hat{\theta}_1$.

Before solving these equations explicitly, we wish to direct the reader's attention to a pattern underlying the emergence of the normal equations. Beginning with the original two-parameter model equation:

$$y_i = \theta_1 x_i + \theta_0 + \epsilon_i$$

a summation across each term yields:

$$\sum_{i=1}^{n} y_i = \theta_1 \sum_{i=1}^{n} x_i + \theta_0 n \tag{16.27}$$

where the last term involving ϵ_i has vanished upon the assumption that n is

sufficiently large so that because $E(\epsilon_i) = 0$, the sum will be close to zero (a point worth keeping in mind to remind the reader that the result of solving the normal equations provide estimates, not "precise" values).

Also, multiplying the model equation by x_i and summing yields:

$$\sum_{i=1}^{n} y_i x_i = \theta_1 \sum_{i=1}^{n} x_i^2 + \theta_0 \sum_{i=1}^{n} x_i \tag{16.28}$$

where once again the last term involving ϵ_i has vanished because of independence with x_i; and the assumption once again that n is sufficiently large that the sum will be close to zero. Note that these two equations are identical to the normal equations; more importantly, as derived by summation from the original model they are the *sample equivalents* of the following expectations:

$$E(Y) = \theta_1 E(x) + \theta_0 \tag{16.29}$$

$$E(Yx) = \theta_1 E(x^2) + \theta_0 E(x) \tag{16.30}$$

which should help put the emergence of the normal equations into perspective.

Returning to the task of computing least-squares estimates of the two model parameters, let us define the following terms:

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{16.31}$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{16.32}$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{16.33}$$

where $\bar{y} = (\sum_{i=1}^{n} y_i)/n$ and $\bar{x} = (\sum_{i=1}^{n} x_i)/n$ represent the usual averages. When expanded out and consolidated, these equations yield:

$$nS_{xx} = n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 \tag{16.34}$$

$$nS_{yy} = n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2 \tag{16.35}$$

$$nS_{xy} = n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right) \tag{16.36}$$

These terms, clearly related to sample variances and covariances, allow us to solve Eqs (16.25) and (16.26) simultaneously to obtain the results:

$$\hat{\theta}_1 = \frac{S_{xy}}{S_{xx}} \tag{16.37}$$

$$\hat{\theta}_0 = \bar{y} - \hat{\theta}_1 \bar{x} \tag{16.38}$$

Nowadays, such computations implied in this derivation are no longer carried out by hand, of course, but by computer programs; the foregoing discussion is therefore intended to acquaint the reader with the principles and mechanics underlying the numbers produced by the statistical software packages.

Maximum Likelihood Estimates

Under the Gaussian assumption, the regression equation, written in the more general form,

$$Y = \eta(x, \boldsymbol{\theta}) + \epsilon, \tag{16.39}$$

implies that the observations $Y_1, Y_2, \ldots, Y_n$ come from a Gaussian distribution with mean η and variance, σ^2; i.e., $Y \sim N(\eta(x, \boldsymbol{\theta}), \sigma^2)$. If the data can be considered as a random sample from this distribution, then the method of maximum likelihood presented in Chapter 14 may be used to estimate $\eta(x, \boldsymbol{\theta})$ and σ^2 in precisely the same manner in which estimates of the $N(\mu, \sigma^2)$ population parameters were determined in Section 14.3.2. The only difference this time is that the population mean, $\eta(x, \boldsymbol{\theta})$, is no longer constant, but a function of x. It can be shown (see Exercise 16.5) that when the variance σ^2 is constant, the maximum likelihood estimate for θ in the one-parameter model,

$$\eta(x, \boldsymbol{\theta}) = \theta x \tag{16.40}$$

and the maximum likelihood estimates for (θ_0, θ_1) in the two-parameter model,

$$\eta(x, \boldsymbol{\theta}) = \theta_0 + \theta_1 x \tag{16.41}$$

are each identical to the corresponding least squares estimates obtained in Eq (16.18) and in Eqs (16.38) and (16.37) respectively. It can also be shown (see Exercise 16.6) that when the variance, σ_i^2, associated with each observation, Y_i, $i = 1, 2, \ldots, n$, differs from observation to observation, the maximum likelihood estimates for the parameters θ in the first case, and for (θ_0, θ_1) in the second case, are the same as the corresponding weighted least squares estimates, with weights related to the reciprocal of σ_i.

Actual Regression Line and Residuals

In the same manner in which the true (constant) mean, μ, of a Gaussian distribution producing the random sample $X_1, X_2, \ldots, X_n$, is not known, only estimated by the sample average $\bar{X}$, the true regression line is also never known but estimated. When the least-squares estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ are introduced into the original model, the result is the estimated observation $\hat{y}$ defined by:

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \tag{16.42}$$

This is not the same as the true theoretical $\mu_{Y|x}$ in Eq (16.21) because, in general $\hat{\theta}_0 \neq \theta_0$ and $\hat{\theta}_1 \neq \theta_1$; $\hat{y}_i$ is the two-parameter model's best estimate

TABLE 16.2:
Density (in gm/cc) and weight percent of ethanol in ethanol-water mixture

Density (g/cc)	Wt % Ethanol
0.99823	0
0.98938	5
0.98187	10
0.97514	15
0.96864	20
0.96168	25
0.95382	30
0.94494	35

(or prediction) of the true but unknown value of the observation y_i (unknown because of the additional random effect, ϵ_i). If we now define as e_i, the error between the actual observation and the estimated value, i.e.,

$$e_i = y_i - \hat{y}_i \tag{16.43}$$

this term is known as the residual error or simply the "residual"; it is our best estimate of the unknown ϵ_i, just as $\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x$ is our best estimate of the true regression line $\mu_{Y|x} = E(Y|x) = \theta_1 x + \theta_0$.

As discussed shortly (Section 16.2.10), the nature of the sequence of residuals provides a great deal of information about how well the model represents the observations.

Example 16.1: DENSITY OF ETHANOL-WATER MIXTURE

An experimental investigation into how the density of an ethanol-water mixture varies with weight percent of ethanol in the mixture yielded the result shown in Table 16.2. Postulate a linear two-parameter model as in Eq (16.19), and use the supplied data to obtain least-squares estimates of the slope and intercept, and also the residuals. Plot the data versus the model and comment on the fit.

Solution:

Given this data set, just about any software package, from Excel to MATLAB and MINITAB, will produce the following estimates:

$$\hat{\theta}_1 = -0.001471; \hat{\theta}_0 = 0.9975 \tag{16.44}$$

so that, if y is the density and x is the wt % of ethanol, the regression model fit to this data is given as:

$$\hat{y} = -0.001471x + 0.9975 \tag{16.45}$$

The model fit to the data is shown in Fig 16.4; and for the given values

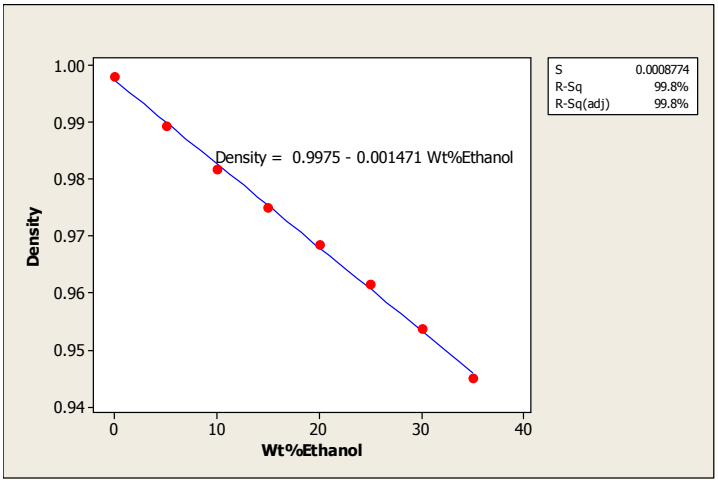

FIGURE 16.4: The fitted straight line to the density versus ethanol weight percent data. The additional terms included in the graph—S, R-Sq, and R-Sq(adj)—are discussed later.

of x, the estimated $\hat{y}$, and the residuals, e, are shown in Table 16.3. Visually, the model seems to fit quite well. This model allows us to predict solution density for any given weight percent of ethanol within the experimental data range but not actually part of the data. For example, for $x = 7.5$, Eq (16.45) estimates $\hat{y} = 0.98647$. How the residuals are analyzed is discussed in Section 16.2.10.

Expressions such as the one obtained in this example, Eq (16.45), are sometimes known as calibration curves. Such curves are used to calibrate measurement devices such as thermocouples, where the raw instrument output (say millivolts) is converted to the actual desired measurement (say temperature in $^{\circ}C$) based on expressions such as the one obtained here. Such expressions are

TABLE 16.3: Density and weight percent of ethanol in ethanol-water mixture: model fit and residual errors

Density (g/cc) y	Wt % Ethanol x	Estimated Density, $\hat{y}$	Residual Errors, e
0.99823	0	0.997500	0.000730
0.98938	5	0.990145	-0.000765
0.98187	10	0.982790	-0.000920
0.97514	15	0.975435	-0.000295
0.96864	20	0.968080	0.000560
0.96168	25	0.960725	0.000955
0.95382	30	0.953370	0.000450
0.94494	35	0.946015	-0.001075

typically generated from standardized experiments where data on instrument output are gathered for various objects with known temperature.

16.2.3 Properties of OLS Estimators

When experiments are repeated for the same fixed values x_i, as a typical consequence of random variation, the corresponding value observed for Y_i will differ each time. The resulting estimates provided in Eqs (16.37) and (16.38) therefore will also change slightly each time. In typical fashion, therefore, the specific parameter estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ are properly considered as realizations of the respective estimators Θ_0 and Θ_1, random variables that depend on the random sample $Y_1, Y_2, \ldots, Y_n$. It will be desirable to investigate the theoretical properties of these estimators defined by:

$$\Theta_1 = \frac{S_{xy}}{S_{xx}} \tag{16.46}$$

$$\Theta_0 = \bar{Y} - \Theta_1 \bar{x} \tag{16.47}$$

Let us begin with the expected values of these estimators. From here, we observe that

$$E(\Theta_1) = E\left(\frac{S_{xy}}{S_{xx}}\right) \tag{16.48}$$

which, from the definitions given above, becomes:

$$E(\Theta_1) = \frac{1}{S_{xx}} E\left[\sum_{i=1}^{n} Y_i(x_i - \bar{x})\right] \tag{16.49}$$

(because $\sum_{i=1}^{n} \bar{Y}(x_i - \bar{x}) = 0$, since $\bar{Y}$ is a constant); and upon introducing Eq (16.19) in for Y_i, we obtain:

$$E(\Theta_1) = \frac{1}{S_{xx}} E\left[\sum_{i=1}^{n} (\theta_1 x_i + \theta_0 + \epsilon_i)(x_i - \bar{x})\right] \tag{16.50}$$

A term-by-term expansion and subsequent simplification results in

$$E(\Theta_1) = \frac{1}{S_{xx}} E\left[\theta_1 \sum_{i=1}^{n} (x_i - \bar{x})\right] \tag{16.51}$$

because $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$ and $E[\sum_{i=1}^{n} \epsilon_i (x_i - \bar{x})] = 0$ since $E(\epsilon_i) = 0$. Hence, Eq (16.51) simplifies to

$$E(\Theta_1) = \frac{1}{S_{xx}} \theta_1 S_{xx} = \theta_1 \tag{16.52}$$

indicating that Θ_1 is an unbiased estimator of θ_1, the true slope.

Similarly, from Eq (16.47), we obtain:

$$E(\Theta_0) = E(\bar{Y} - \Theta_1\bar{x}) = E(\bar{Y}) - E(\Theta_1)\bar{x} \tag{16.53}$$

which by virtue of Eq (16.51) simplifies to:

$$E(\Theta_0) = \theta_1\bar{x} + \theta_0 - \theta_1\bar{x} = \theta_0 \tag{16.54}$$

so that Θ_0 is also an unbiased estimator for θ_0, the true intercept.

In similar fashion, by definition of the variance of a random variable, it is straightforward to show that:

$$Var(\Theta_1) = \sigma^2_{\Theta_1} = \frac{\sigma^2}{S_{xx}} \tag{16.55}$$

$$Var(\Theta_0) = \sigma^2_{\Theta_0} = \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right) \tag{16.56}$$

where σ^2 is the variance of the random component, ϵ. Consequently, the standard error of each estimate, the positive square root of the variance, is given by:

$$SE(\Theta_1) = \frac{\sigma}{\sqrt{S_{xx}}} \tag{16.57}$$

$$SE(\Theta_0) = \sigma\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \tag{16.58}$$

16.2.4 Confidence Intervals

As with all estimation problems, the point estimates obtained above for the regression parameters, θ_0 and θ_1, by themselves are insufficient in making decisions about their true, but unknown values; we must add a measure of how precise these estimates are. Obtaining interval estimates is one option; and such interval estimates are determined for regression parameters essentially by the same procedure as that presented in Chapter 14 for population parameters. This, of course, requires sampling distributions.

Slope and Intercept Parameters

Under the Gaussian distributional assumption for ϵ, with the implication that the sample $Y_1, Y_2, \ldots, Y_n$, possesses the distribution $N(\theta_0 + \theta_1 x, \sigma^2)$, and from the results obtained above about the characteristics of the estimates, it can be shown that the random variables Θ_1 and Θ_0, respectively the slope and the intercept, are distributed as $\Theta_1 \sim N(\theta_1, \sigma^2_{\Theta_1})$ and $\Theta_0 \sim N(\theta_0, \sigma^2_{\Theta_0})$ with the variances as shown in Eqs (16.55) and (16.56), *provided the data variance*, σ^2, *is known*. However, this variance is not known and must be estimated from data. This is done as follows for this particular problem.

Consider residual errors, e_i, our best estimates of ϵ_i; define the residual error sum of squares as

$$\begin{aligned} SS_E &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 && (16.59)\\ &= \sum_{i=1}^{n}[y_i - (\hat{\theta}_1 x_i + \hat{\theta}_0)]^2 \\ &= \sum_{i=1}^{n}[(y_i - \bar{y}) - \hat{\theta}_1(x_i - \bar{x})]^2 && (16.60) \end{aligned}$$

which, upon expansion and simplification reduces to:

$$SS_E = S_{yy} - \hat{\theta}_1 Sxy \tag{16.61}$$

It can be shown that

$$E(SS_E) = (n-2)\sigma^2 \tag{16.62}$$

as a result, the mean squared error, s_e^2, defined as:

$$s_e^2 = \frac{SS_E}{(n-2)} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_1)^2}{n-2} \tag{16.63}$$

is an unbiased estimate of σ^2.

Now, as with previous statistical inference problems concerning normal populations with unknown σ, by substituting s_e^2, the mean residual sum-of-squares, for σ^2, we have the following results: the statistics T_1 and T_0 defined as:

$$T_1 = \frac{\Theta_1 - \theta_1}{s_e/\sqrt{S_{xx}}} \tag{16.64}$$

and

$$T_0 = \frac{\Theta_0 - \theta_0}{s_e\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \tag{16.65}$$

each possess t-distribution with $\nu = n-2$ degrees of freedom. The immediate implications are therefore that

$$\theta_1 = \hat{\theta}_1 \pm t_{\alpha/2}(n-2)\frac{s_e}{\sqrt{S_{xx}}} \tag{16.66}$$

$$\theta_0 = \hat{\theta}_0 \pm t_{\alpha/2}(n-2)s_e\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)} \tag{16.67}$$

constitute $(1-\alpha)\times 100\%$ confidence intervals around the slope and intercept estimates, respectively.

Example 16.2: CONFIDENCE INTERVAL ESTIMATES FOR THE SLOPE AND INTERCEPT OF ETHANOL-WATER MIXTURE DENSITY REGRESSION MODEL

Obtain 95% confidence interval estimates for the slope and intercept of the regression model obtained in Example 16.1 for the ethanol-water mixture density data.

Solution:

In carrying out the regression in Example 16.1 with MINITAB, part of the computer program output is the set of standard errors. In this case, $SE(\Theta_1) = 0.00002708$ for the slope, and $SE(\Theta_0) = 0.000566$ for the intercept. These could also be computed by hand (although not recommended). Since the data set consists of 8 data points, we obtain the required $t_{0.025}(6) = 2.447$ from the cumulative probability feature. The required 95% confidence intervals are therefore obtained as follows:

$$\theta_1 = -0.001471 \pm 0.00006607 \tag{16.68}$$

$$\theta_0 = 0.9975 \pm 0.001385 \tag{16.69}$$

Note that none of these two intervals includes 0.

Regression Line

The actual regression line fit (see for example Fig 16.4), an estimate of the true but unknown regression line, is obtained by introducing into Eq (16.21), the estimates for the slope and intercept parameters to give

$$\hat{\mu}_{Y|x} = \hat{\theta}_1 x + \hat{\theta}_0 \tag{16.70}$$

For any specific value $x = x^*$, the value

$$\hat{\mu}_{Y|x^*} = \hat{\theta}_1 x^* + \hat{\theta}_0 \tag{16.71}$$

is the estimate of the actual response of Y at this point (akin to the sample average estimate of a true but unknown population mean).

In the same manner in which we obtained confidence intervals for sample averages, we can also obtain a confidence interval for $\hat{\mu}_{Y|x^*}$. It can be shown from Eq (16.71) (and Eq (16.56)) that the associated variance is:

$$Var(\hat{\mu}_{Y|x^*}) = \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \tag{16.72}$$

and because of the normality of the random variables Θ_0 and Θ_1, then if σ is known, $\hat{\mu}_{Y|x^*}$ has a normal distribution with mean $(\hat{\theta}_1 x^* + \hat{\theta}_0)$ and variance shown in Eq (16.72). With σ unknown, substituting s_e for it, as in the previous section, leads to the result that the specific statistic,

$$t_{RL} = \frac{(\hat{\mu}_{Y|x^*} - \mu_{Y|x^*})}{s_e \sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]}} \tag{16.73}$$

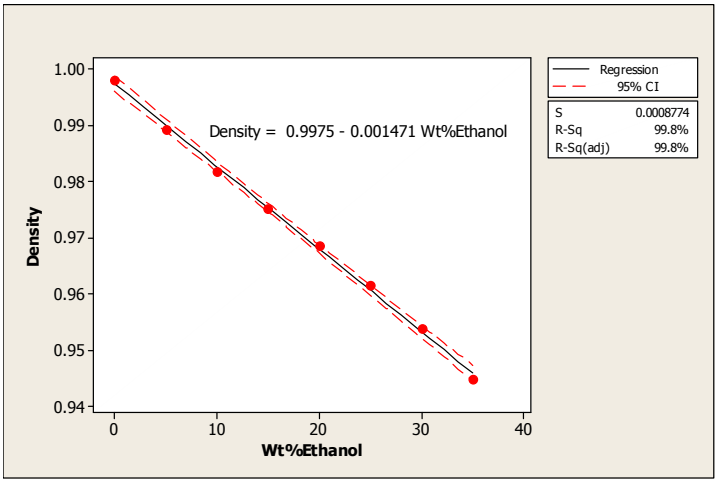

FIGURE 16.5: The fitted regression line to the density versus ethanol weight percent data (solid line) along with the 95% confidence interval (dashed line). The confidence interval is narrowest at $x = \bar{x}$ and widens for values further away from $\bar{x}$.

has a t-distribution with $\nu = (n-2)$ degrees of freedom. As a result, the $(1-\alpha) \times 100\%$ confidence interval on the regression line (mean response) at $x = x^*$, is:

$$\hat{\mu}_{Y|x^*} = (\hat{\theta}_1 x^* + \hat{\theta}_0) \pm t_{\alpha/2}(n-2)s_e\sqrt{\left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]} \tag{16.74}$$

When this confidence interval is computed for all values of x of interest, the result is a confidence interval around the entire regression line. Again, as most statistical analysis software packages have the capability to compute and plot this confidence interval along with the regression line, the primary objective of this discussion is to provide the reader with a fundamental understanding of the theoretical bases for these computer outputs. For example, the 95% confidence interval for the density versus weight percent ethanol problem in Examples 16.1 and 16.2 is shown in Fig 16.5.

By virtue of the $(x^* - \bar{x})^2$ term in Eq (16.74), a signature characteristic of these confidence intervals is that they are narrowest when $x^* = \bar{x}$ and widen for values further away from $\bar{x}$.

16.2.5 Hypothesis Testing

For this class of problems, the hypothesis of concern is whether or not there is a real (and significant) linear functional relationship between x and Y; i.e., whether the slope parameter, $\theta_1 = 0$, in which case the variation in Y is purely random around a constant mean value θ_0 (which may or may

not be zero). This translates to the following hypotheses regarding the slope parameter:

$$\begin{aligned} H_0 &: \theta_1 = 0 \\ H_a &: \theta_1 \neq 0 \end{aligned} \tag{16.75}$$

And from the preceding discussion regarding confidence intervals, the appropriate test statistic for this test, from Eq (16.64), is:

$$t_1 = \frac{\hat{\theta_1}}{s_e/\sqrt{S_{xx}}} \tag{16.76}$$

since the postulated value for the unknown θ_1 is 0; and the decision to reject or not reject H_0 follows the standard two-sided t-test criteria; i.e., at the significance level α, H_0 is rejected when

$$t_1 < -t_{\alpha/2}(n-2), \text{ or } t_1 > t_{\alpha/2}(n-2) \tag{16.77}$$

As with previous results, these conditions are identical to the $(1-\alpha)\times 100\%$ confidence interval on θ_1 not containing zero. When there is sufficient reason to reject H_0, the estimated regression coefficient is said to be "significant," by which we mean that it is significantly different from zero, at the significance level α.

There is nothing to prevent testing hypotheses also about the intercept parameter, θ_0, whether or not its value is significantly different from zero. The principles are precisely as indicated above for the slope parameter; the hypotheses are,

$$\begin{aligned} H_0 &: \theta_0 = 0 \\ H_a &: \theta_0 \neq 0 \end{aligned} \tag{16.78}$$

in this case, with the test statistic (from Eq (16.65)):

$$t_0 = \frac{\hat{\theta_0}}{s_e\sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)}} \tag{16.79}$$

and the rejection criteria,

$$t_0 < -t_{\alpha/2}(n-2), \text{ or } t_0 > t_{\alpha/2}(n-2) \tag{16.80}$$

In addition to computing estimates of the regression coefficient and the associated standard errors, most computer programs will also compute the t-statistics and the associated p-values for each of the two coefficients.

Let us illustrate with the following example.

TABLE 16.4: Cranial circumference and finger lengths for 16 individuals

Cranial Circum (cm)	58.5	54.2	57.2	52.7	55.1	60.7	57.2	58.8
Finger Length (cm)	7.6	7.9	8.4	7.7	8.6	8.6	7.9	8.2
Cranial Circum (cm)	56.2	60.7	53.5	60.7	56.3	58.1	56.6	57.7
Finger Length (cm)	7.7	8.1	8.1	7.9	8.1	8.2	7.8	7.9

Example 16.3: CRANIAL CIRCUMFERENCE AND FINGER LENGTH
A once-popular exercise in the late 19th and early 20th centuries involved attempts at finding mathematical expressions that will allow one to predict, for a population of humans, some physical human attribute on the basis of a different one. The data in Table 16.4 shows the result of a classic example of such an exercise where the cranial circumference (in cms) and the length of the longest finger (in cms) of 16 individuals were determined. Postulate a linear two-parameter model as in Eq (16.19), obtain least-squares estimates of the slope and intercept, and test hypotheses that these parameters are *not* significantly different from zero. Plot the data versus the model fit and comment on the results.

Solution:
If Y is the cranial circumference and x, the finger length, using MINITAB to analyze this data set produces the following results:

Regression Analysis: Cranial Circ(cm) versus Finger Length(cm)

```
The regression equation is
Cranial Circ(cm) = 43.0 + 1.76 Finger Length(cm)
```

Predictor	Coef	SE Coef	T	P
Constant	43.00	17.11	2.51	0.025
Finger Length	1.757	2.126	0.83	0.422

```
S = 2.49655 R-Sq = 4.7% R-Sq(adj) = 0.0%
```

Thus, the regression equation is obtained as

$$\hat{y} = 1.76x + 43.0 \tag{16.81}$$

and the model fit to the data is shown in Fig 16.6. (Again, we defer until the appropriate place, any comment on the terms included in the last line of the MINITAB output.)

It is important to note how, rather than clustering tightly around the regression line, the data shows instead a significant amount of scatter, which, at least visually, calls into question the postulated dependence of cranial circumference on finger length. This question is settled concretely by the computed T statistics for the model parameters and the

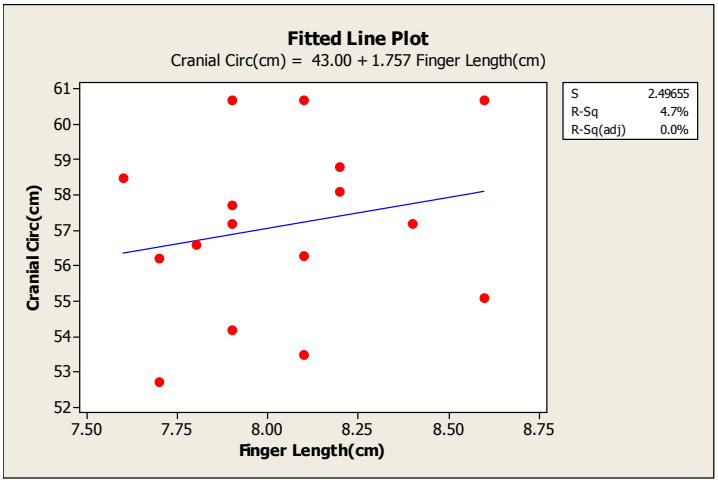

FIGURE 16.6: The fitted straight line to the cranial circumference versus finger length data. Note how the data points are widely scattered around the fitted regression line. (The additional terms included in the graph—S, R-Sq, and R-Sq(adj)—are discussed later.)

associated p-values. The p-value of 0.025 associated with the constant (intercept parameter, θ_0) indicates that we must reject the null hypothesis that $\theta_0 = 0$ in favor of the alternative that the estimated value, 43.0, is significantly different from zero, at the 5% significance level. On the other hand, the corresponding p-value associated with the θ_1, the coefficient of x, the finger length (i.e., the regression line slope), is 0.422, indicating that there is no evidence to reject the null hypothesis. Thus, at the 5% significance level, θ_1 is *not* significantly different from zero and we therefore conclude that there is no discernible relationship between cranial circumference and finger length.

Thus, the implication of the significance of the constant term, and non-significance of the coefficient of the finger length is two-fold: (i) that cranial circumference does not depend on finger length (at least for the 16 individuals in this study), so that the observed variability is purely random, with no systematic component that can be explained by finger length; and consequently, (ii) that the cranial circumference is best characterized for this population of individuals by the mean value (43.0 cm), a value that is significantly different from zero (as one would certainly expect!)

This last example illustrates an important point about regression analysis: one can always fit any postulated model to any given set of data; the real question is: how "useful" is this model? In other words, to what extent is the implied relationship between x and Y representative of the real information contained in the data? These are very important questions that will be

answered systematically in the upcoming sections. For now, we note that at the most basic level, the hypothesis tests discussed here provide an objective assessment of the implied relationship, whether it is "real" or it is merely an artifact of random variability. Anytime we are unable to reject the null hypothesis on the slope parameter, the estimated value, and hence the model itself, are "not significant"; i.e., the real parameter value cannot be distinguished from zero.

16.2.6 Prediction and Prediction Intervals

A model whose parameters are confirmed as "significant" is useful for at least two things:

1. *Estimating Mean Responses*:
   For a given value $x = x^*$, the fitted regression line provides a means of estimating the expected value of the response, $Y|x^*$; i.e.,
   $$E(Y|x^*) = \hat{\theta}_1 x^* + \hat{\theta}_0 \tag{16.82}$$
   This is to be understood as the "least-squares" surrogate of the average of a number of replicate responses obtained when the experiment is repeated for the same fixed value $x = x^*$.

2. *Predicting a New Response:*
   Here, the objective is slightly different: for a given $x = x^*$, we wish to *predict* the response observed from a *single* experiment performed at the specified value. Not surprisingly, the fitted regression line provides the best prediction, $\hat{y}(x^*)$ as
   $$\hat{y}(x^*) = \hat{\theta}_1 x^* + \hat{\theta}_0 \tag{16.83}$$
   which is precisely the same as Eq (16.82). The difference lies not in the value themselves but in the *precision* associated with each value.

When the regression line is used as an estimator of mean (or expected) response, the precision associated with the estimate was given in the form of the variance shown in Eq (16.72), from which we developed the confidence interval around the regression line. When the regression line is used as a prediction of a yet-to-be-observed value $Y(x^*)$, however, the prediction error is given by:

$$E_p = Y(x^*) - \hat{Y}(x^*) \tag{16.84}$$

which, under the normality assumption, possesses the distribution $N(0, \sigma^2_{E_p})$, with the variance obtained from Eq (16.84) as

$$\begin{aligned} \sigma^2_{E_p} &= Var[Y(x^*)] + Var[\hat{Y}(x^*)] \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right] \end{aligned} \tag{16.85}$$

where, we recall, σ^2 as the variance of the random error component, ϵ.

This expression differs from the expression in Eq (16.72) by the presence of the additional term, 1, making $\sigma^2_{E_p} > Var(\hat{\mu}_{Y|x^*})$ always. This mathematical fact is a consequence of the phenomenological fact that the prediction error is a combination of the variability inherent in determining the observed value, $Y(x^*)$, and the regression model error associated with $\hat{Y}(x^*)$, this latter quantity being the only error associated with using the regression model as an estimator of mean response.

We may now use Eq (16.85) to obtain the $(1-\alpha)\times 100\%$ prediction interval for $y(x^*)$ by substituting the data estimate, s_e, for the unknown standard deviation, σ, and from the resulting t-distribution characteristics, i.e.,

$$y(x^*) = \hat{y}(x^*) \pm t_{\alpha/2}(n-2)s_e\sqrt{\left[1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{S_{xx}}\right]} \tag{16.86}$$

As with the confidence intervals around the regression line, these prediction intervals are narrowest when $x^* = \bar{x}$ and widen for values further away from $\bar{x}$, but they are consistently wider than the confidence intervals.

Example 16.4: HIGHWAY GASOLINE MILEAGE AND ENGINE CAPACITY FOR TWO-SEATER AUTOMOBILES
From the data shown in Table 12.5 (Chapter 12) on gasoline mileage for a collection of two-seater cars, postulate a linear two-parameter model for highway mileage (y) as a function of the engine capacity, x; obtain least-squares estimates of the parameters for all the cars, leaving out the Chevrolet Corvette and the Dodge Viper data (these cars were identified Chapter 12 as different from the others in the class because of the material used for their bodies). Show a plot of the fitted regression line, the 95% confidence interval, and the 95% prediction interval.

Solution:
Using MINITAB for this problem produces the following results:

Regression Analysis: MPGHighway versus EngCapacity

```
The regression equation is
MPGHighway = 33.2 - 2.74 EngCapacity
```

Predictor	Coef	SE Coef	T	P
Constant	33.155	1.110	29.88	0.000
EngCapacity	-2.7387	0.2665	-10.28	0.000

```
S = 1.95167 R-Sq = 86.8% R-Sq(adj) = 86.0%
```

Thus, with highway mpg as y, and engine capacity as x, the fitted regression line equation is

$$\hat{y} = -2.74x + 33.2 \tag{16.87}$$

and, since the p-values associated with each parameter are both zero

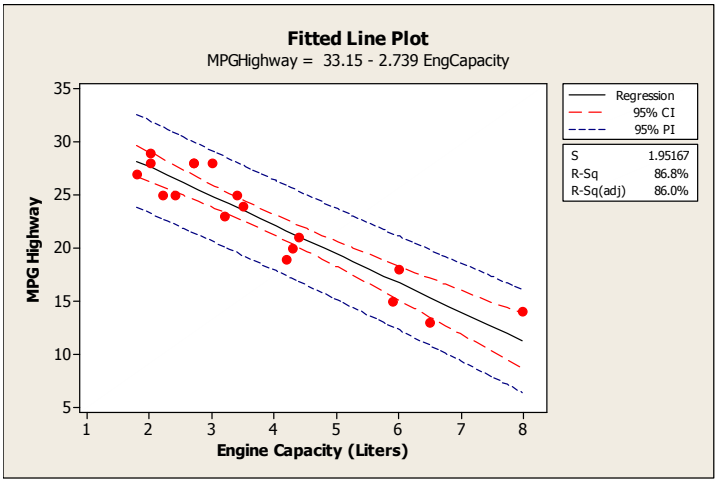

FIGURE 16.7: The fitted straight line to the highway mpg versus engine capacity data of Table 12.5 (leaving out the two "inconsistent" data points) along with the 95% confidence interval (long dashed line) and the 95% prediction interval (short dashed line). (Again, the additional terms—S, R-Sq, and R-Sq(adj)—are discussed later.)

to three decimal places, we conclude that these parameters are "significant." The implication is that for every liter increase in engine capacity, the average two-seater car is expected to lose about 2 and 3/4 miles per gallon on the highway. (As before, we defer until later any comment on the terms in the last line of the MINTAB output.)

The model fit to the data is shown in Fig 16.7 along with the required 95% confidence interval (CI) and the 95% prediction interval (PI). Note how much wider the PI is than the CI at every value of x.

16.2.7 Coefficient of Determination and the F-Test

Beyond hypotheses tests to determine the significance of *individual* estimated parameters, other techniques exist for assessing the *overall* effectiveness of the regression model, based on measures of how much of the total variability in the data has been captured (or explained) by the model.

Orthogonal Decomposition of Variability

The total variability present in the data, represented by $\sum_{i=1}^{n}(y_i - \bar{y})^2$, and defined as S_{yy} in Eq (16.32), may be rearranged as follows, merely by adding and subtracting $\hat{y}_i$:

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}\left[(y_i - \hat{y}_i) - (\bar{y} - \hat{y}_i)\right]^2 \tag{16.88}$$

Upon expanding and simplifying (see Exercise 16.9), one obtains the very important expression:

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ \text{or } S_{yy} &= SS_R + SS_E \end{aligned} \tag{16.89}$$

where we have recalled that the second term on the RHS of the equation is the residual error sum of squares defined in Eq (16.59), and have introduced the term SS_R to represent the regression sum of squares,

$$SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{16.90}$$

a measure of the variability represented in the regression line's estimate of the mean response. The expression in Eq (16.89) represents a decomposition of the total variability in the data into two components: the variability captured by the regression model, SS_R, and what is left in the residual error, SS_E. In fact, we had actually encountered this expression earlier, in Eq (16.61), where what we now refer to as SS_R had earlier been presented as $\hat{\theta}_1 S_{xy}$. If the data vector is represented as $\mathbf{y}$, the corresponding vector of regression model estimates as $\hat{\mathbf{y}}$, and the vector of residual errors between these two as $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, then observe that

$$(\mathbf{y} - \bar{y}) = (\hat{\mathbf{y}} - \bar{y}) + \mathbf{e} \tag{16.91}$$

But from the definition of vector Euclidian norms,

$$||(\mathbf{y} - \bar{y})||^2 = S_{yy} \tag{16.92}$$

$$||(\hat{\mathbf{y}} - \bar{y})||^2 = SS_R \tag{16.93}$$

$$||\mathbf{e}||^2 = SS_E \tag{16.94}$$

with the very important implication that, as a result of the vector representation in Eq (16.91), the expression in Eq (16.89) is an orthogonal decomposition of the data variance vector reminiscent of Pythagoras' Theorem. (If the vector sum in Eq (16.91) holds simultaneously as the corresponding sums of squares expression in Eq (16.89), then the vector $(\hat{\mathbf{y}} - \bar{y})$ must be orthogonal to the vector $\mathbf{e}$.)

Eq (16.89) is in fact known as the *analysis of variance (ANOVA) identity*; and it plays a central role in statistical inference that transcends the restricted role observed here in regression analysis. We shall have cause to revisit this subject in our discussion of the design of experiments in upcoming chapters. For now, we use it to assess the effectiveness of the overall regression model (as a single entity purporting to represent the information contained in the data), first in the form of the coefficient of determination, and later as the basis for an F-test of significance. This latter exercise will constitute a preview of an upcoming, more general discussion of ANOVA.

R^2, The Coefficient of Determination

Let us now consider the ratio defined as:

$$R^2 = \frac{SS_R}{S_{yy}} \tag{16.95}$$

which represents the proportion of the total data variability (around the mean $\bar{y}$) that has been captured by the regression model; its complement,

$$1 - R^2 = SS_E/S_{yy} \tag{16.96}$$

is the portion left unexplained by the regression model. Observe that $0 \leq R^2 \leq 1$, and that if a model adequately captures the relevant information contained in a data set, what will be left unexplained as random variation should be comparatively small, so that the R^2 value will be close to 1. Conversely, a value close to zero indicates a model that is inadequate in capturing the important variability present in the data. R^2 is therefore known as the coefficient of determination; it is a direct measure of the quality of fit provided by the regression model.

Although not directly relevant yet at this point (where we are still discussing the classical two-parameter model), it is possible to improve a model fit by introducing additional parameters. Under such circumstances, the improvement in R^2 may come at the expense of over-fitting (as discussed more fully later). A somewhat more judicious assessment of model adequacy requires adjusting the value of R^2 to reflect the number of parameters that have been used by the model to capture the variability.

By recasting the expression in Eq (16.95) in the equivalent form:

$$R^2 = 1 - \frac{SS_E}{S_{yy}} \tag{16.97}$$

rather than base the metric on the indicated absolute sums of squares, consider using the mean sums of squares instead. In other words, instead of the total residual error sum of squares, SS_E, we employ instead the mean residual error sum of squares, $SS_E/(n-p)$, where p is the number of parameters in the model and n is the total number of experimental data points; also instead of the total data sum of squares, S_{yy}, we employ instead the data variance $S_{yy}/(n-1)$. The resulting quantity, known as R^2_{adj}, and defined as:

$$R^2_{adj} = 1 - \frac{SS_E/(n-p)}{S_{yy}/(n-1)} \tag{16.98}$$

is similar to the coefficient of determination, R^2, but it is adjusted for the the number of parameters contained in the model. It penalizes models that achieve decent values of R^2 via the use of an excessive number of parameters. Relatively high values of R^2 and R^2_{adj} that are also comparable in magnitude

indicate a model that is quite adequate: the variability in the data has been captured adequately without using an excessive number of parameters.

All software packages that carry out regression analysis routinely compute R^2 and R^2_{adj}, sometimes presented not as fractions (as indicated above), but multiplied by 100%. In fact, all the examples and fitted regression line plots encountered thus far in this chapter have shown these values (in percentage form) but we had to defer commenting on them until now. We are only now in a position for such a discussion.

In Figs 16.4, 16.5, 16.6, and 16.7, the value shown for S is the square root of the mean residual sum of squares, i.e., $\sqrt{SS_E/(n-2)}$, an estimate of the unknown data standard deviation, σ; this is accompanied by values for R^2 and R^2_{adj}. Thus, in the density versus ethanol weight percent regression model (Fig 16.4), both R^2 and R^2_{adj} are reported as 99.8%, indicating a model that appears to have explained virtually all the variability in the data, with very little left by way of the residual error (as indicated by the very small value of S). The exact opposite is the case with the cranial circumference versus finger length regression model: R^2 is an incredibly low 4.7% and the R^2_{adj} vanishes entirely (a "perfect" 0.00%), indicating that (a) the model has explained very little of the variability in the data, and (b) when penalized for the parameters employed in achieving even the less that 5% variability captured, the inadequacy of the model is seen to be total. The residual data standard deviation, S, is almost 2.5. With the highway gas mileage versus engine capacity regression model, the R^2 value is reasonably high at 86.8%, with an adjusted value of 86% that is essentially unchanged; the residual standard of $S = 1.95$ is also reasonable. The indication is that while there is still some unexplained variability left, the regression model captures a significant amount of the variability in the data and provides a reasonable mathematical explanation of the information contained in the data.

F-Test for Significance of Regression

Let us return to the ANOVA expression in Eq (16.89):

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ \text{i.e., } S_{yy} &= SS_R + SS_E \end{aligned}$$

and note the following:

1. The total sum of squares, $S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2$, has $(n-1)$ degrees of freedom; the error sum of squares, $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, has $(n-2)$ degrees of freedom, and the regression sum of squares, $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$, has 1 degree of freedom.

2. One informal way to confirm this fact is as follows: (i) of the n independent units of information in the raw data, $y_i; i = 1, 2, \ldots, n$, 1 "degree

of freedom" is "tied up" in obtaining the average, $\bar{y}$, so that $(y_i - \bar{y})$ will have $(n-1)$ degrees of freedom left (i.e., there are now only $(n-1)$ independent quantities in $(y_i - \bar{y})$); (ii) similarly, 2 "degrees of freedom" are "tied up" in obtaining the response estimate $\hat{y}$ (via the two parameters, $\hat{\theta_0}$ and $\hat{\theta_1}$), so that $(y_i - \hat{y}_i)$ has $(n-2)$ degrees of freedom left; and finally (iii) while $\hat{y}_i$ "ties up" 2 degrees of freedom, $\bar{y}$ "ties up" 1, so that $(\hat{y}_i - \bar{y})$ has 1 degree of freedom left.

3. The implication is therefore that, in addition to representing a decomposition of variability, since it is also true that

$$(n-1) = 1 + (n-2) \tag{16.99}$$

Eq (16.89) also represents a concurrent decomposition of the degrees of freedom associated with each sum of squares.

Finally, from the following results (given without proof; e.g., Eq (16.62)),

$$\begin{aligned} E(SS_E) &= (n-2)\sigma^2 \\ E(SS_R) &= \hat{\theta}_1^2 S_{xx} + \sigma^2 \end{aligned} \tag{16.100}$$

we arrive at the following conclusions: Under the null hypothesis $H_0 : \theta_1 = 0$, these two equations suggest $SS_E/(n-2)$ and $SS_R/1$ (respectively the error mean square, MS_E, and the regression mean square, MS_R) as two separate and distinct estimators of σ^2. Furthermore, under the normality assumption for y_i, then the statistic

$$F = \frac{SS_R/1}{SS_E/(n-2)} \tag{16.101}$$

will possess an $F(\nu_1, \nu_2)$ distribution, with $\nu_1 = 1$, and $\nu_2 = (n-2)$, if H_0 is true that $\theta_1 = 0$. However, if H_0 is not true, then the numerator in Eq (16.101) will be inflated by the term $\hat{\theta}_1^2 S_{xx}$ as indicated in Eq (16.100). Hence, at the significance level of α, we reject H_0 (that the regression as a whole is *not* significant), when the actual computed statistic

$$f > f_\alpha(\nu_1, \nu_2) \tag{16.102}$$

where $f_\alpha(\nu_1, \nu_2)$ is the usual $F(\nu_1, \nu_2)$-distribution variate with upper tail area α. Equivalently, one computes the p-value associated with the computed f, as

$$p = P(F \geq f) \tag{16.103}$$

and reject or fail to reject the null hypothesis on the basis of the actual p-value.

These results are typically presented in what is referred to as an ANOVA table as shown in Table 16.5. They are used to carry out F-tests for the significance of the entire regression model as a single entity; if the resulting p-value is low, we reject the null hypothesis and conclude that the regression is significant, i.e., the relationship implied by the regression model is meaningful.

TABLE 16.5: ANOVA table for testing significance of regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F	p
Regression	SS_R	1	MS_R	$\frac{MS_R}{MS_E}$	
Error	SS_E	$(n-2)$	MS_E		
Total	S_{yy}	$(n-1)$			

Alternatively, if the p-value exceeds a pre-specified threshold (say, 0.05), we fail to reject the null hypothesis and conclude that the regression model is not significant—that the implied relationship is purely random.

All computer programs that perform regression analysis produce such ANOVA tables. For example, the MINITAB output for Example 16.4 above (involving the regression model relating engine capacity to the highway mpg rating) includes the following ANOVA table.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	402.17	402.17	105.58	0.000
Residual Error	16	60.94	3.81		
Total	17	463.11			

The indicated p-value of 0.000 implies that we must reject the null hypothesis and conclude that the regression model is "significant."

On the other hand, the ANOVA table produced by MINITAB for the cranial circumference versus finger length regression problem of Example 16.3 is as shown below:

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	4.259	4.259	0.68	0.422
Residual Error	14	87.259	6.233		
Total	15	91.517			

In this case, the p-value associated with the F-test is so high (0.422) that we reject the null hypothesis and conclude that the regression is *not* significant.

Of course, these conclusions agree perfectly with our earlier conclusions concerning each of these problems.

In general, we tend to de-emphasize these ANOVA-based F-tests for significance of the regression. This is for the simple reason that they are coarse tests of the *overall* regression model, adding little or nothing to the individual t-tests presented earlier for each parameter. These individual parameter tests are preferred because they are finer-grained.

From this point on, we will no longer refer to these ANOVA tests of significance for regression. Nevertheless, these same concepts take center stage in Chapter 19 where they are central to analysis of designed experiments.

16.2.8 Relation to the Correlation Coefficient

In Chapter 5, we defined the correlation coefficient between two jointly distributed random variables X and Y as:

$$\rho = \frac{E[XY]}{\sigma_X \sigma_Y} \tag{16.104}$$

The sample version, obtained from data, known as the Pearson product moment correlation coefficient, is:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{16.105}$$

If we now recall the expressions in Eqs (16.31)–(16.33), we immediately obtain, in the context of regression analysis:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}} \tag{16.106}$$

And now, in terms of the slope parameter estimate, we obtain from Eq (16.37), first that

$$r = \hat{\theta}_1 \frac{\sqrt{S_{xx}}}{\sqrt{S_{yy}}} \tag{16.107}$$

an expression we shall return to shortly. For now, let us return to the expression for R^2 in Eq (16.97); if we introduce Eq (16.61) for SS_E, we obtain

$$R^2 = 1 - \frac{S_{yy} - \hat{\theta}_1 S_{xy}}{S_{yy}} = \hat{\theta}_1 \frac{S_{xy}}{S_{yy}} \tag{16.108}$$

Upon introducing Eq (16.37) for $\hat{\theta}_1$, we obtain the result that:

$$R^2 = \frac{(S_{xy})^2}{S_{xx} S_{yy}} \tag{16.109}$$

which, when compared with Eq (16.106) establishes the important result that R^2, the coefficient of determination, is the square of the sample correlation coefficient, r; i.e.,

$$R^2 = r^2 \tag{16.110}$$

16.2.9 Mean-Centered Model

As obtained previously, the estimated observation from the regression model is given by

$$\hat{y} = \hat{\theta}_0 + \hat{\theta}_1 x \tag{16.111}$$

and, from a rearrangement of Eq (16.38) used to estimate $\hat{\theta}_0$, we obtain

$$\bar{y} = \hat{\theta}_0 + \hat{\theta}_1 \bar{x} \tag{16.112}$$

If we now subtract the latter equation from the former, we obtain:

$$(\hat{y} - \bar{y}) = \hat{\theta}_1 (x - \bar{x}) \tag{16.113}$$

which is a "mean-centered" version of the regression model.

If we now rearrange Eq (16.107) to express $\hat{\theta}_1$ in terms of r and introduce it into Eq (16.113), we obtain:

$$(\hat{y} - \bar{y}) = r \left(\frac{s_y}{s_x} \right) (x - \bar{x}) \tag{16.114}$$

where $s_y = \sqrt{S_{yy}}/(n-1)$ and $s_x = \sqrt{S_{xx}}/(n-1)$ are respectively sample estimates of the data standard deviation for y and for x. Alternatively, Eq (16.114) could equivalently be written as

$$(\hat{y} - \bar{y}) = \sqrt{R^2 \left(\frac{S_{yy}}{S_{xx}} \right)} (x - \bar{x}) \tag{16.115}$$

This equation provides the clearest indication of the impact of R^2 on how "strongly" the mean-centered value of the predictor, x, is connected by the model to—and hence can be used to estimate—the mean-centered response. Observe that in the density and ethanol weight percent example, with $R^2 = 0.998$, the connection between the predictor and response estimate is particularly strong; with the cranial circumference–finger length example, the connection is extremely weak, and the best estimate of the response (cranial circumference) for any value of the predictor (finger length) is essentially the mean value, $\bar{y}$.

16.2.10 Residual Analysis

While the statistical significance of the estimated parameters gives us some information about the usefulness of a model, and while the R^2 and R^2_{adj} values provide a measure of how much of the data's variability has been captured by the overall model, how well the model represents the data is most directly determined from the difference between the actual observation, y_i, and the model estimate, $\hat{y}_i$, i.e., $e_i = y_i - \hat{y}_i, i = 1, 2, \ldots, n$. These quantities, identified earlier as residual errors (or simply as "residuals"), provide n different samples

of how closely the model matches the data. If the model's representation of the data is adequate, the residuals should be nothing but purely random variation. Any departure from pure random variation in the residuals is an indication of some form or the other of model inadequacy. Residual analysis therefore allows us to do the following:

1. Check model assumptions, specifically that $\epsilon \sim N(0, \sigma^2)$, with σ^2 constant;
2. Identify any "left over" systematic structural discrepancies between the model and data; and
3. Identify which data points might be inconsistent with the others in the set.

By formulation, the least-squares estimation technique *always* produces a model for which $\sum_{i=1}^{n} e_i = 0$, making the desired zero mean characteristics of the model error sequence a non-issue. Residual analysis is therefore concerned mostly with the following activities:

1. Formal and informal tests of normality of e_i;
2. Graphical/visual evaluation of residual patterns;
3. Graphical/visual and numerical evaluation of individual residual magnitude.

Other than formal normality tests (which involve techniques discussed in the next chapter) residual analysis is more of an art involving various graphical plots. When there is sufficient data, histograms of the residuals provide great visual clues regarding normality quite apart from what formal tests show. In many cases, however, available data is usually modest. Regardless of the data size, plots of the residuals themselves versus the fitted value, $\hat{y}$, or versus data order, or versus x, are not only capable of indicating model adequacy, they also provide clues about the nature of the implied model inadequacy.

It is often recommended that residual plots be based not on the residual themselves, but on the standardized residual,

$$e_i^* = \frac{e_i}{s_e} \tag{16.116}$$

where, s_e, as we recall, is the estimate of σ, the data standard deviation. This is because if the residuals are truly normally distributed, then $-2 < e_i^* < 2$ for approximately 95% of the standardized residuals; and some 99% should lie between -3 and 3. As a general rule-of-thumb, therefore, $|e_i^*| > 3$ indicates a value that is considered a potential "outlier" because it is inconsistent with the others.

When a model appears inadequate, the recommendation is to look for clues within the residuals for what to do next. The following examples illustrate residual analysis for practical problems in engineering.

TABLE 16.6: Thermal conductivity measurements at various temperatures for a metal

k (W/m-$^\circ C$)	Temperature ($^\circ C$)
93.228	100
92.563	150
99.409	200
101.590	250
111.535	300
115.874	350
119.390	400
126.615	450

Example 16.5: TEMPERATURE DEPENDENCE OF THERMAL CONDUCTIVITY
To characterize how the thermal conductivity, k (W/m-$^\circ C$), of a metal varies with temperature, 8 independent experiments were performed at the temperatures, $T^\circ C$, shown in Table 16.6 along with the measured thermal conductivities. A two-parameter model as in Eq 16.19 has been postulated for the relationship between k and T. Obtain a least-squares estimate of the parameters and evaluate the model fit to the data.

Solution:
We use MINITAB and obtain the following results:

Regression Analysis: k versus Temperature

```
The regression equation is
k = 79.6 + 0.102 Temperature
```

Predictor	Coef	SE Coef	T	P
Constant	79.555	2.192	36.29	0.000
Temperature	0.101710	0.007359	13.82	0.000

```
S = 2.38470 R-Sq = 97.0% R-Sq(adj) = 96.4%
```

Therefore, as before, representing the thermal conductivity as y, and temperature as x, the fitted regression line equation is

$$\hat{y} = 0.102x + 79.6 \tag{16.117}$$

The p-values associated with each parameter is zero, implying that both parameters are significantly different from zero. The estimate of the data standard deviation is shown as S; and the R^2 and R^2_{adj} values indicate that the model captures a reasonable amount of the variability in the data.

The actual model fit to the data is shown in the top panel of Fig 16.8 while the standardized residuals versus fitted value, $\hat{y}_i$, is shown in the bottom panel. With only 8 data points, there are not enough residuals for a histogram plot. Nevertheless, upon visual examination of the residual plots, there appears to be no discernible pattern, nor is

there any reason to believe that the residuals are anything but purely random. Note that no standardized residual value exceeds ± 2.

The model is therefore considered to provide a reasonable representation of how the thermal conductivity of this metal varies with temperature.

The next example illustrates a practical circumstance where the residuals not only expose the inadequacy of a linear regression model, but also provide clues concerning how to rectify the inadequacy.

Example 16.6: BOILING POINT OF HYDROCARBONS
It has been proposed to represent with a linear two-parameter model, the relationship between the number of carbon atoms in the hydrocarbon compounds listed in Table 16.1 and the respective boiling points. Evaluate a least-squares fit of this model to the data.

Solution:
Using MINITAB produces the following results for this problem:

Regression Analysis: Boiling Point versus n

```
The regression equation is
Boiling Point = - 172.8 + 39.45 n
```

Predictor	Coef	SE Coef	T	P
Constant	-172.79	13.26	-13.03	0.000
n	39.452	2.625	15.03	0.000

```
S = 17.0142 R-Sq = 97.4 % R-Sq(adj) = 97.0%
```

Therefore, as before, the fitted regression line equation is

$$\hat{y} = 39.45x - 172.8 \tag{16.118}$$

with the hydrocarbon compound boiling point as y, and the number of carbon atoms it contains as x.

We notice, once again, that for this model, the parameter values are all significantly different from zero because the p-values are zero in each case; furthermore, the R^2 and R^2_{adj} values are quite good. The error standard deviation is obtained as $S = 17.0142$. By themselves, nothing in these results seem out of place; one might even be tempted by the excellent R^2 and R^2_{adj} values to declare that this is a very good model. However, the model fit to the data, shown in the top panel of Fig 16.9, tells a different story; and the normalized residuals versus fitted value, $\hat{y}_i$, shown in the bottom panel, is particularly revealing. The model fit shows a straight line model that very consistently overestimates BP at the extremes and underestimates it in the middle. The standardized residual versus model fit quantifies this under- and over-estimation and shows a clear "left over" quadratic structure.

The implication clearly is that while approximately 97% of the relationship between n, the number of carbon atoms, and the hydrocarbon

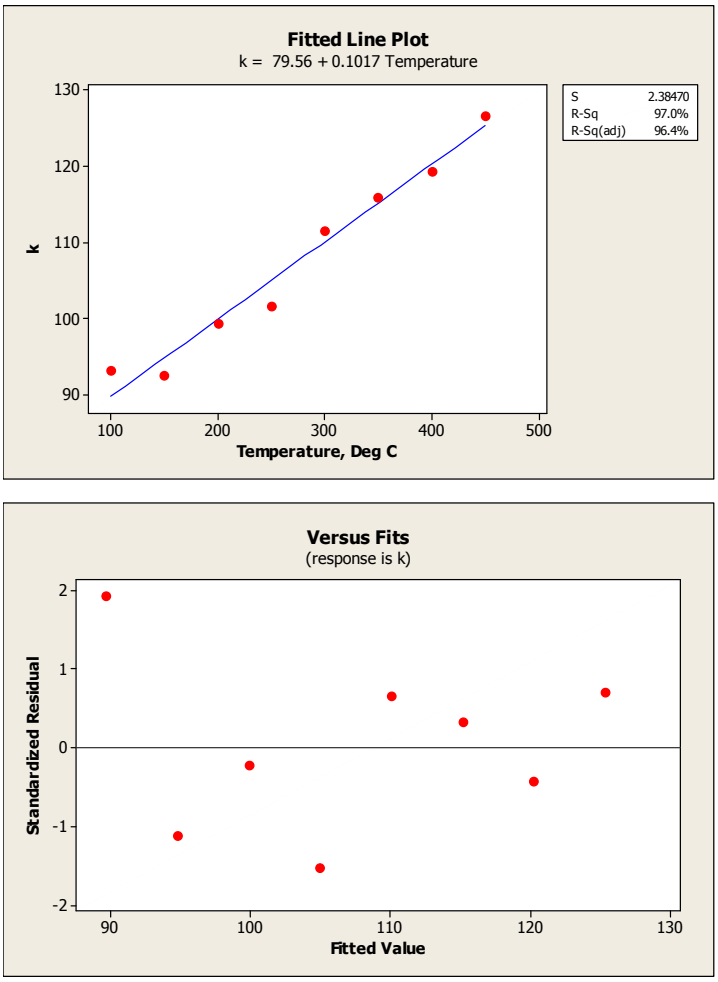

FIGURE 16.8: Modeling the temperature dependence of thermal conductivity. Top: fitted straight line to the thermal conductivity (k) versus temperature ($T^{\circ}C$) data in Table 16.6; Bottom: standardized residuals versus fitted value, $\hat{y}_i$.

BP has been captured by a linear relationship, there remains an unexplained, possibly quadratic, component that is clearly discernible. The suggestion: consider a revised model of the type

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 \tag{16.119}$$

Such a model is a bit more complicated, but the residual structure seems to suggest that the additional term is warranted. How to obtain a model of this kind is discussed shortly.

We revisit the problem illustrated in this example after completing a discussion of more complicated regression models in the upcoming sections.

16.3 "Intrinsically" Linear Regression

16.3.1 Linearity in Regression Models

In estimating the vector of parameters $\boldsymbol{\theta}$ contained in the general regression model,

$$Y = g(x;\boldsymbol{\theta}) + \epsilon$$

it is important to clarify what the term "linear" in linear regression refers to. While it is true that the model:

$$Y = \theta_0 + \theta_1 x + \epsilon$$

is a linear equation because it represents a straight line relationship between Y and x, what is actually of relevance in regression analysis is that this functional form is linear with respect to the unknown parameter vector $\boldsymbol{\theta} = (\theta_0, \theta_1)$. It must be kept in mind in regression, that x is known and given; the parameters $\boldsymbol{\theta}$ are the unknowns to be determined by the regression exercise.

Thus, what is of importance in determining whether a regression problem is linear or not is the functional form of $g(x;\boldsymbol{\theta})$ with respect to the vector of parameters $\boldsymbol{\theta}$, *not with respect to* x. For example, if the regression model is given as

$$Y = \theta_1 x^n + \epsilon, \tag{16.120}$$

clearly this is a nonlinear function of x; however, so long as n is known, for any given value of x, x^n is also known, say $x^n = z$; this equation is therefore exactly equivalent to

$$Y = \theta_1 z + \epsilon, \tag{16.121}$$

which is clearly linear. Thus, even though nonlinear in x, Eq (16.120) nevertheless represents a linear regression problem because the model equation is

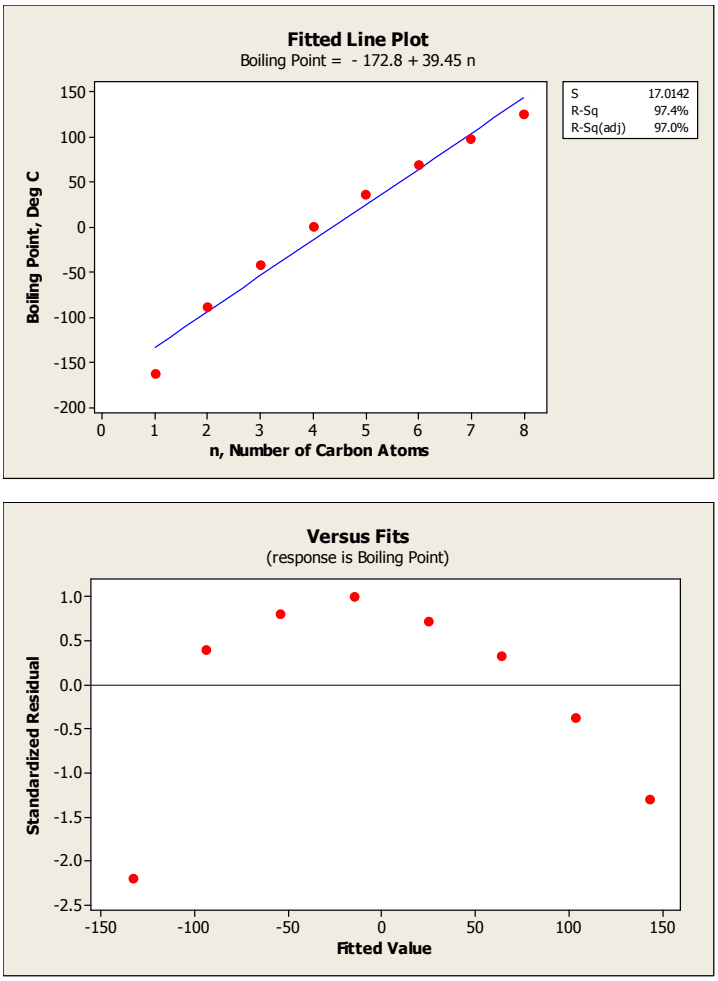

FIGURE 16.9: Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound. Top: fitted straight line of BP versus n, the number of carbon atoms; Bottom: standardized residuals versus fitted value $\hat{y}_i$. Notice the distinctive quadratic structure "left over" in the residuals exposing the linear model's over-estimation at the extremes and the under-estimation in the middle.

linear in the unknown parameter θ. On the other hand, the model representing how the concentration $C(t)$ of a reactant undergoing a first order kinetic reaction in a batch reactor changes with time,

$$C(t) = \theta_0 e^{-\theta_1 t} \tag{16.122}$$

with time, t, as the independent variable, along with θ_0 and θ_1 respectively as the unknown initial concentration and kinetic reaction rate constant, represents a truly *nonlinear* regression model. This is because one of the unknown parameters, θ_1, enters the model nonlinearly; the model is linear in the other parameter θ_0.

As far as regression is concerned, therefore, whether the problem at hand is linear or nonlinear, depends on whether the parameter sensitivity function,

$$\mathcal{S}_{\theta_i} = \frac{\partial g}{\partial \theta_i} \tag{16.123}$$

is a function of θ_i or not. For linear regression problems S_{θ_i} is independent of θ_i for all i; the defining characteristics of nonlinear regression is that $\mathcal{S}_{\theta_i}$ depends on θ_i for at least one i.

Example 16.7: LINEAR VERSUS NONLINEAR REGRESSION PROBLEMS

Which of the following three models presents a linear or nonlinear regression problem in estimating the indicated unknown parameters, θ_i?

$$(1)\ Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \epsilon \tag{16.124}$$

$$(2)\ Y = \theta_1 {\theta_2}^x + \epsilon \tag{16.125}$$

$$(3)\ Y = \theta_1 e^{x_1} + \theta_2 \ln x_2 + \frac{\theta_3}{x_3} + \epsilon \tag{16.126}$$

Solution:

Model (1) presents a linear regression problem because each of the sensitivities, $\mathcal{S}_{\theta_0} = 1$; $\mathcal{S}_{\theta_1} = x$; $\mathcal{S}_{\theta_2} = x^2$; and $\mathcal{S}_{\theta_3} = x^3$, is free of the unknown parameter on which it is based, i.e., $\mathcal{S}_{\theta_i}$ is not a function of θ_i for $i = 0, 1, 2, 3$. In fact, all the sensitivities are entirely free of all parameters.

Model (2) on the other hand presents a nonlinear regression problem: $\mathcal{S}_{\theta_1} = {\theta_2}^x$ *does not* depend on θ_1, but

$$\mathcal{S}_{\theta_2} = \theta_1 x {\theta_2}^{x-1} \tag{16.127}$$

depends on θ_2. Thus, while this model is linear in θ_1 (because the sensitivity to θ_1 does not depend on θ_1), it is nonlinear in θ_2; therefore, it presents a nonlinear regression problem.

Model (3) presents a linear regression problem: $\mathcal{S}_{\theta_1} = e^{x_1}$; $\mathcal{S}_{\theta_2} = \ln x_2$, are both entirely free of unknown parameters.

16.3.2 Variable Transformations

A restricted class of truly nonlinear regression models may be converted to linear models via appropriate variable transformations; linear regression analysis can then be carried out in terms of the transformed variables. For example, observe that even though the reactant concentration model in Eq (16.122) has been identified as nonlinear in the parameters, a logarithmic transformation results in:

$$\ln C(t) = \ln \theta_0 - \theta_1 t \tag{16.128}$$

In this case, observe that if we now let $Y = \ln C(t)$, and let $\theta_0^* = \ln \theta_0$, then Eq (16.128) represents a linear regression model.

Such cases abound in chemical engineering. For example, the equilibrium vapor mole fraction, y, as a function of liquid mole fraction, x, of a compound with relative volatility α is given by the expression:

$$y = \frac{\alpha x}{1 + (\alpha - 1)x} \tag{16.129}$$

It is an easy exercise to show that by inverting this equation, we obtain:

$$\frac{1}{y} = \theta \frac{1}{x} + (1 - \theta) \tag{16.130}$$

so that $1/y$ versus $1/x$ produces a linear regression problem.

Such models are said to be "intrinsically" linear because while nonlinear in their original variables, they are linear in a different set of transformed variables; the task is to find the required transformation. Nevertheless, the careful reader would have noticed something missing from these model equations: we have carefully avoided introducing the error terms, ϵ. This is for the simple reason that in virtually all cases, if the error term is additive, then even the obvious transformations are no longer possible. For example, if each actual concentration measurement, $C(t_i)$, observed at time t_i has associated with it the additive error term ϵ_i, then Eq (16.122) must be rewritten as

$$C(t_i) = \theta_0 e^{-\theta_1 t_i} + \epsilon_i \tag{16.131}$$

and the logarithmic transformation is no longer possible.

Under such circumstances, most "practitioners" will suspend the addition of the error term until after the function has been appropriately transformed; i.e., instead of writing the model as in Eq (16.131), it would be written as:

$$\ln C(t_i) = \ln \theta_0 - \theta_1 t_i + \epsilon_i \tag{16.132}$$

But this implies that the error is multiplicative in the original variable. It is important, before taking such a step, to take time to consider whether such a multiplicative error structure is reasonable or not.

Thus, in employing transformations to deal with these so-called intrinsically linear models, the most important issue lies in determining the proper error structure. Such transformations should be used with care; alternatively, the parameter estimates obtained from such an exercise should be considered as approximations that may require further refinement by using more advanced nonlinear regression techniques. Notwithstanding, many engineering problems involving models of this kind have benefited from the sort of linearizing transformations discussed here.

16.4 Multiple Linear Regression

In many cases, the response variable Y depends on several independent variables, $x_1, x_2, \ldots, x_m$. Under these circumstances, the simplest possible regression model is:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m + \epsilon \tag{16.133}$$

with m independent predictor variables, $x_i; i = 1, 2, \ldots, m$, and $m+1$ unknown parameters, $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_m)$, the "regression coefficients." Eq (16.133) represents a multiple linear regression model. An example is when Y, the conversion obtained from a catalytic process depends on the temperature, x_1, and pressure, x_2, at which the reactor is operated, according to:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \epsilon \tag{16.134}$$

Just as the expected value of the response in Eq (16.19) represents a straight line, the expected value in Eq (16.133) represents an m-dimensional hyperplane.

However, there is no reason to restrict the model to the form in Eq (16.133). With more than one independent variable, it is possible for the response variable to depend on higher order powers of, as well as interactions between, some of the variables; for example, a better representation of the relationship between yield and temperature and pressure might be:

$$Y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_{11} x_1^2 + \theta_{22} x_2^2 + \theta_{12} x_1 x_2 + \epsilon \tag{16.135}$$

Such models as these are often specifically referred to as "response surface" models and we shall have cause to revisit them later on in upcoming chapters. For now, we note that in general, most multiple regression models are, for the most part, justified as approximations to the more general expression,

$$Y = g(x_1, x_2, \ldots, x_m; \boldsymbol{\theta}) + \epsilon \tag{16.136}$$

where neither the true form of $g(x_1, x_2, \ldots, x_m; \boldsymbol{\theta})$, nor the parameters, $\boldsymbol{\theta}$

are known. If $g(.)$ is analytic, then by taking a Taylor series expansion and truncating after a pre-specified number of terms, the result will be a multiple regression model. The multiple regression function is therefore often justified as a Taylor series approximation of an unknown, and perhaps more complex, function. For example, Eq (16.135) arises when the Taylor expansion only goes up to the second order.

Regardless of what form the regression model takes, keep in mind that so long as the values of the independent variables are known, all such models can always be recast in the form shown in Eq (16.133). For example, even though there are two actual predictor variables x_1 and x_2 in the model in Eq (16.135), if we define "new" variables $x_3 = x_1^2; x_4 = x_2^2; x_5 = x_1 x_2$, then Eq (16.135) immediately becomes like Eq (16.133) with $m = 5$. Thus, it is without loss of generality that we consider Eq (16.133) as the general multiple regression model.

Observe that the model in Eq (16.133) is a direct generalization of the two-parameter model in Eq (16.19); we should therefore expect the procedure for estimating the increased number of parameters to be similar to the procedure discussed earlier. While this is true in principle, the analysis is made more tractable by using matrix methods, as we now show.

16.4.1 General Least Squares

Obtaining estimates of the m parameters, $\theta_i, i = 1, 2, \ldots, m$, from data $(y_i; x_{1i}, x_{2i}, \ldots, x_{mi})$ via the least-squares technique involves minimizing the sum-of-squares function,

$$S(\boldsymbol{\theta}) = \sum_{i=1}^{n} [y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \cdots + \theta_m x_{mi})]^2 \tag{16.137}$$

The technique calls for taking derivatives with respect to each parameter, setting the derivative to zero and solving the resultant equations for the unknown parameters, i.e.,

$$\frac{\partial S}{\partial \theta_0} = -2 \sum_{i=1}^{n} [y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \cdots + \theta_m x_{mi})] = 0 \tag{16.138}$$

and for $1 \leq j \leq m$,

$$\frac{\partial S}{\partial \theta_j} = -2 \sum_{i=1}^{n} x_{ji} [y_i - (\theta_0 + \theta_1 x_{1i} + \theta_2 x_{2i} + \cdots + \theta_m x_{mi})] = 0 \tag{16.139}$$

These expressions rearrange to give the general linear regression normal equations:

$$\begin{aligned}
\sum_{i=1}^{n} y_i &= \theta_0 n + \theta_1 \sum_{i=1}^{n} x_{1i} + \theta_2 \sum_{i=1}^{n} x_{2i} + \cdots + \theta_m \sum_{i=1}^{n} x_{mi} \\
\sum_{i=1}^{n} y_i x_{1i} &= \theta_0 \sum_{i=1}^{n} x_{1i} + \theta_1 \sum_{i=1}^{n} x_{1i}^2 + \theta_2 \sum_{i=1}^{n} x_{2i}x_{1i} + \cdots + \theta_m \sum_{i=1}^{n} x_{mi}x_{1i} \\
\sum_{i=1}^{n} y_i x_{ji} &= \theta_0 \sum_{i=1}^{n} x_{ji} + \theta_1 \sum_{i=1}^{n} x_{1i}x_{ji} + \theta_2 \sum_{i=1}^{n} x_{2i}x_{ji} + \cdots + \theta_m \sum_{i=1}^{n} x_{mi}x_{ji}
\end{aligned}$$

$m+1$ linear equations to be solved simultaneously to produce the least-squares estimates for the $m+1$ unknown parameters. Even with a modest number of parameters, such problems are best solved using matrices.

16.4.2 Matrix Methods

For specific data sets, $(y_i; x_{1i}, x_{2i}, \ldots, x_{mi}); i = 1, 2, \ldots, n$, the multiple regression model equation in Eq (16.133) may be written as,

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{x2} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_m \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix} \tag{16.140}$$

which may be consolidated into the explicit matrix form,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{16.141}$$

where $\mathbf{y}$ is the n-dimensional vector of response observations, with $\mathbf{X}$ as the $n \times m$ matrix of values of the predictor variables used to generate the n observed responses; $\boldsymbol{\theta}$ is the m-dimensional vector of unknown parameters, and $\boldsymbol{\epsilon}$ is the n-dimensional vector of random errors associated with the response observations. Obtaining the least-squares estimate of the parameter vector involves the same principle of minimizing the sum of squares, which, this time is given by

$$S(\boldsymbol{\theta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\theta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) \tag{16.142}$$

where the superscript, T, represents the vector or matrix transpose. Taking derivatives with respect to the parameter vector $\boldsymbol{\theta}$ and setting the result to zero yields:

$$\frac{\partial S}{\partial \boldsymbol{\theta}} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\theta}) = 0 \tag{16.143}$$

resulting finally in the matrix form of the normal equations:

$$\mathbf{X}^T\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\theta} \tag{16.144}$$

(compare with series of equations shown earlier). This matrix equation is easily solved for the unknown parameter vector to produce the least-squares solution:

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{16.145}$$

Properties of the Estimates

To characterize the estimates, we begin by introducing Eq (16.141) into Eq (16.145) for $\mathbf{y}$ to obtain:

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\boldsymbol{\theta}+\boldsymbol{\epsilon}) \\ &= \boldsymbol{\theta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\end{aligned} \tag{16.146}$$

We may now use this expression to obtain the mean and variance of these estimates as follows. First, by taking expectations, we obtain:

$$\begin{aligned}E(\hat{\boldsymbol{\theta}}) &= \boldsymbol{\theta} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E(\boldsymbol{\epsilon}) \\ &= \boldsymbol{\theta}\end{aligned} \tag{16.147}$$

because $\mathbf{X}$ is known and $E(\boldsymbol{\epsilon}) = 0$. Thus, the least-squares estimate $\hat{\boldsymbol{\theta}}$ as given in Eq (16.145) is unbiased for $\boldsymbol{\theta}$. As for the co-variance of the estimates, first, by definition, $E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T) = \boldsymbol{\Sigma}$ is the random error covariance matrix; then from the assumption that each ϵ_i is independent, and identically distributed, with the same variance, σ^2, we have that

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I} \tag{16.148}$$

where σ^2 is the variance associated with each random error, and $\mathbf{I}$ is the identity matrix. As a result,

$$Var(\hat{\boldsymbol{\theta}}) = E\left[(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})^T\right] \tag{16.149}$$

which, from Eq (16.146) becomes

$$\begin{aligned}Var(\hat{\boldsymbol{\theta}}) &= E\left[\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right] \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T)\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2\end{aligned} \tag{16.150}$$

Thus, the covariance matrix of the estimates, $\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}}$ is given by

$$\mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma^2 \tag{16.151}$$

As usual, σ^2 must be estimated from data. With $\hat{\mathbf{y}}$, the model estimate of the response data vector $\mathbf{y}$ now given by:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\theta}} \tag{16.152}$$

the residual error vector, $\mathbf{e}$, is therefore defined as:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \tag{16.153}$$

so that the residual error sum-of-squares will be given by

$$SS_E = \mathbf{e}^T\mathbf{e} \tag{16.154}$$

It can be shown that with $p = m + 1$ as the number of estimated parameters, the mean error sum-of-squares

$$s_e = \frac{\mathbf{e}^T\mathbf{e}}{n-p} \tag{16.155}$$

is an unbiased estimate of σ.

Thus, following the typical normality assumption on the random error component of the regression model, we now conclude that the least-squares estimate vector, $\hat{\boldsymbol{\theta}}$, has a multivariate normal distribution, $MVN(\boldsymbol{\theta}, \mathbf{\Sigma}_{\hat{\boldsymbol{\theta}}})$, with the covariance matrix as given in Eq (16.151). This fact is used to test hypotheses regarding the significance or otherwise of the parameters in precisely the same manner as before. The t-statistic arises directly from substituting data estimate s_e for σ in Eq (16.151).

Thus, when cast in matrix form, the multiple regression problem is seen to be merely a higher dimensional form of the earlier simple linear regression problem; the model equations are structurally similar:

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\epsilon} \\ y &= \theta x + \epsilon \end{aligned}$$

as are the least-squares solutions:

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ \hat{\theta} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \\ \text{or } \hat{\theta} &= \frac{S_{xy}}{S_{xx}} \end{aligned} \tag{16.156}$$

The computations for multiple regression problems become rapidly more complex, but all the results obtained earlier for the simpler regression problem transfer virtually intact, including hypothesis tests of significance for the parameters, the values for the coefficient of determination, R^2 (and its "adjusted" variant, R^2_{adj}) for assessing the model adequacy in capturing the data information. Fortunately, these computations are routinely carried out by computer software packages. Nevertheless, the reader is reminded that the availability of these computer programs has relieved us *only* of the computational burden; the task of *understanding* what these computations are based upon remains very much an important responsibility of the practitioner.

Residuals Analysis

The residuals in multiple linear regression are given by Eq (16.153). Obtaining the standardized version of residuals in this case requires the introduction of a new matrix, $\mathbf{H}$, the so-called "hat matrix." Introducing the least-squares estimate into the expression for the vector of model estimates in Eq (16.152) yields:

$$\hat{\mathbf{y}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y} \tag{16.157}$$

where

$$\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \tag{16.158}$$

is called the "hat" matrix because it relates the actual observations vector, $\mathbf{y}$, to the vector of model fits, $\hat{\mathbf{y}}$. This matrix has some unique characteristics: for example, it is an idempotent matrix, meaning that

$$\mathbf{H}\mathbf{H} = \mathbf{H} \tag{16.159}$$

The residual vector, $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}}$, may therefore be represented as:

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} \tag{16.160}$$

($(\mathbf{I} - \mathbf{H})$ is also idempotent). If h_{ii} represents the diagonal elements of $\mathbf{H}$, the standardized residuals are obtained for multiple regression problems as:

$$e_i^* = \frac{e_i}{s_e\sqrt{1 - h_{ii}}} \tag{16.161}$$

These standardized residuals are the exact equivalents of the ones shown in Eq (16.116) for the simple linear regression case.

The next example illustrates an application of these results.

Example 16.8: QUANTIFYING TEMPERATURE AND PRESSURE EFFECTS ON YIELD

In an attempt to quantify the effect of temperature and pressure on the yield obtained from a laboratory scale catalytic process, the data shown in Table 16.7 was obtained from a series of designed experiments where

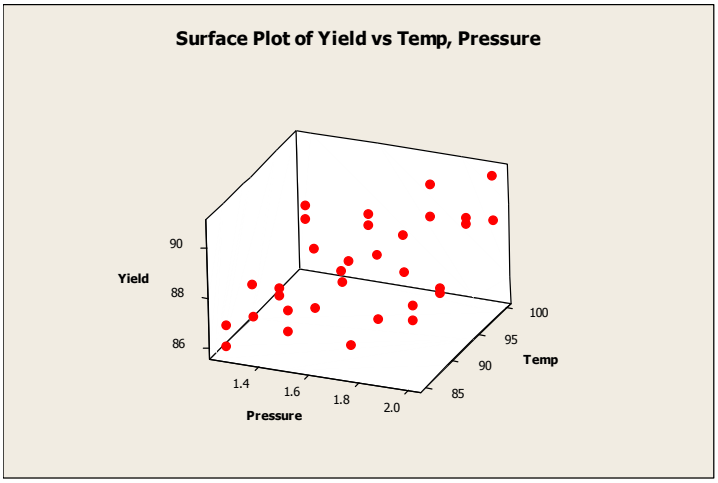

FIGURE 16.10: Catalytic process yield data of Table 16.7.

temperature was varied over a relatively narrow range, from $85°C$ to $100\ °C$, and pressure from 1.25 atmospheres to 2 atmospheres. If yield is y, temperature is x_1, and pressure is x_2, obtain a regression model of the type in Eq (16.135) and evaluate the model fit.

Solution:
A 3-dimensional scatter plot of the data is shown in Fig 16.10, where it appears as if the data truly fall on a plane.

The results from an analysis carried out using MINITAB is as follows:

Regression Analysis: Yield versus Temp, Pressure

```
The regression equation is
75.9 + 0.0757 Temp + 3.21 Pressure
```

Predictor	Coef	SE Coef	T	P
Constant	75.866	2.924	25.95	0.000
Temp	0.07574	0.02977	2.54	0.017
Pressure	3.2120	0.5955	5.39	0.000

```
S = 0.941538 R-Sq = 55.1 % R-Sq(adj) = 52.0%
```

Thus, the fitted regression line equation is, in this case

$$\hat{y} = 75.9 + 0.0757x_1 + 3.21x_2 \tag{16.162}$$

The p-values associated with all the parameters indicate significance; the estimate of the error standard deviation is as shown (0.94) with the R^2 and R^2_{adj} values indicating that the model explanation of the variation in the data is only modest.

The fitted plane represented by Eq (16.162) and the standardized

TABLE 16.7: Laboratory experimental data on yield obtained from a catalytic process at various temperatures and pressures

Yield (%)	Temp ($^{\circ}C$)	Pressure (Atm)
86.8284	85	1.25
87.4136	90	1.25
86.2096	95	1.25
87.8780	100	1.25
86.9892	85	1.50
86.8632	90	1.50
86.8389	95	1.50
88.0432	100	1.50
86.8420	85	1.75
89.3775	90	1.75
87.6432	95	1.75
90.0723	100	1.75
88.8353	85	2.00
88.4265	90	2.00
90.1930	95	2.00
89.0571	100	2.00
85.9974	85	1.25
86.1209	90	1.25
85.8819	95	1.25
88.4381	100	1.25
87.8307	85	1.50
89.2073	90	1.50
87.2984	95	1.50
88.5071	100	1.50
90.1824	85	1.75
86.8078	90	1.75
89.1249	95	1.75
88.7684	100	1.75
88.2137	85	2.00
88.2571	90	2.00
89.9551	95	2.00
90.8301	100	2.00

residual errors are shown in Fig 16.11. There is nothing unusual about the residuals but the relatively modest values of R^2 and R^2_{adj} seem to suggest that the true model might be somewhat more complicated than the one we have postulated and fitted here; it could also mean that there is significant noise associated with the measurement, or both.

16.4.3 Some Important Special Cases

Weighted Least Squares

For a wide variety of reasons, ranging from x_i variables with values that are orders of magnitude apart (e.g., if x_1 is temperature in the 100s of degrees, while x_2 is mole fraction, naturally scaled between 0 and 1), to measurement errors with non-constant variance-covariance structures, it is often necessary to modify the basic multiple linear regression problem by placing more or less weight on different observations. Under these circumstances, the regression model equation in Eq (16.141) is modified to:

$$\mathbf{Wy} = \mathbf{WX}\boldsymbol{\theta} + \mathbf{W}\boldsymbol{\epsilon} \tag{16.163}$$

where $\mathbf{W}$ is an appropriately chosen $(n \times n)$ weighting matrix. Note that the pre-multiplication in Eq (16.163) has not changed the model itself; it merely allows a re-scaling of the $\mathbf{X}$ matrix and/or the error vector. However, the introduction of the weights does affect the solution to the least-squares optimization problem. It can be shown that in this case, the sum-of-squares

$$S(\boldsymbol{\theta}) = (\mathbf{Wy} - \mathbf{WX}\boldsymbol{\theta})^T(\mathbf{Wy} - \mathbf{WX}\boldsymbol{\theta}) \tag{16.164}$$

is minimized by

$$\hat{\boldsymbol{\theta}}_{WLS} = \left(\mathbf{X}^T\mathbf{W}^T\mathbf{WX}\right)^{-1}\mathbf{X}^T\mathbf{W}^T\mathbf{Wy} \tag{16.165}$$

This is known as the weighted least-squares (WLS) estimate, and it is easy to establish that regardless of the choice of $\mathbf{W}$,

$$E\left(\hat{\boldsymbol{\theta}}_{WLS}\right) = \boldsymbol{\theta} \tag{16.166}$$

so that this is also an unbiased estimate of $\boldsymbol{\theta}$.

In cases where the motivation for introducing weights arises from the structure of the error covariance matrix, $\boldsymbol{\Sigma}$, it is recommended that $\mathbf{W}$ be chosen such that

$$\mathbf{W}^T\mathbf{W} = \boldsymbol{\Sigma}^{-1} \tag{16.167}$$

Under these circumstances, the covariance matrix of the WLS estimate can be shown to be given by:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\theta}}_{WLS}} = \left(\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X}\right)^{-1} \tag{16.168}$$

making it comparable to Eq (16.151). All regression packages provide an option for carrying out WLS instead of ordinary least squares.

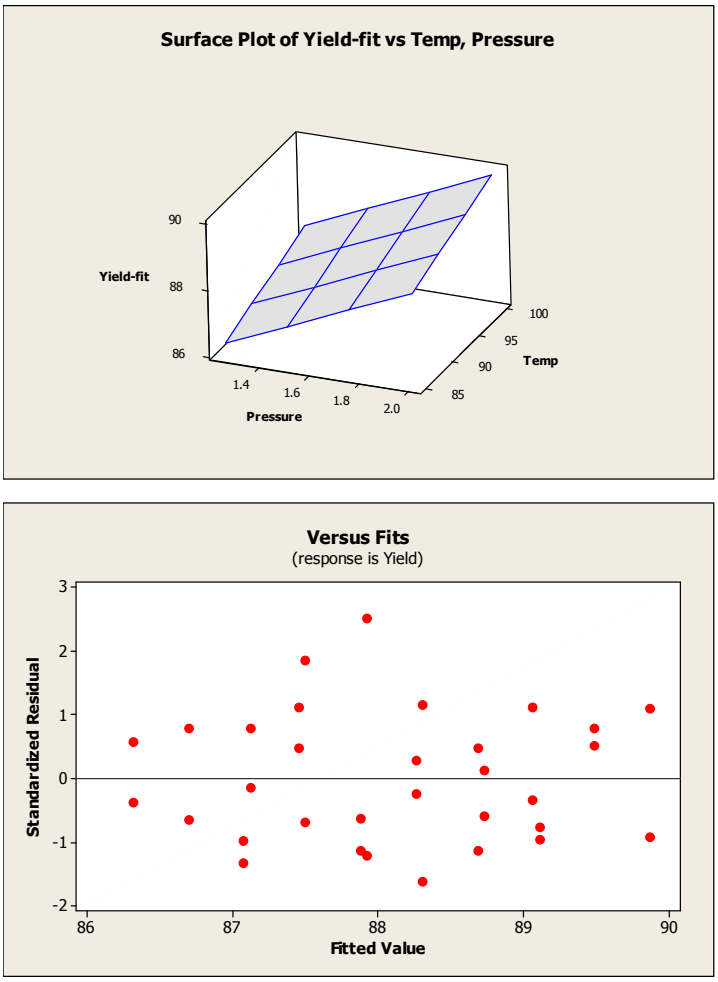

FIGURE 16.11: Catalytic process yield data of Table 16.1. Top: fitted plane of yield as a function of temperature and pressure; Bottom: standardized residuals versus fitted value $\hat{y}_i$. Nothing appears unusual about these residuals.

Constrained Least Squares

Occasionally, one encounters regression problems in engineering where the model parameters are subject to a set of linear equality constraints. For example, in a blending problem where the unknown parameters, θ_1, θ_2, and θ_3 to be estimated from experimental data, are the mole fractions of the three component materials, clearly

$$\theta_1 + \theta_2 + \theta_3 = 1 \tag{16.169}$$

In general such linear constraints are of the form:

$$\mathbf{L}\boldsymbol{\theta} = \mathbf{v} \tag{16.170}$$

When subject to such constraints, obtaining the least-squares estimate of the parameters in the Eq (16.141) model now requires attaching these constraint equations to the original problem of minimizing the sum-of-squares function $S(\boldsymbol{\theta})$. It can be shown that when the standard tools of Lagrange multipliers are used to solve this constrained optimization problem, the solution is:

$$\hat{\boldsymbol{\theta}}_{CLS} = \hat{\boldsymbol{\theta}} + \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{L}^T\left[\mathbf{L}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{L}^T\right]^{-1}(\mathbf{v} - \mathbf{L}\hat{\boldsymbol{\theta}}) \tag{16.171}$$

the constrained least-squares (CLS) estimate, where $\hat{\boldsymbol{\theta}}$ is the normal, unconstrained least-squares estimate in Eq (16.145).

If we define a "gain" matrix:

$$\boldsymbol{\Gamma} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{L}^T\left[\mathbf{L}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{L}^T\right]^{-1} \tag{16.172}$$

then Eq (16.171) may be rearranged to:

$$\hat{\boldsymbol{\theta}}_{CLS} = \hat{\boldsymbol{\theta}} + \boldsymbol{\Gamma}(\mathbf{v} - \mathbf{L}\hat{\boldsymbol{\theta}}) \tag{16.173}$$

$$\text{or } \hat{\boldsymbol{\theta}}_{CLS} = \boldsymbol{\Gamma}\mathbf{v} + (\mathbf{I} - \boldsymbol{\Gamma}\mathbf{L})\hat{\boldsymbol{\theta}} \tag{16.174}$$

where the former (as in Eq (16.171)) emphasizes how the constraints provide a correction to the unconstrained estimate, and the latter emphasizes that $\hat{\boldsymbol{\theta}}_{CLS}$ provides a compromise between unconstrained estimates and the constraints.

Ridge Regression

The ordinary least-squares estimate, $\hat{\boldsymbol{\theta}}$, given in Eq (16.145), will be unacceptable for "ill-conditioned" problems for which $\mathbf{X}^T\mathbf{X}$ is nearly singular typically because the determinant, $|\mathbf{X}^T\mathbf{X}| \approx 0$. This will occur, for example, when there is near-linear dependence in some of the predictor variables, x_i, or when some of the x_i variables are orders of magnitude different from others,

and the problem has not been re-scaled accordingly. The problem created by ill-conditioning manifests in the form of overly inflated values for the elements of the matrix inverse $(\mathbf{X}^T\mathbf{X})^{-1}$ as a result of the vanishingly small determinant. Consequently, the norm of the estimate vector, $\hat{\boldsymbol{\theta}}$, will be too large, and the uncertainty associated with the estimates (see Eq (16.151)) will be unacceptably large.

One solution is to augment the original model equation as follows:

$$\begin{bmatrix} \mathbf{y} \\ \cdots \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \cdots \\ k\mathbf{I} \end{bmatrix} \boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{16.175}$$

or,

$$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\theta} + \boldsymbol{\epsilon} \tag{16.176}$$

where $\mathbf{0}$ is an m-dimensional vector of zeros, k is a constant, and $\mathbf{I}$ is the identity matrix. Instead of minimizing the original sum of squares function, minimizing the sum of squares based on the augmented Eq (16.176) results in the so-called ridge regression estimate:

$$\hat{\boldsymbol{\theta}}_{RR} = (\mathbf{X}^T\mathbf{X} + k^2\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{16.177}$$

As is evident from Eq (16.175), the purpose of the augmentation is to use the constant k to force the estimated parameters to be close to 0, preventing over-inflation. As the value chosen for k increases, the conditioning of the matrix $(\mathbf{X}^T\mathbf{X} + k^2\mathbf{I})$ improves, reducing the otherwise inflated estimates $\hat{\boldsymbol{\theta}}$. However, this improvement is achieved at the expense of introducing bias into the resulting estimate vector. Still, even though $\hat{\boldsymbol{\theta}}_{RR}$ is biased, its variance is much better than that of the original $\hat{\boldsymbol{\theta}}$. (See Hoerl (1962)[1] and Hoerl and Kennard (1970a,[2] 1970b[3]).) Selecting an appropriate value of k is an art. (See Marquardt (1970).[4])

16.4.4 Recursive Least Squares

Problem Formulation

A case often arises in engineering where the experimental data used to estimate parameters in the model in Eq (16.141) are available sequentially. After accumulating a set of n observations, $y_i; i = 1, 2, \ldots, n$, and, subsequently

[1]Hoerl, A.E. (1962). "Application of ridge analysis to regression problems," *Chem. Eng. Prog.* 55, 54–59.

[2]Hoerl A.E., and R.W. Kennard. (1970). "Ridge regression. Biased estimation for nonorthogonal problems," *Technometrics*, 12, 55–67.

[3]Hoerl A.E., and R.W. Kennard. (1970). "Ridge regression. Applications to nonorthogonal problems," *Technometrics*, 12, 69–82.

[4]Marquardt, D.W. (1970). "Generalized inverses, Ridge regression, Biased linear estimation, and Nonlinear estimation," *Technometrics*, 12, 591–612.

using this n-dimensional data vector, $\mathbf{y}_n$, to obtain the parameter estimates as:

$$\hat{\boldsymbol{\theta}}_n = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}_n \tag{16.178}$$

suppose that a new, single observation y_{n+1} then becomes available. This new data can be combined with past information to obtain an updated estimate that reflects the additional information about the parameter contained in the data, information represented by:

$$y_{n+1} = \mathbf{x}_{n+1}^T\boldsymbol{\theta} + \epsilon_{n+1} \tag{16.179}$$

where $\mathbf{x}_{n+1}^T$ is the m-dimensional vector of the independent predictor variables used to generate this new $(n+1)^{th}$ observation, and ϵ_{n+1} is the associated random error component.

In principle, we can append this new information to the old to obtain:

$$\begin{bmatrix} \mathbf{y}_n \\ \cdots \\ y_{n+1} \end{bmatrix} = \begin{bmatrix} \mathbf{X} \\ \cdots \\ \mathbf{x}_{n+1}^T \end{bmatrix}\boldsymbol{\theta} + \begin{bmatrix} \boldsymbol{\epsilon}_n \\ \cdots \\ \epsilon_{n+1} \end{bmatrix} \tag{16.180}$$

or, more compactly,

$$\mathbf{y}_{n+1} = \mathbf{X}_{n+1}\boldsymbol{\theta} + \boldsymbol{\epsilon}_{n+1} \tag{16.181}$$

so that the new $\mathbf{X}$ matrix, $\mathbf{X}_{n+1}$, is now an $(n+1)\times m$ matrix, with the data vector $\mathbf{y}_{n+1}$ now of dimension $(n+1)$. From here, we can use these new matrices and vectors to obtain the new least-squares estimate directly, as before, to give:

$$\hat{\boldsymbol{\theta}}_{n+1} = \left(\mathbf{X}_{n+1}^T\mathbf{X}_{n+1}\right)^{-1}\mathbf{X}_{n+1}^T\mathbf{y}_{n+1} \tag{16.182}$$

Again, in principle, we can repeat this exercise each time a new observation becomes available. However, such a strategy requires that we recompute the estimates from scratch every time as if for the first time. While it is true that the indicated computational burden is routinely borne nowadays by computers, the fact that the information is available recursively raises a fundamental question: *instead of having to recompute the estimate $\hat{\boldsymbol{\theta}}_{n+1}$ all over again as in Eq (16.182) every time new information is available, is it possible to determine it by judiciously updating $\hat{\boldsymbol{\theta}}_n$ directly with the new information?* The answer is provided by the recursive least-squares technique whereby $\hat{\boldsymbol{\theta}}_{n+1}$ is obtained recursively as a function of $\hat{\boldsymbol{\theta}}_n$.

Recursive Least-Squares Estimation

We begin by obtaining the least-squares estimate, $\hat{\boldsymbol{\theta}}_{n+1}$, directly from the partitioned matrices in Eq (16.180), giving the result

$$\hat{\boldsymbol{\theta}}_{n+1} = \left[\mathbf{X}^T\mathbf{X} + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\right]^{-1}\left[\mathbf{X}^T\mathbf{y}_n + \mathbf{x}_{n+1}y_{n+1}\right] \tag{16.183}$$

Now, let us define

$$\mathbf{P}_{n+1} = \left[\mathbf{X}^T\mathbf{X} + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\right]^{-1} \tag{16.184}$$

so that by analogy with Eq (16.178),

$$\mathbf{P}_n = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}; \tag{16.185}$$

then,

$$\mathbf{P}_{n+1}^{-1} = \mathbf{X}^T\mathbf{X} + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T$$

From here, on the one hand, by simple rearrangement,

$$\mathbf{X}^T\mathbf{X} = \mathbf{P}_{n+1}^{-1} - \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T \tag{16.186}$$

and on the other, using Eq (16.185),

$$\mathbf{P}_{n+1}^{-1} = \mathbf{P}_n^{-1} + \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T \tag{16.187}$$

Now, as a result of Eq (16.184), the least-squares estimate in Eq (16.183) may be written as:

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{n+1} &= \mathbf{P}_{n+1}\left[\mathbf{X}^T\mathbf{y}_n + \mathbf{x}_{n+1}y_{n+1}\right] \\ &= \mathbf{P}_{n+1}\mathbf{X}^T\mathbf{y}_n + \mathbf{P}_{n+1}\mathbf{x}_{n+1}y_{n+1}\end{aligned} \tag{16.188}$$

Returning briefly to Eq (16.178) from which

$$(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\theta}}_n = \mathbf{X}^T\mathbf{y}_n,$$

upon introducing Eq (16.186), we obtain

$$\left(\mathbf{P}_{n+1}^{-1} - \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\right)\hat{\boldsymbol{\theta}}_n = \mathbf{X}^T\mathbf{y}_n \tag{16.189}$$

Introducing this into Eq (16.188) yields

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{n+1} &= \mathbf{P}_{n+1}\left(\mathbf{P}_{n+1}^{-1} - \mathbf{x}_{n+1}\mathbf{x}_{n+1}^T\right)\hat{\boldsymbol{\theta}}_n + \mathbf{P}_{n+1}\mathbf{x}_{n+1}y_{n+1} \\ \text{or, } \hat{\boldsymbol{\theta}}_{n+1} &= \hat{\boldsymbol{\theta}}_n + \mathbf{P}_{n+1}\mathbf{x}_{n+1}\left(y_{n+1} - \mathbf{x}_{n+1}^T\hat{\boldsymbol{\theta}}_n\right)\end{aligned} \tag{16.190}$$

This last equation is the required recursive expression for determining $\hat{\boldsymbol{\theta}}_{n+1}$ from $\hat{\boldsymbol{\theta}}_n$ and new response data, y_{n+1}, generated with the new values $\mathbf{x}_{n+1}^T$ for the predictor variables. The "gain matrix," $\mathbf{P}_{n+1}$, is itself generated recursively from Eq (16.187). And now, several points are worth noting:

1. The matrix $\mathbf{P}_n$ is related to the covariance matrix of the estimate, $\hat{\boldsymbol{\theta}}_n$ (see Eq (16.151)), so that Eq (16.187) represents the recursive evolution of the covariance of the estimates as n increases;

2. The term in parentheses in Eq (16.190) resembles a correction term, the discrepancy between the actual observed response, y_{n+1}, and an *à-priori* value predicted for it (before the new data is available) using the previous estimate, $\hat{\boldsymbol{\theta}}_n$, and the new predictor variables, $\mathbf{x}_{n+1}^T$;

3. This recursive procedure allows us to begin with an initial estimate, $\hat{\boldsymbol{\theta}}_0$, along with a corresponding (scaled) covariance matrix, $\mathbf{P}_0$, and proceed recursively to estimate the true value of the parameters one data point at a time, using Eq (16.187) first to obtain an updated covariance matrix, and Eq (16.190) to update the parameter estimate;

4. Readers familiar with Kalman filtering in dynamical systems theory will immediately recognize the structural similarity between the combination of Eqs (16.187) and (16.190) and the discrete Kalman filter.

16.5 Polynomial Regression

16.5.1 General Considerations

A special case of multiple linear regression occurs when, in Eq (16.133), the response Y depends on powers of a single predictor variable, x, i.e.,

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \cdots + \theta_m x^m + \epsilon \tag{16.191}$$

in which case $x_i = x^i$. An example was given earlier in Eq (16.135), with Y as a quadratic function of x.

This class of regression models is important because, in engineering, many unknown functional relationships, $y(x)$, can be approximated by polynomials. Because Eq (16.191) is a special case of Eq (16.133), all the results obtained earlier for the more general problem transfer directly, and there is not much to add for this restricted class of problems. However, in terms of practical application, there are some peculiarities unique to polynomial regression analysis.

In many practical problems, the starting point in polynomial regression is often a low order linear model; when residual analysis indicates that the simple model is inadequate, the model complexity is then increased, typically by adding the next higher power of x, until the model is deemed "adequate." But one must be careful: fitting an m^{th} order polynomial to $m+1$ data points (e.g., fitting a straight line to 2 points) will produce a perfect $R^2 = 1$ but the parameter estimates will be unreliable. The primary pitfall to avoid in such an exercise is therefore "overfitting," whereby the polynomial model is of an order higher than can be realistically supported by the data. Under such circumstances, the improvement in R^2 must be cross-checked against the corresponding R^2_{adj} value.

The next examples illustrate the application of polynomial regression.

Example 16.9: BOILING POINT OF HYDROCARBONS: REVISITED

In Example 16.6, a linear two-parameter model was postulated for the relationship between the number of carbon atoms in the hydrocarbon compounds listed in Table 16.1 and the respective boiling points. Upon evaluation, however, the model was found to be inadequate; specifically, the residuals indicated the potential for a "left over" quadratic component. Postulate the following quadratic model,

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \epsilon \tag{16.192}$$

and evaluate a least-squares fit of this model to the data. Compare this model fit to the simple linear model obtained in Example 16.6.

Solution:

Once more, when we use MINITAB for this problem, we obtain the following results:

Regression Analysis: Boiling Point versus n, n2

```
The regression equation is
Boiling Point = - 218 + 66.7 n - 3.02 n2
```

Predictor	Coef	SE Coef	T	P
Constant	-218.143	8.842	-24.67	0.000
n	66.667	4.508	14.79	0.000
n2	-3.0238	0.4889	-6.18	0.002

```
S = 6.33734 R-Sq = 99.7% R-Sq(adj) = 99.6%
```

Thus, the fitted quadratic regression line equation is

$$\hat{y} = -218.14 + 66.67x - 3.02x^2 \tag{16.193}$$

where, as before, y is the hydrocarbon compound boiling point, and the number of carbon atoms it contains is x.

Note how the estimates for θ_0 and θ_1 are now different from the respective values obtained when these were the only two parameters in the model. This is a natural consequence of adding a new component to the model; the responsibility for capturing the variability in the data is now being shared by three parameters instead of two, and the best estimates of the model parameter set will change accordingly.

Before inspecting the model fit and the residuals, we note first that the three parameters in this case also are all significantly different from zero (the p-values are zero for the constant term and the linear term and 0.002 for the quadratic coefficient). As expected, there is an improvement in R^2 for this more complicated model (99.7% versus 97.4% for the simpler linear model); furthermore, this improvement was also accompanied by a commensurate improvement in R^2_{adj} (99.6% versus 97.0% for the simpler model). Thus, the improved model performance was not achieved at the expense of overfitting, indicating that the added

quadratic term is truly warranted. The error standard deviation also shows an almost three-fold improvement from $S = 17.0142$ for the linear model to $S = 6.33734$, again indicating that more of the variability in the data has been captured by the more complicated model.

The model fit to the data, shown in the top panel of Fig 16.12, indicates a much-improved fit, compared with the one in the top panel of Fig 16.9. This is also consistent with everything we have noted so far. However, the normalized residuals versus fitted values, plotted in the bottom panel of Fig 16.12 shows that there is still some "left over" structure, the improved fit notwithstanding. The implication is that perhaps an additional cubic term might be necessary to capture the remaining structural information still visible in the residual plot. This suggests further revising the model as follows:

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \epsilon \tag{16.194}$$

The problem of fitting an adequate regression model to the data in Table 16.1 concludes with this next example.

Example 16.10: BOILING POINT OF HYDROCARBONS: PART III

As a follow up to the analysis in the last example, fit the cubic equation

$$Y = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \epsilon \tag{16.195}$$

to the data in Table 16.1, evaluate the model fit, and compare it to the fit obtained in Example 16.9.

Solution:

This time around, the MINITAB results are as follows:

Regression Analysis: Boiling Point versus n, n2, n3

```
The regression equation is
Boiling Point = - 244 + 93.2 n - 9.98 n2 + 0.515 n3
```

Predictor	Coef	SE Coef	T	P
Constant	-243.643	8.095	-30.10	0.000
n	93.197	7.325	12.72	0.000
n2	-9.978	1.837	-5.43	0.006
n3	0.5152	0.1348	3.82	0.019

```
S = 3.28531 R-Sq = 99.9% R-Sq(adj) = 99.9%
```

The fitted cubic regression equation is

$$\hat{y} = -243.64 + 93.20x - 9.98x^2 + 0.515x^3 \tag{16.196}$$

Note that the estimates for θ_0 and θ_1 have changed once again, as has the estimate for θ_2. Again, this is a natural consequence of adding the new parameter, θ_3, to the model.

As a result of the p-values, we conclude once again, that all the four

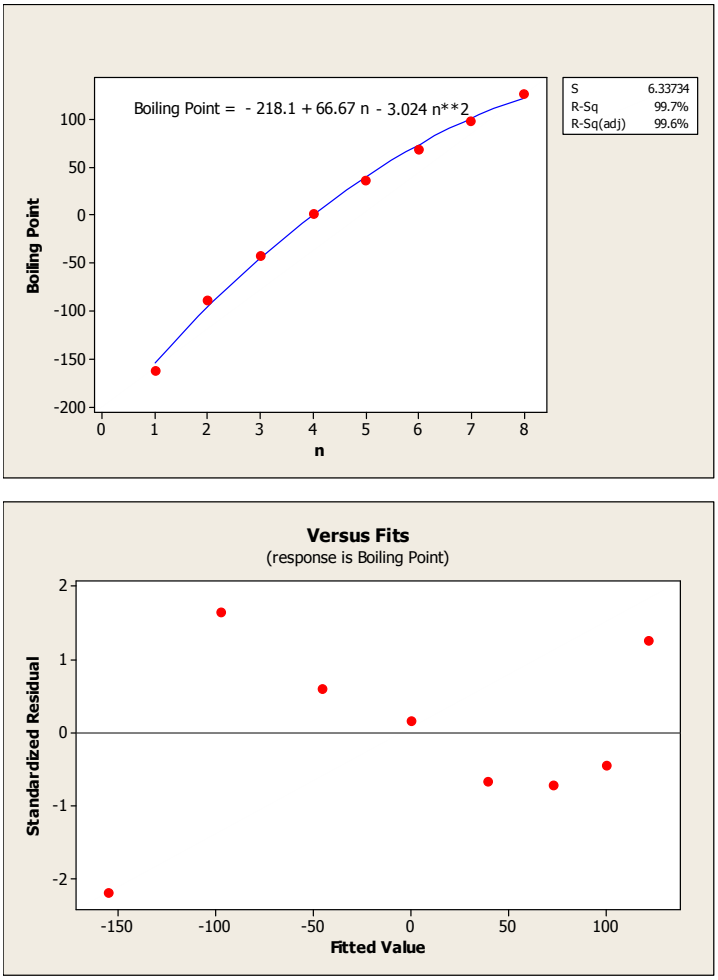

FIGURE 16.12: Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound. Top: Fitted quadratic curve of BP versus n, the number of carbon atoms; Bottom: standardized residuals versus fitted value $\hat{y}_i$. Despite the good fit, the visible systematic structure still "left over" in the residuals suggests adding one more term to the model.

parameters in this model are significantly different from zero; the R^2 and R^2_{adj} values are virtually perfect and identical, indicating that the expenditure of four parameters in this model is justified.

The error standard deviation has improved further by a factor of almost 2 (from $S = 6.33734$ for the quadratic model to $S = 3.28531$ for this cubic model) and the model fit to the data shows this improvement graphically in the top panel of Fig 16.13. This time, the residual plot in the bottom panel of Fig 16.13 shows no significant "left over" structure. Therefore, in light of all the factors considered above, we conclude that the cubic fit in Eq (16.196) appears to provide an adequate fit to the data; and that this has been achieved without the expenditure of an excessive number of parameters.

A final word: While polynomial models may provide adequate representations of data (as in the last example), this should not be confused with a fundamental scientific *explanation* of the underlying relationship between x and Y. Also, it is not advisable to extrapolate the model prediction outside of the range covered by the data used to fit the model. For example, the model in this last example should not be used to predict the boiling point of hydrocarbons with carbon atoms ≥ 9.

16.5.2 Orthogonal Polynomial Regression

Let $p_i(x); i = 0, 1, 2, \ldots, m$, be a sequence of i^{th} order polynomials in the single independent variable, x, defined on an interval $[x_L, x_R]$ on the real line. For our purposes here, these polynomials take on *discrete* values $p_i(x_k)$ at equally spaced values $x_k; k = 1, 2, \ldots, n$, in the noted interval. The sequence of polynomials constitutes an orthogonal set if the following conditions hold:

$$\sum_{k=1}^{n} p_i(x_k)p_j(x_k) = \begin{cases} \psi_i^2 & i = j; \\ 0 & i \neq j \end{cases} \tag{16.197}$$

An Example: Gram Polynomials

Without loss of generality, let the independent variable x be defined in the interval $[-1, 1]$ (for variables defined on $[x_l, x_R]$, a simple scaling transformation is all that is required to obtain a corresponding variable defined instead on $[-1, 1]$); furthermore, let this interval be divided into n equal discrete intervals, $k = 1, 2, \ldots, n$, to provide n values of x at $x_1, x_2, \ldots, x_n$, such that $x_1 = -1$, $x_n = 1$, and in general,

$$x_k = \frac{2(k-1)}{n-1} - 1 \tag{16.198}$$

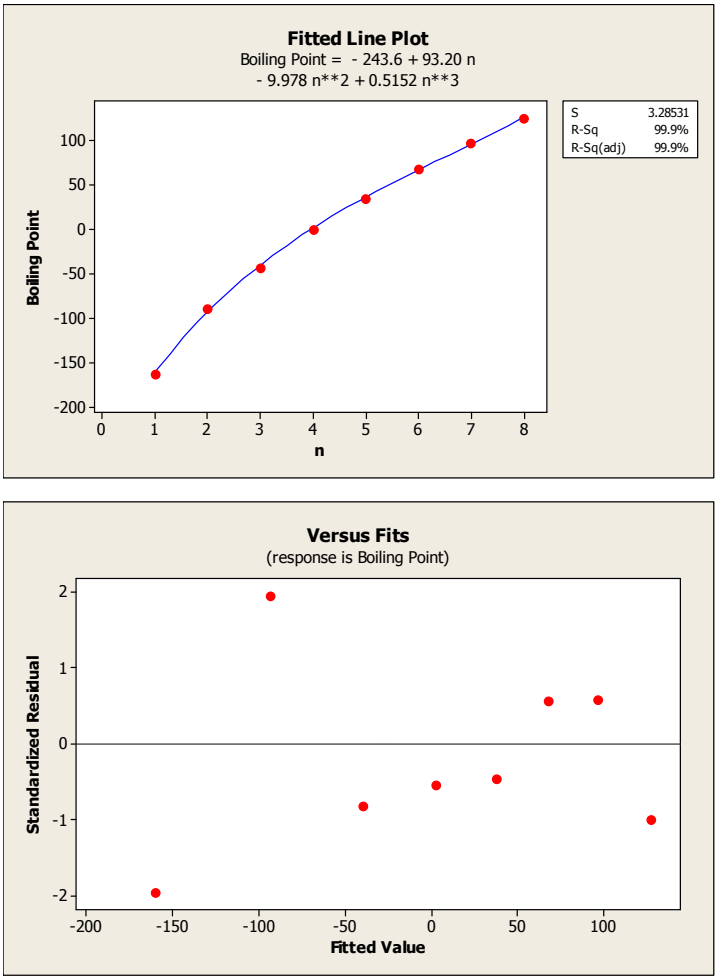

FIGURE 16.13: Modeling the dependence of the boiling points (BP) of hydrocarbon compounds in Table 16.1 on the number of carbon atoms in the compound. Top: fitted cubic curve of BP versus n, the number of carbon atoms; Bottom: standardized residuals versus fitted value $\hat{y}_i$. There appears to be little or no systematic structure left in the residuals, suggesting that the cubic model provides an adequate description of the data.

The set of Gram polynomials defined on $[-1, 1]$ for x_k as given above is,

$$\begin{aligned}
p_0(x_k) &= 1 \\
p_1(x_k) &= x_k \\
p_2(x_k) &= x_k^2 - \frac{(n+1)}{3(n-1)} \\
p_3(x_k) &= x_k^3 - \left[\frac{(3n^2-7)}{5(n-1)^2}\right] x_k \\
&\vdots \quad\quad \vdots \\
p_{\ell+1}(x) &= x_k\, p_\ell(x_k) - \left[\frac{\ell^2(n^2-\ell^2)}{(4\ell^2-1)(n-1)^2}\right] p_{\ell-1}(x_k)
\end{aligned} \tag{16.199}$$

where each polynomial in the set is generated from the "recurrence relation" in Eq (16.199), given the initial two, $p_0(x_k) = 1$ and $p_1(x_k) = x_k$.

Example 16.11: ORTHOGONALITY OF GRAM POLYNOMIALS
Obtain the first four Gram polynomials determined at $n = 5$ equally spaced values, x_k, of the independent variable, x, on the interval $-1 \leq x \leq 1$. Show that these polynomials are mutually orthogonal.

Solution:
First, from Eq (16.198), the values x_k at which the polynomials are to be determined in the interval $1 \leq x \leq 1$ are:

$$x_1 = -1;\ x_2 = -0.5;\ x_3 = 0;\ x_4 = 0.5;\ x_5 = 1.$$

Next, let the 5-dimensional vector, $\mathbf{p}_i; i = 0, 1, 2, 3$, represent the values of the polynomial $p_i(x_k)$ determined at these 5 x_k values: i.e.,

$$\mathbf{p}_i = \begin{bmatrix} p_i(x_1) \\ p_i(x_2) \\ \vdots \\ p_i(x_5) \end{bmatrix} \tag{16.200}$$

Then, from the expressions given above, we obtain, using $n = 5$, that:

$$\mathbf{p}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix};\ \mathbf{p}_1 = \begin{bmatrix} -1 \\ -0.5 \\ 0 \\ 0.5 \\ 1 \end{bmatrix};\ \mathbf{p}_2 = \begin{bmatrix} 0.50 \\ -0.25 \\ -0.50 \\ -0.25 \\ 0.50 \end{bmatrix};\ \mathbf{p}_3 = \begin{bmatrix} -0.15 \\ 0.30 \\ 0.00 \\ -0.30 \\ 0.15 \end{bmatrix}$$

A plot of these values is shown in Fig 16.14 where we see that $p_0(x_k)$ is a constant, $p_1(x_k)$ is a straight line, $p_2(x_k)$ is a quadratic and $p_3(x_k)$ is a cubic, each one evaluated at the indicated discrete points.

To establish orthogonality, we compute inner products, $\mathbf{p}_i^T \mathbf{p}_j$, for

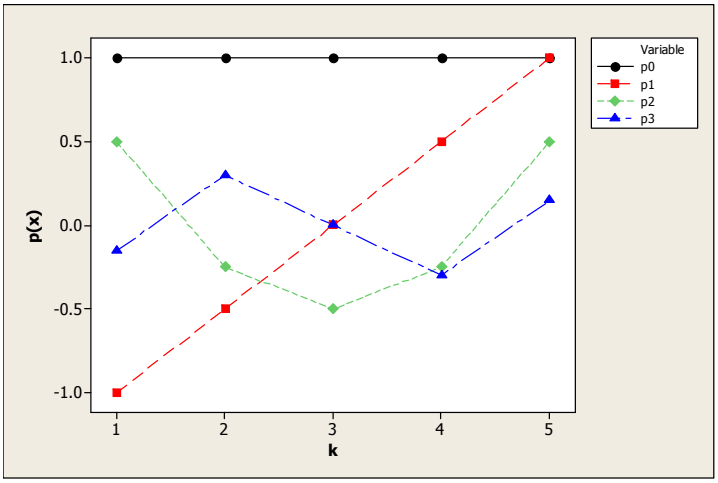

FIGURE 16.14: Gram polynomials evaluated at 5 discrete points $k = 1, 2, 3, 4, 5$; p_0 is the constant; p_1, the straight line; p_2, the quadratic; and p_3, the cubic.

all combinations of $i \neq j$. First, we note that $\mathbf{p}_0^T \mathbf{p}_j$ is simply a sum of all the elements in each vector, which is uniformly zero in all cases, i.e.,

$$\mathbf{p}_0^T \mathbf{p}_j = \sum_{k=1}^{5} p_j(x_k) = 0 \tag{16.201}$$

Next, we obtain:

$$\mathbf{p}_1^T \mathbf{p}_2 = \sum_{k=1}^{5} p_1(x_k) p_2(x_k) = -0.500 + 0.125 + 0.000 - 0.125 + 0.500 = 0$$

$$\mathbf{p}_1^T \mathbf{p}_3 = \sum_{k=1}^{5} p_1(x_k) p_3(x_k) = 0.15 - 0.15 + 0.00 - 0.15 + 0.15 = 0$$

$$\mathbf{p}_2^T \mathbf{p}_3 = \sum_{k=1}^{5} p_2(x_k) p_3(x_k) = -0.075 - 0.075 + 0.000 + 0.075 + 0.075 = 0$$

For completeness, the sums of squares, ψ_i^2, are obtained below (note monotonic decrease):

$$\psi_0^2 = \mathbf{p}_0^T \mathbf{p}_0 = \sum_{k=1}^{5} p_0(x_k) p_0(x_k) = 5$$

$$\psi_1^2 = \mathbf{p}_1^T \mathbf{p}_1 = \sum_{k=1}^{5} p_1(x_k) p_1(x_k) = 2.5$$

$$\psi_2^2 = \mathbf{p}_2^T \mathbf{p}_2 = \sum_{k=1}^{5} p_2(x_k) p_2(x_k) = 0.875$$

$$\psi_3^2 = \mathbf{p}_3^T \mathbf{p}_3 = \sum_{k=1}^{5} p_3(x_k) p_3(x_k) = 0.225$$

Application in Regression

Among many attractive properties possessed by orthogonal polynomials, the following is the most relevant to the current discussion:

Orthogonal Basis Function Expansion: Any m^{th} order polynomial, $U(x)$, can be expanded in terms of an orthogonal polynomial set, $p_0(x), p_1(x), \ldots, p_m(x)$, as the basis functions, i.e.,

$$U(x) = \sum_{i=0}^{m} \alpha_i p_i(x) \tag{16.202}$$

This result has some significant implications for polynomial regression involving the single independent variable, x. Observe that as a consequence of this result, the original m^{th} order polynomial regression model in Eq (16.191) can be rewritten as

$$Y(x) = \alpha_0 p_0(x) + \alpha_1 p_1(x) + \alpha_2 p_2(x) + \cdots + \alpha_m p_m(x) + \epsilon \tag{16.203}$$

where we note that, given any specific set of orthogonal polynomial basis, the one-to-one relationship between the original parameters, θ_i, and the new set, α_i, is easily determined. Regression analysis is now concerned with estimating the new set of parameters, α_i, instead of the old set, θ_i, a task that is rendered dramatically easier because of the orthogonality of the basis set, $p_i(x)$, as we now show.

Suppose that the data $y_i; i = 1, 2, \ldots, n$, have been acquired using equally spaced values $x_k; k = 1, 2, \ldots, n$, in the range $[x_L, x_R]$ over which the orthogonal polynomial set, $p_i(x)$, is defined. In this case, from Eq (16.203), we will have:

$$\begin{aligned}
y(x_1) &= \alpha_0 p_0(x_1) + \alpha_1 p_1(x_1) + \alpha_2 p_2(x_1) + \cdots + \alpha_m p_m(x_1) + \epsilon_1 \\
y(x_2) &= \alpha_0 p_0(x_2) + \alpha_1 p_1(x_2) + \alpha_2 p_2(x_2) + \cdots + \alpha_m p_m(x_2) + \epsilon_2 \\
&\vdots \qquad\qquad \vdots \\
y(x_n) &= \alpha_0 p_0(x_n) + \alpha_1 p_1(x_n) + \alpha_2 p_2(x_n) + \cdots + \alpha_m p_m(x_n) + \epsilon_n
\end{aligned}$$

which, when written in matrix form, becomes:

$$\mathbf{y} = \mathbf{P}\boldsymbol{\alpha} + \boldsymbol{\epsilon} \tag{16.204}$$

The matrix $\mathbf{P}$ consists of vectors of the orthogonal polynomials computed at the discrete values x_k, just as we showed in Example 16.10 for the Gram polynomials. The least-squares solution to this equation,

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{P}^T\mathbf{P}\right)^{-1}\mathbf{P}^T\mathbf{y}, \tag{16.205}$$

looks very much like what we have seen before, until we recall that, as a result of the orthogonality of the constituent elements of $\mathbf{P}$, the matrix $\mathbf{P}^T\mathbf{P}$ is diagonal, with elements ψ_i^2, because all the off-diagonal terms vanish identically (see Eq (16.197) and Example 16.10)). As a result, the expression in Eq (16.205) is nothing but a collection of n isolated algebraic equations,

$$\hat{\alpha}_i = \frac{\sum_{k=1}^{n} p_i(x_k)y(x_k)}{\psi_i^2} \tag{16.206}$$

where

$$\psi_i^2 = \sum_{k=1}^{n} p_i(x_k)p_i(x_k); i = 0, 1, \ldots m. \tag{16.207}$$

This approach has several additional advantages beyond the dramatically simplified computation:

1. Each parameter estimate, $\hat{\alpha}_i$, is independent of the others, and its value remains unaffected by the order chosen for the polynomial model. In other words, after obtaining the first m parameter estimates for an m^{th} order polynomial model, should we decide to increase the polynomial order to $m+1$, the new parameter estimate, $\hat{\alpha}_{m+1}$, is simply obtained as
$$\hat{\alpha}_{m+1} = \frac{\sum_{k=1}^{n} p_{m+1}(x_k)y(x_k)}{\xi_{m+1}^2} \tag{16.208}$$
using the very same data set $y(x_k)$, and only introducing $p_{m+1}(k)$, a precomputed vector of the $(m+1)^{th}$ polynomial. All the previously obtained values for $\hat{\alpha}_i; i = 1, 2, \ldots, m$, remain unchanged. This is very convenient indeed. Recall that this is not the case with regular polynomial regression (see Examples 16.9 and 16.10) where a change in the order of the polynomial model mandates a change in the values estimated for the new set of parameters.

2. From earlier results, we know that the variance associated with the estimates, $\hat{\alpha}_i$, is given by:
$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\alpha}}} = \left(\mathbf{P}^T\mathbf{P}\right)^{-1}\sigma^2 \tag{16.209}$$
But, by virtue of the orthogonality of the elements of the $\mathbf{P}$ matrix, this reduces to:
$$\sigma_{\hat{\alpha}_i}^2 = \frac{\sigma^2}{\psi^2} \tag{16.210}$$

and since the value for the term ψ_i^2, defined as in Eq (16.207), is determined strictly by the placement of the "design" points, x_k, where the data is obtained, Eq (16.210) indicates that this approach makes it possible to select experimental points such as to influence the variance of the estimated parameters favorably, with obvious implications for strategic design of experiments.

3. Finally, it can be shown that ψ_i^2 decreases monotonically with i, indicating that the precision with which coefficients of higher order polynomials are estimated worsens with increasing order. This is also true for regular polynomial regression, but it is not as obvious.

An example of how orthogonal polynomial regression has been used in engineering applications may be found in Kristinsson and Dumont (1993,[5] 1996[6]).

16.6 Summary and Conclusions

The fitting of simple empirical mathematical expressions to data is an activity with which many engineers and scientists are very familiar, and perhaps have been engaged in even since high school. This chapter has therefore been more or less a re-introduction of the reader to regression analysis, especially to the fundamental principles behind the mechanical computations that are now routinely carried out with computer software. We have shown regression analysis to be a direct extension of estimation to cases where the mean of the random variation in the observations is no longer constant (as in our earlier discussions) but now varies as a function of one or more independent variables. The primary problem in regression analysis is therefore the determination of the unknown parameters contained in the functional relationship (the regression model equation), given appropriate experimental data. The primary method discussed in this chapter for carrying out this task is the method of least squares. However, when regression analysis is cast as the probabilistic estimation problem that it truly is fundamentally, one can also employ the method of maximum likelihood to determine the unknown parameters. However, this requires the explicit specification of a probability distribution for the observed random variability—something not explicitly required by the method of least squares. Still, under the normal distribution

[5] Kristinsson, K. and G. A. Dumont. (1993),"Paper Machine Cross Directional Basis Weight Control Using Gram Polynomials," *Proceedings of the Second IEEE Conference on Control Applications*, p 235–240, September 13–16, Vancouver, B.C.

[6] Kristinsson K. and G. A. Dumont. (1996), "Cross-directional control on paper machines using Gram polynomials," *Automatica* 32 (4), 533–548.

assumption, maximum likelihood estimates of the regression model parameters coincide precisely with least squares estimates (see Exercises 16.5 and 16.6).

In addition to the familiar, we have also presented some results for specialized problems, for example, when variances are not uniform across observations (weighted least squares); when the parameters are not independent but are subject to (linear) constraints (constrained least squares); when the data matrix is poorly conditioned perhaps because of collinearity (ridge regression); and when information is available sequentially (recursive least squares). Space constraints compel us to limit the illustration and application of these techniques to a handful of end-of-chapter exercises and application problems, which are highly recommended to the reader.

It bears repeating, in conclusion, that since all the computations required for regression analysis are now routinely carried out with the aid of computers, it is all the more important to concentrate on understanding the principles behind these computations, so that computer-generated results can be *interpreted* appropriately. In particular, first, the well-informed engineer should understand the implications of the following on the problem at hand:

- the results of hypothesis tests on the significance of estimated parameters;
- the R^2 and R^2_{adj} values as measures of how much of the information contained in the data has been adequately explained by the regression model, and with the "expenditure" of how many significant parameters;
- the value computed for the standard error of the residuals.

These will always be computed by any regression analysis software as a matter of course. Next, other quantities such as confidence and prediction intervals, and especially *residuals*, can be generated upon request. It is highly recommended that every regression analysis be accompanied by a thorough analysis of the residuals as a matter of routine "diagnostics." The principles—and mechanics—behind how the assumption (explicit or implicit) of the normality of residuals are validated systematically and rigorously is discussed in the next chapter as part of a broader discussion of probability model validation.

REVIEW QUESTIONS

1. In regression analysis, what is an independent variable and what is a dependent variable?

2. In regression analysis as discussed in this chapter, which variable is deterministic

and which is random?

3. In regression analysis, what is a "predictor" and what is a response variable?

4. Regression analysis is concerned with what tasks?

5. What is the principle of least squares?

6. In simple linear regression, what is a one-parameter model; what is a two-parameter model?

7. What are the two main assumptions underlying regression analysis?

8. In simple linear regression, what are the "normal equations" and how do they arise?

9. In simple linear regression, under what conditions are the least squares estimates identical to the maximum likelihood estimates?

10. In regression analysis, what are residuals?

11. What does it mean that OLS estimators are unbiased?

12. Why is the confidence interval around the regression line curved? Where is the interval narrowest?

13. What does hypothesis testing entail in linear regression? What is H_0 and what is H_a in this case?

14. What is the difference between using the regression line to estimate mean responses and using it to predict a new response?

15. Why are prediction intervals consistently wider than confidence intervals?

16. What is R^2 and what is its role in regression?

17. What is R^2_{adj} and what differentiates it from R^2?

18. Is an R^2 value close to 1 always indicative of a good regression model? By the same token, is an R^2_{adj} value close to 1 always indicative of a good regression model?

19. In the context of simple linear regression, what is an F-test used for?

20. What is the connection between R^2, the coefficient of determination, and the correlation coefficient?

21. If a regression model represents a data set adequately, what should we expect of the residuals?

22. What does residual analysis allow us to do?

23. What activities are involved in residual analysis?

24. What are standardized residuals?

25. Why is it recommended for residual plots to be based on standardized residuals?

26. The term "linear" in linear regression refers to what?

27. As far as regression is concerned, how does one determine whether the problem is linear or nonlinear?

28. What is an "intrinsically linear" model?

29. In employing variable transformations to convert nonlinear regression problems to linear ones, what important issue should be taken into consideration?

30. What is multiple linear regression?

31. What is the "hat" matrix and what is its role in multiple linear regression?

32. What are some reasons for using weights in regression problems?

33. What is constrained least squares and what class of problems require this approach?

34. What is ridge regression and under what condition is it recommended?

35. What is the principle behind recursive least squares?

36. What is polynomial regression?

37. What is special about orthogonal polynomial regression?

38. What is the orthogonal basis function expansion result and what are its implications for polynomial regression?

EXERCISES

16.1 Given the one-parameter model,

$$y_i = x_i\theta + \epsilon_i$$

where $\{y_i\}_{i=1}^n$ is the specific sample data set, and ϵ_i, the random error component, has zero mean and variance σ^2, it was shown in Eq (16.18) that the least squares

Chapter 17

Probability Model Validation

17.1 Introduction 728
17.2 Probability Plots 729
17.2.1 Basic Principles 729
17.2.2 Transformations and Specialized Graph Papers 730
17.2.3 Modern Probability Plots 732
17.2.4 Applications 733
Safety Data 733
Yield Data 733
Residual Analysis for Regression Model 733
Others 735
17.3 Chi-Squared Goodness-of-Fit Test 735
17.3.1 Basic Principles 735
17.3.2 Properties and Application 738
Poisson Model Validation 738
Binomial Special Case 741
17.4 Summary and Conclusions 741
REVIEW QUESTIONS 742
EXERCISES 743
APPLICATION PROBLEMS 746

An abstract analysis which is accepted
without any synthetic examination
of the question under discussion
is likely to surprise rather than enlighten us.

Daniel Bernoulli (1700–1782)

In his pithy statement, "All models are wrong but some are useful," the legendary George E. P. Box of Wisconsin was employing hyperbole to make a subtle but important point. The point, well-known to engineers, is that perfection is not a prerequisite for usefulness in modeling (in fact, it can be an impediment). If complex, real-world problems are to become tractable, idealizing assumptions are inevitable. But what is thus given up in "perfection" is more than made up for in usefulness, *so long as the assumptions can be validated as reasonable*. As a result, assessing the reasonableness of inevitable assumptions is—or ought to be—an important part of the modeling exercise; and this chapter is concerned with presenting some techniques for doing just that—validating distributional assumptions. We focus specifically on probability plots and the Chi-squared goodness-of-fit test, two time-tested techniques that also happen to complement each other perfectly in such a way that,

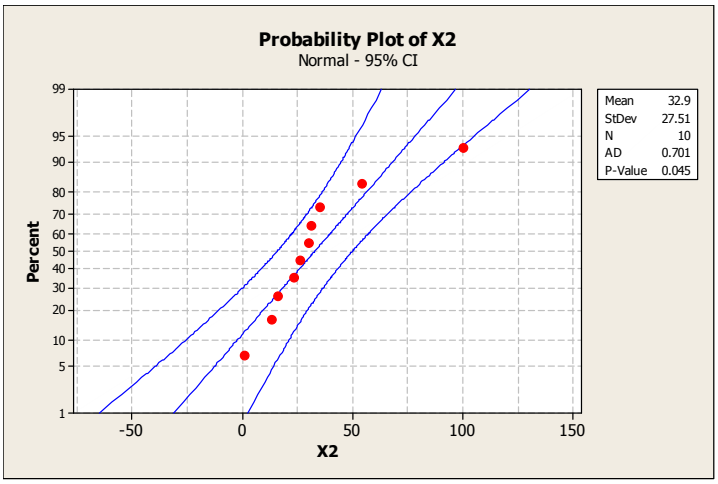

FIGURE 17.2: Probability plot for safety data S_2 wrongly postulated to be normally distributed. The departure from the linear fit does not appear too severe, but the low/borderline p-value (0.045) objectively compels us to reject H_0 at the 0.05 significance level and conclude that the Gaussian model is inadequate for this data.

in this section have been subjected. This is left to the reader as an exercise (see Exercise 17.1).

Others

In addition to the probability plots illustrated above for exponential and Gaussian distributions, MINITAB can also generate probability plots for several other distributions, including lognormal, gamma, and Weibull distributions, all continuous distributions.

Probability plots are not used for discrete probability models in part because the associated cdfs consist of a series of discontinuous "step" functions, not smooth curves like continuous random variable cdfs. To check the validity of discrete distributions such as the binomial and Poisson, it is necessary to use the more versatile technique discussed next.

17.3 Chi-Squared Goodness-of-Fit Test

17.3.1 Basic Principles

While the probability plot is fundamentally a graphically-based approach, the Chi-squared goodness-of-fit test is fundamentally computational.

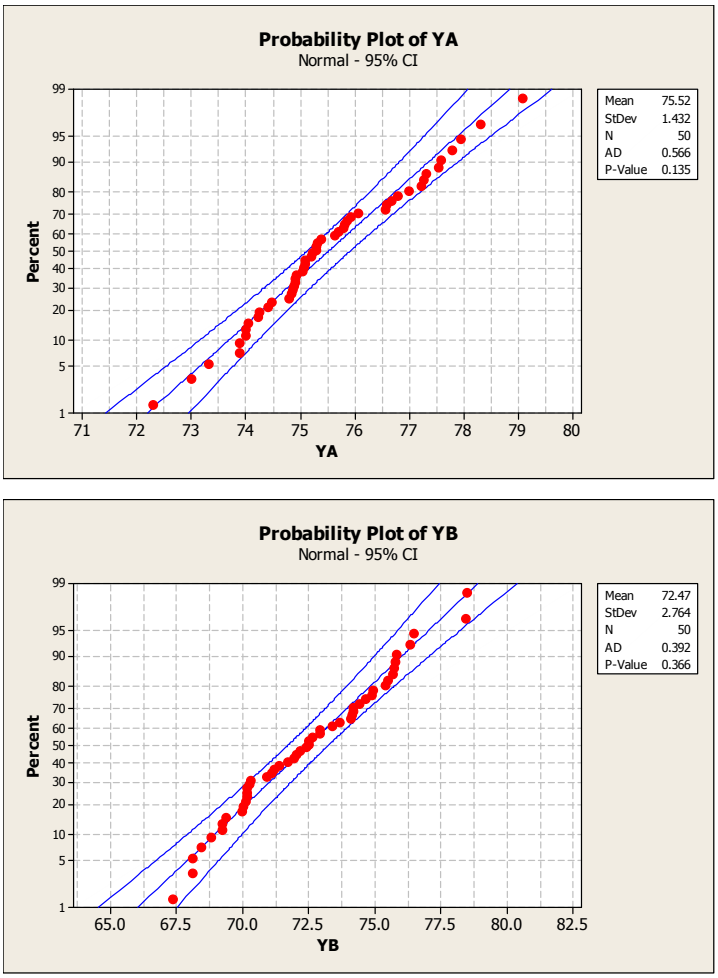

FIGURE 17.3: Probability plots for yield data sets Y_A and Y_B postulated to be normally distributed. The 95% confidence intervals around the fitted line, along with the indicated p-values, strongly suggest that the distributional assumptions appear to be valid.

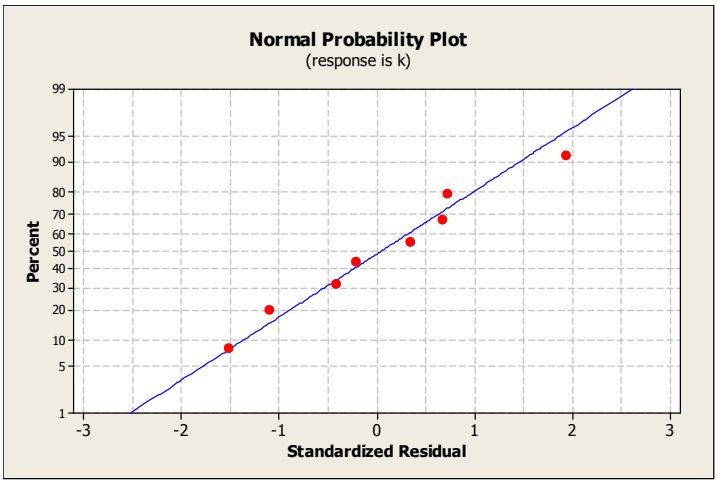

FIGURE 17.4: Normal probability plot for the residuals of the regression analysis of the dependence of thermal conductivity, k, on temperature in Example 16.5. The postulated model, a two-parameter regression model with Gaussian distributed zero mean errors, appears valid.

We begin by classifying the sample data, $x_1, x_2, \ldots, x_n$, into m groups (or "bins"), and obtaining from there a frequency distribution with f_i^o as the resulting observed frequency associated with the i^{th} group—precisely how histograms are generated (see Chapter 12). From the postulated probability model, and its p parameters estimated from the sample data, the theoretical (i.e., expected) frequency associated with each of the m groups, $\varphi_i; i = 1, 2, \ldots, m$, is then computed. If the postulated model is correct, the observed and expected frequencies should be close. Because the observed frequencies are subject to random variability, their "closeness" to the corresponding theoretical expectations, quantified by,

$$C^2 = \sum_{i=1}^{m} \frac{\left(f_i^o - \varphi_i\right)^2}{\varphi_i} \tag{17.9}$$

is a statistic that can be shown to have an approximate $\chi^2(\nu)$ distribution with $\nu = m - p - 1$ degrees of freedom—an approximation that improves rapidly with increasing n.

The Chi-squared goodness-of-fit test is a hypothesis test based on this test statistic; the null hypothesis, H_0, that the data follow the postulated probability model, is tested, at the α significance level, against the alternative that the data do not follow the model. H_0 is rejected if

$$C^2 > \chi_\alpha^2(\nu) \tag{17.10}$$

17.3.2 Properties and Application

The Chi-squared goodness-of-fit test is versatile in the sense that it can be applied to both discrete and continuous random variables. With the former, the data already occur naturally in discrete groups; with the latter, theoretical frequencies must be computed by discretizing the continuous intervals. The test is also transparent, logical (as evident from Eq (17.9)), and relatively easy to perform. However, it has some important weaknesses also:

1. To be valid, the test requires that the expected frequency associated with each bin must be at least 5. Where this is not possible, it is recommended that adjacent bins be combined appropriately. This has the drawback that the test will not be very sensitive to tails of postulated models where, by definition, expected observations are few.

2. In general, the test lacks sensitivity in detecting inadequate models when n is small.

3. Even though recommendations are available for how best to construct discrete intervals for continuous random variables, both the number as well as the nature of these discretized intervals are largely arbitrary and can (and often do) affect the outcome of the test. Therefore, even though applicable in principle, this test is not considered the best option for continuous random variables.

Poisson Model Validation

The following example illustrates the application of the Chi-squared test to the glass manufacturing data presented in Chapter 1 and revisited in various chapters including Chapters 8 (Example 8.8), Chapter 14 (Example 14.13), and Chapter 15 (Example 15.15).

Example 17.2: VALIDATING THE POISSON MODEL FOR INCLUSIONS DATA
The number of *inclusions* found in each of 60 square-meter sheets of manufactured glass and presented in Table 1.2 in Chapter 1, was postulated to be a Poisson random variable with the single parameter, λ. Perform a Chi-squared goodness-of-fit test on this data to evaluate the reasonableness of this postulate.

Solution:
Recall from our various encounters with this data set that the Poisson model parameter, estimated from the data mean, is $\hat{\lambda} = 1.017$. If the data are now arranged into frequency groups for 0, 1, 2, and 3+ inclusions, we obtain the following table:

Data Group (Inclusions)	Observed Frequency	Poisson $f(x\|\lambda)$	Expected Frequency
0	22	0.3618	21.708
1	23	0.3678	22.070
2	11	0.1870	11.219
≥ 3	4	0.0834	5.004

with the expected frequency obtained from $60 \times f(x|\lambda)$. We may now compute the desired C^2 statistic as:

$$\begin{aligned} C^2 &= \frac{(22-21.708)^2}{21.708} + \frac{(23-22.070)^2}{22.070} + \frac{(11-11.219)^2}{11.219} + \frac{(4-5.004)^2}{5.004} \\ &= 0.249 \end{aligned} \tag{17.11}$$

The associated degrees of freedom is $4 - 1 - 1 = 2$, so that from the $\chi^2(2)$ distribution, we obtain

$$P(\chi^2(2) > 0.249) = 0.883 \tag{17.12}$$

As a result, we have no evidence to reject the null hypothesis, and hence conclude that the Poisson model for this data set appears adequate.

Of course, it is unnecessary to carry out any of the indicated computations by hand, even the frequency grouping. Programs such as MINITAB have Chi-squared test features that can be used for problems of this kind.

When MINITAB is used on this last example, upon entering the raw data into a column labeled "Inclusions," the sequence, `Stat > Basic Stats > Chi-Sq Goodness-of-Fit-Test for Poisson` opens a self-explanatory dialog box; and making the required selections produces the following results, just as we had obtained earlier:

Goodness-of-Fit Test for Poisson Distribution

```
Data column: Inclusions
Poisson mean for Inclusions = 1.01667
```

Inclusions	Observed	Poisson Probability	Expected	Contribution to Chi-Sq
0	22	0.361799	21.7079	0.003930
1	23	0.367829	22.0697	0.039212
2	11	0.186980	11.2188	0.004267
≥ 3	4	0.083392	5.0035	0.201279

N	DF	Chi-Sq	P-Value
60	2	0.248687	0.883

The right-most column in the MINITAB output shows the individual contributions from each group to the Chi-squared statistic—an indication of how the "lack-of-fit" is distributed among the groups. For example, the group of 3 or more inclusions contributed by far the largest to the discrepancy between observation and model prediction; but even this is not sufficient to jeopardize the model adequacy. MINITAB also produces graphical representations of these results, as shown in Fig 17.5.

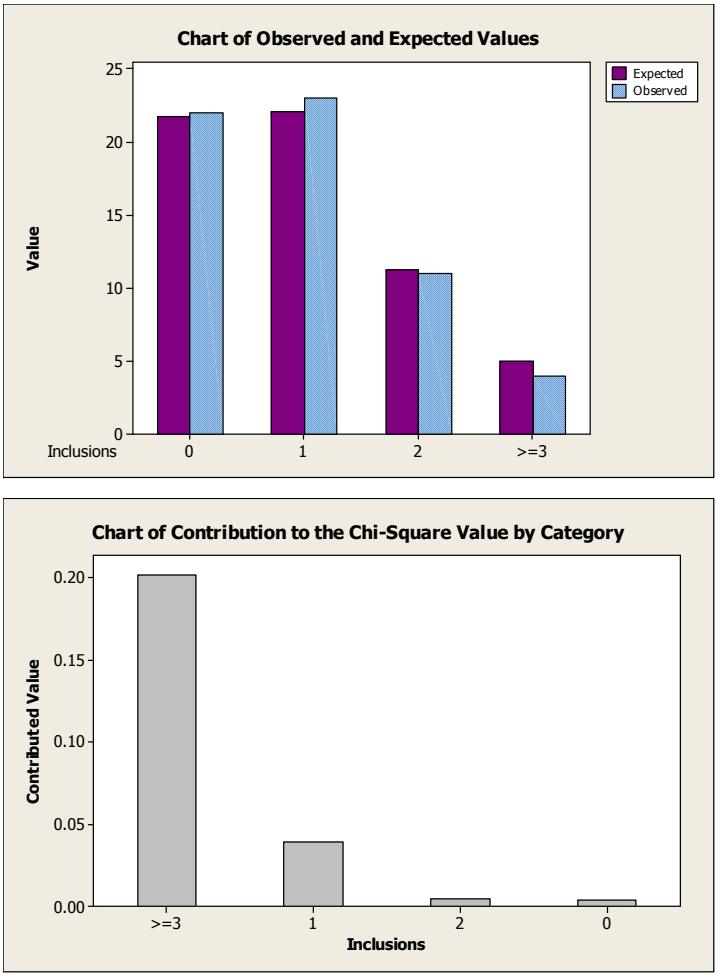

FIGURE 17.5: Chi-squared test results for inclusions data and a postulated Poisson model. Top panel: Bar chart of "Expected" and "Observed" frequencies, which shows how well the model prediction matches observed data; Bottom panel: Bar chart of contributions to the Chi-squared statistic, showing that the group of 3 or more inclusions is responsible for the largest model-observation discrepancy, by a wide margin.

Binomial Special Case

For the binomial case, where there are only two categories (x "successes" and $n - x$ "failures" observed in n independent trials being the observed frequencies in each respective category), for a postulated $Bi(n, p)$ model, the Chi-squared statistic reduces to:

$$C^2 = \frac{(x - np)^2}{np} + \frac{[(n - x) - nq]^2}{nq} \tag{17.13}$$

where $q = 1 - p$, as usual. When this expression is consolidated to

$$C^2 = \frac{q(x - np)^2 + p[(n - x) - nq]^2}{npq}$$

upon introducing $q = 1 - p$ for the first term in the numerator and taking advantage of the "difference of two squares" result in algebra, the right-hand side of the equation rearranges easily to give the result:

$$C^2 = \frac{(x - np)^2}{npq} \tag{17.14}$$

which, if we take the positive square root, reduces to:

$$Z = \frac{x - np}{\sqrt{npq}} \tag{17.15}$$

This, of course, is immediately recognized as the z-statistic for the (large sample) Gaussian approximation to the binomial random variable used to carry out the z-test of the observed mean against a postulated mean np, as discussed in Chapter 15. Thus, the Chi-squared test for the binomial model is identical to the standard z-test when the population parameter p is specified independently.

17.4 Summary and Conclusions

This chapter has been primarily concerned with examining two methods for validating probability models: modern probability plots and the Chi-squared goodness-of-fit test. While we presented the principles behind these methods, we concentrated more on applying them, particularly with the aid of computer programs. With some perspective, we may now observe the following as the main points of the chapter:

- Probability plots augmented with theoretical model fits and p-values are most appropriate for continuous models.

- Chi-squared tests, on the other hand, are more naturally suited to discrete models (although they can also be applied to continuous models after appropriate discretization).

As a practical matter, it is important to keep in mind that, just as with other hypotheses tests, a postulated probability model can never be completely *proven* adequate by these tests (on the basis of finite sample data), but inadequate models can be successfully identified as such. Still, it can be difficult to identify inadequate models with these tests when sample sizes are small; our chances of identifying inadequate models correctly as inadequate improve significantly as $n \to \infty$. Therefore, as much sample data as possible should be used to validate probability models; and wherever possible, the data set used to validate a model should be collected independently of that used to estimate the parameters. Some of the end-of-chapter exercises and application problems are used to reinforce these points.

Finally, it must be kept in mind always that no model is (or can ever be) perfect. The final decision about the validity of the model assumptions rests with the practitioner—the person who will ultimately use these models for problem solving—and these tests should be considered properly only as objective guides, not as final and absolute arbiters.

REVIEW QUESTIONS

1. What is the primary question of interest in probability model validation?

2. What are the two approaches discussed in this chapter for validating probability models?

3. Which approach is better suited to continuous probability models and which one is applicable most directly to discrete probability models?

4. What is the fundamental principle behind probability plots?

5. What is the fundamental concept behind old-fashioned probability plots?

6. What hypothesis test accompanies modern probability plots?

7. What does a modern probability plot consist of?

8. Why are probability plots not used for discrete probability models?

9. What is a Chi-squared goodness-of-fit test?