UNIT I

INFERENTIAL STATISTICS I

Populations – samples – random sampling – probability and statistics Sampling distribution – creating a sampling distribution – mean of all sample means – standard error of the mean – other sampling distributions Hypothesis testing – z-test – z-test procedure – statement of the problem – null hypothesis – alternate hypotheses – decision rule – calculations – decisions - interpretations

POPULATIONS

Any complete set of observations (or potential observations) may be characterized as a population. Accurate descriptions of populations specify the nature of the observations to be taken. For example, a population might be described as "attitudes toward abortion of currently enrolled students at Bucknell University" or as "SAT critical reading scores of currently enrolled students at Rutgers University."

REAL POPULATION

Pollsters, such as the Gallup Organization, deal with real populations. A real population is one in which all potential observations are accessible at the time of sampling. Examples of real populations include the two described in the previous paragraph, as well as the ages of all visitors to Disneyland on a given day, the ethnic backgrounds of all current employees of the U.S. Postal Department, and presidential preferences of all currently registered voters in the United States. Incidentally, federal law requires that a complete survey be taken every 10 years of the real population of all U.S. households—at considerable expense, involving thousands of data collectors—as a means of revising election districts for the House of Representatives. (An estimated undercount of millions of people, particularly minorities, in both the 2000 and 2010 censuses has revived a suggestion, long endorsed by statisticians, that the entire U.S. population could be estimated more accurately if a highly trained group of data collectors focused only on a random sample of households.

HYPOTHETICAL POPULATION

Insofar as research workers concern themselves with populations, they often invoke the notion of a hypothetical population. A hypothetical population is one in which all potential observations are not accessible at the time of sampling. In most experiments, subjects are selected from very small, uninspiring real populations: the lab rats housed in the local animal colony or student volunteers from general psychology classes. Experimental subjects often are viewed, nevertheless, as a sample from a much larger hypothetical population, loosely described as "the scores of all similar animal subjects (or student volunteers) who could conceivably undergo the present experiment." According to the rules of inferential statistics, generalizations should be made only to real populations that, in fact, have been

sampled. Generalizations to hypothetical populations should be viewed, therefore, as provisional conclusions based on the wisdom of the researcher rather than on any logical or statistical necessity. In effect, it's an open question—often answered only by additional experimentation— whether or not a given experimental finding merits the generality assigned to it by the researcher

***SAMPLES**

Any subset of observations from a population may be characterized as a sample. In typical applications of inferential statistics, the sample size is small relative to the population size. For example

less than 1 percent of all U.S. worksites are included in the Bureau of Labour Statistics' monthly survey to estimate the rate of unemployment. And although, only 1475 likely voters had been sampled in the final poll for the 2012 presidential election by the NBC News/Wall Street Journal

OPTIMAL SAMPLE SIZE

There is no simple rule of thumb for determining the best or optimal sample size for any particular situation. Often sample sizes are in the hundreds or even the thousands for surveys, but they are less than 100 for most experiments. Optimal sample size depends on the answers to a number of questions, including "What is the estimated variability among observations?" and "What is an acceptable amount of error in our conclusion?" Once these types of questions have been answered, with the aid of guidelines such as those discussed in Section 11.11, specific procedures can be followed to determine the optimal sample size for any situation.

***RANDOM SAMPLINGS:**

The valid use of techniques from inferential statistics requires that samples be random.

Random sampling occurs if, at each stage of sampling, the selection process guarantees that all potential observations in the population have an equal chance of being included in the sample

It's important to note that randomness describes the selection process—that is, the conditions under which the sample is taken—and not the particular pattern of observations in the sample. Having established that sampling is random, you still can't predict anything about the unique pattern of observations in that sample. The observations in the sample should be representative of those in the population, but there is no guarantee that they actually will be

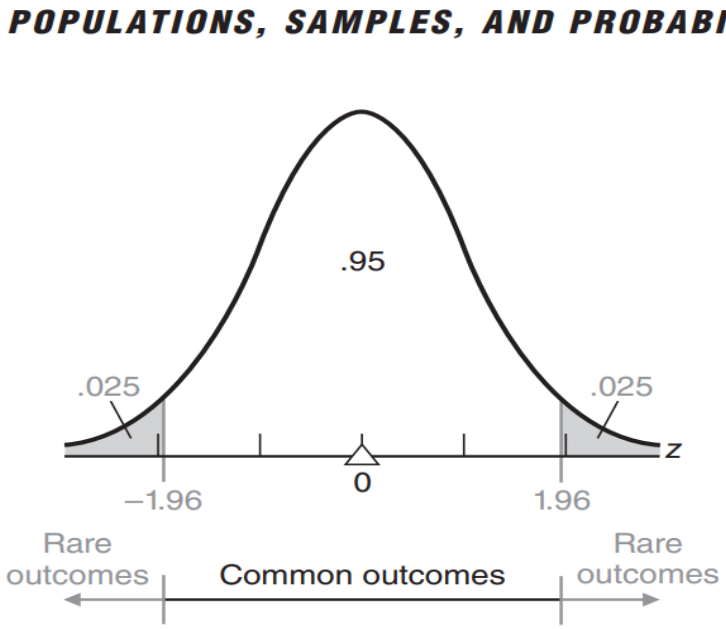
CASUAL OR HAPHAZARD NOT RANDOM

A casual or haphazard sample doesn't qualify as a random sample. Not every student at UC San Diego has an equal chance of being sampled if, for instance, a pollster casually selects only students

who enter the student union. Obviously excluded from this sample are all those students (few, we hope) who never enter the student union. Even the final selection of students from among those who do enter the student union might reflect the pollster's various biases, such as an unconscious preference for attractive students who are walking alone

***PROBABILITY AND STATISTICS**

Probability assumes a key role in inferential statistics including, for instance, the important area known as hypothesis testing. Because of the inevitable variability that accompanies any observed result, such as a mean difference between two groups, its value must be viewed within the context of the many possible results that could have occurred just by chance. With the aid of some theoretical curve, such as the normal curve, and a provisional assumption, known as the null hypothesis, that chance can reasonably account for the result, probabilities are assigned to the one observed mean difference. If this probability is very small, the result is viewed as a rare outcome, and we conclude that something real—that is, something that can't reasonably be attributed to chance—has occurred. On the other hand, if this probability isn't very small, the result is viewed as a common outcome, and we conclude that something transitory—that is, something that can reasonably be attributed to chance—has occurred.



COMMON OUTCOMES:

Common outcomes signify, most generally, a lack of evidence that something special has occurred. For instance, they suggest that the observed mean difference— whatever its value—might signify that the true mean difference could equal zero and, therefore, that any comparable study would just as

likely produce either a positive or negative mean difference. Therefore, the observed mean difference should not be taken seriously because, in the language of statistics, it lacks statistical significance.

RARE OUTCOMES:

On the other hand, rare outcomes signify that something special has occurred. For instance, they suggest that the observed mean difference probably signifies a true mean difference equal to some nonzero value and, therefore, that any comparable study would most likely produce a mean difference with the same sign and a value in the neighborhood of the one originally observed. Therefore, the observed mean difference should be taken seriously because it has statistical significance

COMMON RARE:

As an aid to determining whether observed results should be viewed as common or rare, statisticians interpret different proportions of area under theoretical curves, such as the normal curve shown in Figure 8.2, as probabilities of random outcomes. For instance, the standard normal table indicates that .9500 is the proportion of total area between z scores of –1.96 and +1.96. (Verify this proportion by referring to Table A in Appendix C and, if necessary, to the latter part of Section 5.6.) Accordingly, the probability of a randomly selected z score anywhere between ±1.96 equals .95. Because it should happen about 95 times out of 100, this is often designated as a common event signifying that, once variability is considered, nothing special is happening. On the other hand, since the standard normal curve indicates that .025 is the proportion of total area above a z score of +1.96, and also that .025 is the proportion of total area below a z score of –1.96, then the probability of a randomly selected z score anywhere beyond either +1.96 or –1.96 equals .05 (from .025 + .025, thanks to the addition rule). Because it should happen only about 5 times in 100, this is often designated as a rare outcome signifying that something special is happening. At this point, you're not expected to understand the rationale behind this perspective, but merely that, once identified with a particular result, a specified sector of area under a curve will be interpreted as the probability of that outcome. Furthermore, since the probability of an outcome has important implications for generalizing beyond actual results, probabilities play a key role in inferential statistics

***SAMPLING DISTRUBTION:**

There's a good chance that you've taken the SAT test, and you probably remember your scores. Assume that the SAT math scores for all college-bound students during a recent year were distributed around a mean of 500 with a standard deviation of 110. An investigator at a university wishes to test the claim that, on the average, the SAT math scores for local freshmen equals the national average of

500. His task would be straightforward if, in fact, the math scores for all local freshmen were readily available. Then, after calculating the mean score for all local freshmen, a direct comparison would indicate whether, on the average, local freshmen score below, at, or above the national average. Assume that it is not possible to obtain scores for the entire freshman class. Instead, SAT math scores are obtained for a random sample of 100 students from the local population of freshmen, and the mean score for this sample equals 533. If each sample were an exact replica of the population, generalizations from the sample to the population would be most straightforward. Having observed a mean score of 533 for a sample of 100 freshmen, we could have concluded, without even a pause, that the mean math score for the entire freshman class also equals 533 and, therefore, exceeds the national average.

What Is sampling Distrubtion ?

Random samples rarely represent the underlying population exactly. Even a mean math score of 533 could originate, just by chance, from a population of freshmen whose mean equals the national average of 500. Accordingly, generalizations from a single sample to a population are much more tentative. Indeed, generalizations are based not merely on the single sample mean of 533 but also on its distribution—a distribution of sample means for all possible random samples. Representing the statistician's model of random outcomes

the sampling distribution of the mean refers to the probability distribution of means for all possible random samples of a given size from some population.

In effect, this distribution describes the variability among sample means that could occur just by chance and thereby serves as a frame of reference for generalizing from a single sample mean to a population mean

The sampling distribution of the mean allows us to determine whether, given the variability among all possible sample means, the one observed sample mean can be viewed as a common outcome or as a rare outcome (from a distribution centered in this case, about a value of 500). If the sample mean of 533 qualifies as a common outcome in this sampling distribution, then the difference between 533 and 500 isn't large enough, relative to the variability of all possible sample means, to signify that anything special is happening in the underlying population. Therefore, we can conclude that the mean math score for the entire freshman class could be the same as the national average of 500. On the other hand, if the sample mean of 533 qualifies as a rare outcome in this sampling distribution, then the difference between 533 and 500 is large enough, relative to the variability of all possible sample means, to signify that something special probably is happening in the underlying population. Therefore, we can conclude that the mean math score for the entire freshman class probably exceeds the national average of 500.

All Possible Random Samples

When attempting to generalize from a single sample mean to a population mean, we must consult the sampling distribution of the mean. In the present case, this distribution is based on all possible random samples, each of size 100 that can be taken from the local population of freshmen. All possible random samples refer not to the number of samples of size 100 required to survey completely the local population of freshmen but to the number of different ways in which a single sample of size 100 can be selected from this population

"All possible random samples" tends to be a huge number. For instance, if the local population contained at least 1,000 freshmen, the total number of possible random samples, each of size 100, would be astronomical in size. The 301 digits in this number would dwarf even the national debt. Even with the aid of a computer, it would be a horrendous task to construct this sampling distribution from scratch, itemizing each mean for all possible random samples.

Fortunately, statistical theory supplies us with considerable information about the sampling distribution of the mean, as will be discussed in the remainder of this chapter. Armed with this information about sampling distributions, we'll return to the current example in the next chapter and test the claim that the mean math score for the local population of freshmen equals the national average of 500. Only at that point—and not at the end of this chapter—should you expect to understand completely the role of sampling distributions in practical applications.

***CREATING A SAMPLE DISTRUBTION:**

Let's establish precisely what constitutes a sampling distribution by creating one from scratch under highly simplified conditions. Imagine some ridiculously small population of four observations with values of 2, 3, 4, and 5, as shown in Figure 9.1. Next, itemize all possible random samples, each of size two, that could be taken from this population. There are four possibilities on the first draw from the population and also four possibilities on the second draw from the population, as indicated in Table 9.1.* The two sets of possibilities combine to yield a total of 16 possible samples. At this point, remember, we're clarifying the notion of a sampling distribution of the mean. In practice, only a single random sample, not 16 possible samples, would be taken from the population; the sample size would be very small relative to a much larger population size, and, of course, not all observations in the population would be known. For each of the 16 possible samples, Table 9.1 also lists a sample mean (found by adding the two observations and dividing by 2) and its probability of occurrence (expressed as $1/16$, since each of the 16 possible samples is equally likely). When cast into a relative frequency or probability distribution, as in Table 9.2, the 16 sample means constitute the sampling distribution of the mean, previously defined as the probability distribution of means for all possible random samples of a given size from some population. Not all values of the sample mean occur with equal probabilities in Table 9.2 since some values occur more than once among the 16 possible

samples. For instance, a sample mean value of 3.5 appears among 4 of 16 possibilities and has a probability of 4/16

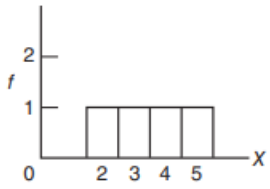


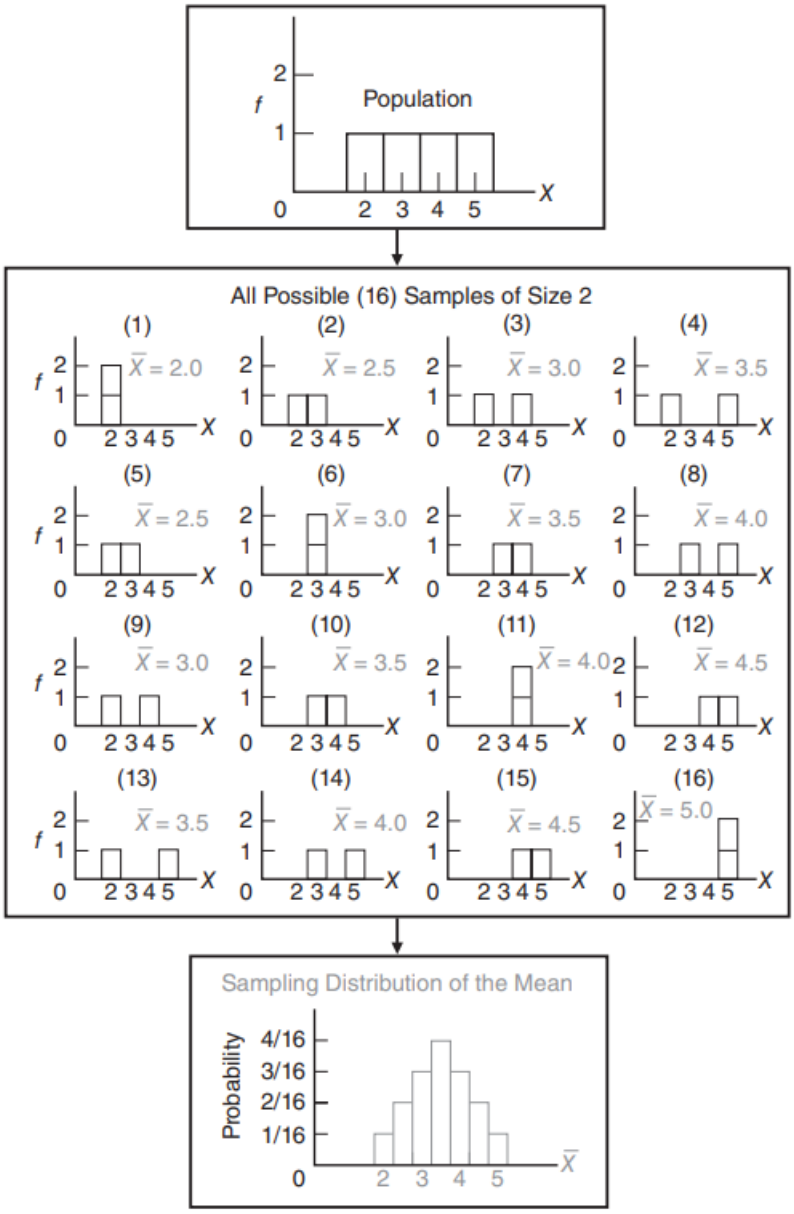
FIGURE 9.1
Graph of a miniature population.

Table 9.1
ALL POSSIBLE SAMPLES OF SIZE TWO FROM A MINIATURE POPULATION

	ALL POSSIBLE SAMPLES	MEAN ($\overline{X}$)	PROBABILITY
(1)	2,2	2.0	$\frac{1}{16}$
(2)	2,3	2.5	$\frac{1}{16}$
(3)	2,4	3.0	$\frac{1}{16}$
(4)	2,5	3.5	$\frac{1}{16}$
(5)	3,2	2.5	$\frac{1}{16}$
(6)	3,3	3.0	$\frac{1}{16}$
(7)	3,4	3.5	$\frac{1}{16}$
(8)	3,5	4.0	$\frac{1}{16}$
(9)	4,2	3.0	$\frac{1}{16}$
(10)	4,3	3.5	$\frac{1}{16}$
(11)	4,4	4.0	$\frac{1}{16}$
(12)	4,5	4.5	$\frac{1}{16}$
(13)	5,2	3.5	$\frac{1}{16}$
(14)	5,3	4.0	$\frac{1}{16}$
(15)	5,4	4.5	$\frac{1}{16}$
(16)	5,5	5.0	$\frac{1}{16}$

Table 9.2 SAMPLING DISTRIBUTION OF THE MEAN (SAMPLES OF SIZE TWO FROM A MINIATURE POPULATION)

SAMPLE MEAN ($\bar{X}$)	PROBABILITY
5.0	$\frac{1}{16}$
4.5	$\frac{2}{16}$
4.0	$\frac{3}{16}$
3.5	$\frac{4}{16}$
3.0	$\frac{3}{16}$
2.5	$\frac{2}{16}$
2.0	$\frac{1}{16}$



***MEAN OF ALL THE SAMPLE MEAN**

The distribution of sample means itself has a mean.

The mean of the sampling distribution of the mean always equals the mean of the population.

SAMPLING DISTRIBUTION OF THE MEAN

Expressed in symbols, we have

MEAN OF SAMPLING DISTRUBTION

$$\mu_{\bar{X}} = \mu$$

where μX represents the mean of the sampling distribution and μ represents the mean of the population.

INTERCHANGABLE MEANS:

Since the mean of all sample means (μX) always equals the mean of the population (μ), these two terms are interchangeable in inferential statistics. Any claims about the population mean can be transferred directly to the mean of the sampling distribution, and vice versa. If, as claimed, the mean math score for the local population of freshmen equals the national average of 500, then the mean of the sampling distribution also automatically will equal 500. For the same reason, it's permissible to view the one observed sample mean of 533 as a deviation either from the mean of the sampling distribution or from the mean of the population. It should be apparent, therefore, that whether an expression involves μX or μ, it reflects, at most, a difference in emphasis on either the sampling distribution or the population, respectively, rather than any difference in numerical value

***STANDARD ERROR OF THE MEAN**

The distribution of sample means also has a standard deviation, referred to as the standard error of the mean.

The standard error of the mean equals the standard deviation of the population divided by the square root of the sample size

Expressed in symbols,

STANDARD ERROR OF THE MEAN

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad (9.2)$$

Where doe sy? X represents the standard error of the mean; σ represents the standard deviation of the population; and n represents the sample size.

SPECIAL TYPE OF STANDARD DEVIATION

The standard error of the mean serves as a special type of standard deviation that measures variability in the sampling distribution. It supplies us with a standard, much like a yardstick, that describes the

amount by which sample means deviate from the mean of the sampling distribution or from the population mean. The error in standard error refers not to computational errors, but to errors in generalizations attributable to the fact that, just by chance, most random samples aren't exact replicas of the population.

You might find it helpful to think of the standard error of the mean as a rough measure of the average amount by which sample means deviate from the mean of the sampling distribution or from the population mean.

Insofar as the shape of the distribution sample means approximates a normal curve, as described in the next section, about 68 percent of all sample means deviate less than one standard error from the mean of the sampling distribution, whereas only about 5 percent of all sample means deviate more than two standard errors from the mean of this distribution.

Effect of Sample Size

A most important implication of Formula 9.2 is that whenever the sample size equals two or more, the variability of the sampling distribution is less than that in the population. A modest demonstration of this effect appears in Figure 9.2, where the means of all possible samples cluster closer to the population mean (equal to 3.5) than do the four original observations in the population. A more dramatic demonstration occurs with larger sample sizes. Earlier in this chapter, for instance, 110 was given as the value of σ, the population standard deviation for SAT scores. Much smaller is the variability in the sampling distribution of mean SAT scores, each based on samples of 100 freshmen. According to Formula 9.2, in the present example,

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = \frac{110}{10} = 11$$

there is a tenfold reduction in variability, from 110 to 11, when our focus shifts from the population to the sampling distribution. According to Formula 9.2, any increase in sample size translates into a smaller standard error and, therefore, into a new sampling distribution with less variability. With a larger sample size, sample means cluster more closely about the mean of the sampling distribution and about the mean of the population and, therefore, allow more precise generalizations from samples to populations

SAMPLING DISTRIBUTION OF THE MEAN

It's not surprising that variability should be smaller in sampling distributions than in populations. The population standard deviation reflects variability among individual observations, and it is directly

affected by any relatively large or small observations within the population. On the other hand, the standard error of the mean reflects variability among sample means, each of which represents a collection of individual observations. The appearance of relatively large or small observations within a particular sample tends to affect the sample mean only slightly, because of the stabilizing presence in the same sample of other, more moderate observations or even extreme observations in the opposite direction. This stabilizing effect becomes even more pronounced with larger sample sizes.

***OTHER SAMPLING DISTRIBUTIONS**

For the Mean

There are many different sampling distributions of means. A new sampling distribution is created by a switch to another population. Furthermore, for any single population, there are as many different sampling distributions as there are possible sample sizes. Although each of these sampling distributions has the same mean, the value of the standard error always differs and depends upon the size of the sample.

For Other Measures

There are sampling distributions for measures other than a single mean. For instance, there are sampling distributions for medians, proportions, standard deviations, variances, and correlations, as well as for differences between pairs of means, pairs of proportions, and so forth. We'll have occasion to work with some of these distributions in later chapters

Important Terms

Mean of the sampling distribution of the mean ($\mu_{\bar{X}}$)
Sampling distribution of the mean
Standard error of the mean ($\sigma_{\bar{X}}$)
Central limit theorem

Key Equations

SAMPLING DISTRIBUTION MEAN

$$\mu_{\bar{X}} = \mu$$

STANDARD ERROR

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

***HYPOTHESIS TESTING**

In the previous chapter, we postponed a test of the hypothesis that the mean SAT math score for all local freshmen equals the national average of 500. Now, given a mean math score of 533 for a random sample of 100 freshmen, let's test the hypothesis that, with respect to the national average, nothing special is happening in the local population. Insofar as an investigator usually suspects just the opposite—namely, that something special is happening in the local population—he or she hopes to reject the hypothesis that nothing special is happening, henceforth referred to as the null hypothesis and defined more formally in a later section

Hypothesized Sampling Distribution

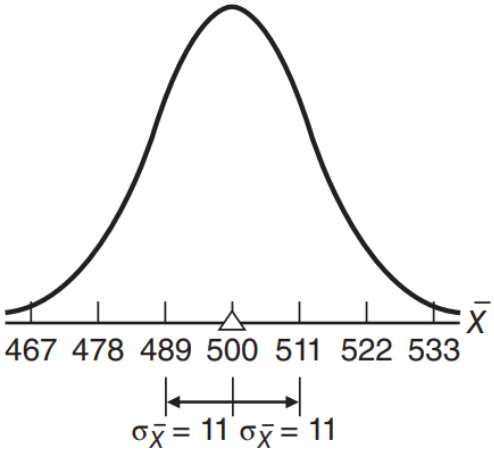
If the null hypothesis is true, then the distribution of sample means—that is, the sampling distribution of the mean for all possible random samples, each of size 100, from the local population of freshmen—will be centered about the national average of 500. (Remember, the mean of the sampling distribution always equals the population mean.) In Figure 10.1, this sampling distribution is referred to as the hypothesized sampling distribution, since its mean equals 500, the hypothesized mean reading score for the local population of freshmen.

Anticipating the key role of the hypothesized sampling distribution in our hypothesis test, let's focus on two more properties of this distribution

1. In Figure 10.1, vertical lines appear, at intervals of size 11, on either side of the hypothesized population mean of 500. These intervals reflect the size of the standard error of the mean, X . To verify this fact, originally demonstrated in Chapter 9, substitute 110 for the population standard deviation, σ, and 100 for the sample size, n, in Formula 9.2 to obtain

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = \frac{110}{10} = 11$$

2. Notice that the shape of the hypothesized sampling distribution in Figure 10.1 approximates a normal curve, since the sample size of 100 is large enough to satisfy the requirements of the central limit theorem. Eventually, with the aid of normal curve tables, we will be able to construct boundaries for common and rare outcomes under the null hypothesis.



***Z TEST FOR A POPULATION MEAN**

For the hypothesis test with SAT math scores, it is customary to base the test not on the hypothesized sampling distribution of X shown in Figure 10.2, but on its standardized counterpart, the hypothesized sampling distribution of z shown in Figure 10.3. Now z represents a variation on the familiar standard score, and it displays all of the properties of standard scores described in Chapter 5. Furthermore, like the sampling distribution of X, the sampling distribution of z represents the distribution of z values that would be obtained if a value of z were calculated for each sample mean for all possible random samples of a given size from some population

The conversion from X to z yields a distribution that approximates the standard normal curve in Table A of Appendix C, since, as indicated in Figure 10.3, the original hypothesized population mean

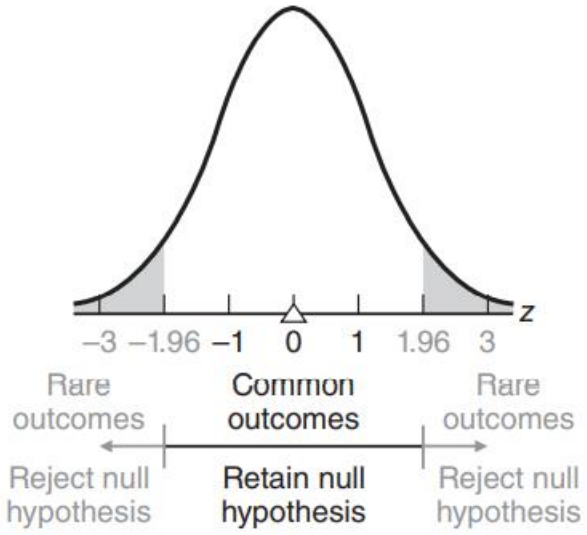
(500) emerges as a z score of 0 and the original standard error of the mean (11) emerges as a z score of 1. The shift from X to z eliminates the original units of measurement and standardizes the hypothesis test across all situations without, however, affecting the test results

: Converting a Raw Score to Z

To convert a raw score into a standard score (also described in Chapter 5), express the raw score as a distance from its mean (by subtracting the mean from the raw score), and then split this distance into standard deviation units (by dividing with the standard deviation). Expressing this definition as a word formula, we have

$$Standard\ score = \frac{raw\ score - mean}{standard\ deviation}$$

in which, of course, the standard score indicates the deviation of the raw score in standard deviation units, above or below the mean.



Converting a Sample Mean to Z

The z for the present situation emerges as a slight variation of this word formula: Replace the raw score with the one observed sample mean X; replace the mean with the mean of the sampling distribution, that is, the hypothesized population mean μ hy and replace the standard deviation with the standard error of the mean x. Now

z RATIO FOR A SINGLE POPULATION MEAN

$$z = \frac{\bar{X} - \mu_{\text{hyp}}}{\sigma_{\bar{x}}}$$

where z indicates the deviation of the observed sample mean in standard error units, above or below the hypothesized population mean.

To test the hypothesis for SAT scores, we must determine the value of z from Formula 10.1. Given a sample mean of 533, a hypothesized population mean of 500, and a standard error of 11, we find

$$z = \frac{533 - 500}{11} = \frac{33}{11} = 3$$

The observed z of 3 exceeds the value of 1.96 specified in the hypothesized sampling distribution in Figure 10.3. Thus, the observed z qualifies as a rare outcome under the null hypothesis, and the null hypothesis is rejected. The results of this test with z are the same as those for the original hypothesis test with X

Assumptions of z Test

When a hypothesis test evaluates how far the observed sample mean deviates, in standard error units, from the hypothesized population mean, as in the present example, it is referred to as a z test or, more accurately, as a z test for a population mean. This z test is accurate only when (1) the population is normally distributed or the sample size is large enough to satisfy the requirements of the central limit theorem and (2) the population standard deviation is known. In the present example, the z test is appropriate because the sample size of 100 is large enough to satisfy the central limit theorem and the population standard deviation is known to be 110.

***S T E P - B Y- S T E P P R O C E D U R E**

Having been exposed to some of the more important features of hypothesis testing, let's take a detailed look at the test for SAT scores. The test procedure lends itself to a step-by-step description, beginning with a brief statement of the problem that inspired the test and ending with an interpretation of the test results. The following box summarizes the step-by-step procedure for the current hypothesis test. Whenever appropriate, this format will be used in the remainder of the book. Refer to it while reading the remainder of the chapter.

***STATEMENT OF THE RESEARCH PROBLEM**

The formulation of a research problem often represents the most crucial and exciting phase of an investigation. Indeed, the mark of a skillful investigator is to focus on an important research problem that can be answered. Do children from broken families score lower on tests of personal adjustment? Do aggressive TV cartoons incite more disruptive behavior in preschool children? Does profit sharing increase the productivity of employees? Because of our emphasis on hypothesis testing, research problems appear in this book as finished products, usually in the first one or two sentences of a new example.

HYPOTHESIS TEST SUMMARY: *z* TEST FOR A POPULATION MEAN (SAT SCORES)

Research Problem

Does the mean SAT math score for all local freshmen differ from the national average of 500?

Statistical Hypotheses

$$H_0 : \mu = 500$$
$$H_1 : \mu \neq 500$$

Decision Rule

Reject H_0 at the .05 level of significance if $z \geq 1.96$ or if $z \leq -1.96$.

Calculations

Given

$$\bar{X} = 533;\ \mu_{hyp} = 500;\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{110}{\sqrt{100}} = 11$$

$$z = \frac{533 - 500}{11} = 3$$

Decision

Reject H_0 at the .05 level of significance because $z = 3$ exceeds 1.96.

Interpretation

The mean SAT math score for all local freshmen does not equal—it exceeds—the national average of 500.

***NULL HYPOTHESIS (H0)**

Once the problem has been described, it must be translated into a statistical hypothesis regarding some population characteristic. Abbreviated as H0 , the null hypothesis becomes the focal point for the entire test procedure (even though we usually hope to reject it). In the test with SAT scores, the null hypothesis asserts that, with respect to the national average of 500, nothing special is happening to the mean score for the local population of freshmen. An equivalent statement, in symbols, reads:

$$H0 : \mu = 500$$

where H0 represents the null hypothesis and μ is the population mean for the local freshman class

Generally speaking, the null hypothesis (H0) is a statistical hypothesis that usually asserts that nothing special is happening with respect to some characteristic of the underlying population. Because the hypothesis testing procedure requires that the hypothesized sampling distribution of the mean be centered about a single number (500), the null hypothesis equals a single number (H0 : μ 500). Furthermore, the null hypothesis always makes a precise statement about a characteristic of the population, never about a sample. Remember, the purpose of a hypothesis test is to determine whether a particular outcome, such as an observed sample mean, could have reasonably originated from a population with the hypothesized charact

Finding the Single Number for H0

The single number actually used in H0 varies from problem to problem. Even for a given problem, this number could originate from any of several sources. For instance, it could be based on available information about some relevant population other than the target population, as in the present example in which 500 reflects the mean SAT math scores for all college-bound students during a recent year. It also could be based on some existing standard or theory—for example, that the mean math score for the current population of local freshmen should equal 540 because that happens to be the mean score achieved by all local freshmen during recent years.

If, as sometimes happens, it's impossible to identify a meaningful null hypothesis, don't try to salvage the situation with arbitrary numbers. Instead, use another entirely different technique, known as estimation

ALTERNATIVE HYPOTHESIS (H1)

n the present example, the alternative hypothesis asserts that, with respect to the national average of 500, something special is happening to the mean math score for the local population of freshmen

(because the mean for the local population doesn't equal the national average of 500). An equivalent statement, in symbols, reads

$$H1 : \mu \neq 500$$

where H1 represents the alternative hypothesis, μ is the population mean for the local freshman class, and signifies, "is not equal to."

The alternative hypothesis (H1) asserts the opposite of the null hypothesis. A decision to retain the null hypothesis implies a lack of support for the alternative hypothesis, and a decision to reject the null hypothesis implies support for the alternative hypothesis.

As will be described in the next chapter, the alternative hypothesis may assume any one of three different forms, depending on the perspective of the investigator. In its present form, H1 specifies a range of possible values about the single number (500) that appears in H0 .

***D E C I S I O N R U L E**

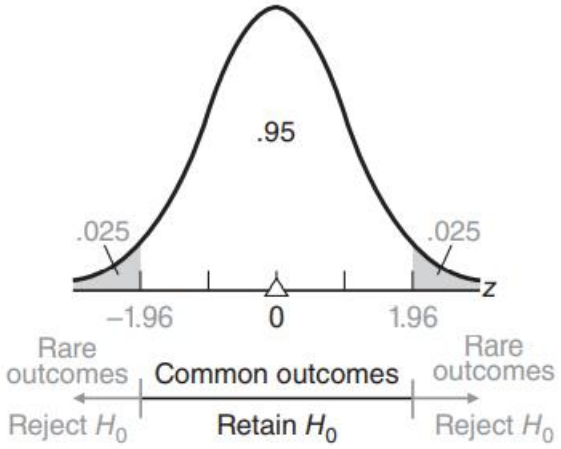
A decision rule specifies precisely when H0 should be rejected (because the observed z qualifies as a rare outcome). There are many possible decision rules, as will be seen in Section 11.3. A very common one, already introduced in Figure 10.3, specifies that H0 should be rejected if the observed z equals or is more positive than 1.96 or if the observed z equals or is more negative than –1.96. Conversely, H0 should be retained if the observed z falls between ± 1.96.

Critical z Scores

indicates that z scores of ± 1.96 define the boundaries for the middle .95 of the total area (1.00) under the hypothesized sampling distribution for z. Derived from the normal curve table, as you can verify by checking Table A in Appendix C, these two z scores separate common from rare outcomes and hence dictate whether H0 should be retained or rejected. Because of their vital role in the decision about H0 , these scores are referred to as critical z scores.

Level of Significance (α)

total area that is identified with rare outcomes. Often referred to as the level of significance of the statistical test, this proportion is symbolized by the Greek letter α (alpha) and discussed more thoroughly in Section 11.4. In the present example, the level of significance, α, equals 05.



The level of significance (α) indicates the degree of rarity required of an observed outcome in order to reject the null hypothesis (H0). For instance, the .05 level of significance indicates that H0 should be rejected if the observed z could have occurred just by chance with a probability of only .05 (one chance out of twenty) or less.

***CALCULATIONS**

We can use information from the sample to calculate a value for z. As has been noted previously, use Formula 10.1 to convert the observed sample mean of 533 into a z of 3.

***DECISIONS**

Either retain or reject H0 , depending on the location of the observed z value relative to the critical z values specified in the decision rule. According to the present rule, H0 should be rejected at the .05 level of significance because the observed z of 3 exceeds the critical z of 1.96 and, therefore, qualifies as a rare outcome, that is, an unlikely outcome from a population centered about the null hypothesis.

Retain or Reject H0?

If you are ever confused about whether to retain or reject H0 recall the logic behind the hypothesis test. You want to reject H0 only if the observed value of z qualifies as a rare outcome because it deviates too far into the tails of the sampling distribution. Therefore, you want to reject H0 only if the observed value of z equals or is more positive than the upper critical z (1.96) or if it equals or is more negative than the lower critical z (–1.96). Before deciding, you might find it helpful to sketch the hypothesized sampling distribution, along with its critical z values and shaded rejection regions, and then use some mark, such as an arrow (^), to designate the location of the observed value of z (3) along the z scale. If this mark is located in the shaded rejection region—or farther out than this region, as in Figure 10.4—then H0 should be rejected

***INTERPRETATION**

Finally, interpret the decision in terms of the original research problem. In the present example, it can be concluded that, since the null hypothesis was rejected, the mean SAT math score for the local freshman class probably differs from the national average of 500.

Although not a strict consequence of the present test, a more specific conclusion is possible. Since the sample mean of 533 (or its equivalent z of 3) falls in the upper rejection region of the hypothesized sampling distribution, it can be concluded that the population mean SAT math score for all local freshmen probably exceeds the national average of 500. By the same token, if the observed sample mean or its equivalent z had fallen in the lower rejection region of the hypothesized sampling distribution, it could have been concluded that the population mean for all local freshmen probably is below the national average

If the observed sample mean or its equivalent z had fallen in the retention region of the hypothesized sampling distribution, it would have been concluded (somewhat weakly, as discussed in Section 11.2) that there is no evidence that the population mean for all local freshmen differs from the national average of 500.

XXXXX

UNIT – II

INFERENTIAL STATISTICS II

Why hypothesis tests? – Strong or weak decisions – one-tailed and two-tailed tests – case studies Influence of sample size – power and sample size 46 Estimation – point estimate – confidence interval – level of confidence – effect of sample size

*WHY HYPOTHESIS TESTS?

There is a crucial link between hypothesis tests and the need of investigators, whether pollsters or researchers, to generalize beyond existing data. If the 100 freshmen in the SAT example of the previous chapter had been not a sample but a census of the entire freshman class, there wouldn't have been any need to generalize beyond existing data, and it would have been inappropriate to conduct a hypothesis test. Now, the observed difference between the newly observed population mean of 533 and the national average of 500, by itself, would have been sufficient grounds for concluding that the mean SAT math score for all local freshmen exceeds the national average. Indeed, any observed difference in favor of the local freshmen, regardless of the size of the difference, would have supported this conclusion. If we must generalize beyond the 100 freshmen to a larger local population, as was actually the case, the observed difference between 533 and 500 cannot be interpreted at face value. The basic problem is that the sample mean for a second random sample of 100 freshmen probably would differ, just by chance, from the sample mean of 533 for the first sample. Accordingly, the variability among sample means must be considered when we attempt to decide whether the observed difference between 533 and 500 is real or merely transitory

Importance of the Standard Error

To evaluate the effect of chance, we use the concept of a sampling distribution, that is, the concept of the sample means for all possible random outcomes. A key element in this concept is the standard error of the mean, a measure of the average amount by which sample means differ, just by chance, from the population mean. Dividing the observed difference (533−500) by the standard error (11) to obtain a value of z (3) locates the original observed difference along a z scale of either common outcomes (reasonably attributable to chance) or rare outcomes (not reasonably attributable to chance). If, when expressed as z, the ratio of the observed difference to the standard error is small enough to be reasonably attributed to chance, we retain H0. Otherwise, if the ratio of the observed difference to the standard error is too large to be reasonably attributed to chance, as in the SAT example, we reject H0. Before generalizing beyond the existing data, we must always measure the effect of chance; that is, we must obtain a value for the standard error. To appreciate the vital role of the standard error in the SAT example, increase its value from 11 to 33 and note that even though the observed difference

remains the same (533−500), we would retain, not reject, H0 because now z would equal 1 (rather than 3) and be less than the critical z of 1.96.

Possibility of Incorrect Decisions

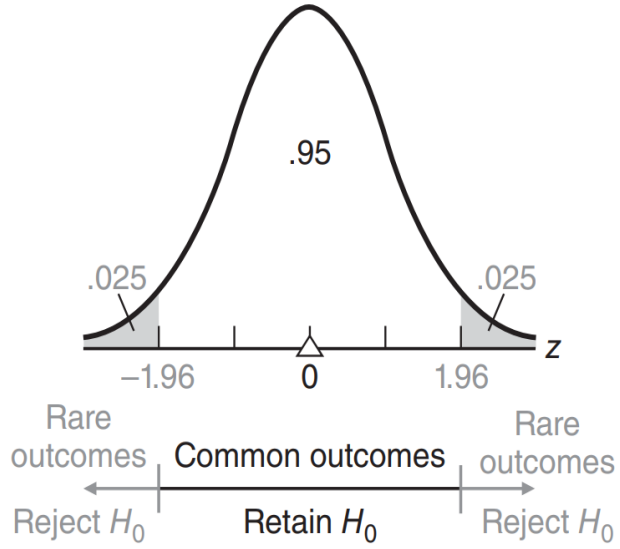
Having made a decision about the null hypothesis, we never know absolutely whether that decision is correct or incorrect, unless, of course, we survey the entire population. Even if H0 is true (and, therefore, the hypothesized distribution of z about H0 also is true), there is a slight possibility that, just by chance, the one observed z actually originates from one of the shaded rejection regions of the hypothesized distribution of z, thus causing the true H0 to be rejected. This type of incorrect decision—rejecting a true H0—is referred to as a type I error or a false alarm. On first impulse, it might seem desirable to abolish the shaded rejection regions in the hypothesized sampling distribution to ensure that a true H0 never is rejected. A most unfortunate consequence of this strategy, however, is that no H0, not even a radically false H0, ever would be rejected. This second type of incorrect decision—retaining a false H0—is referred to as a type II error or a miss. Both type I and type II errors are described in more detail later in this chapter

Minimizing Incorrect Decisions

Traditional hypothesis-testing procedures, such as the one illustrated in Figure 11.1, tend to minimize both types of incorrect decisions. If H0 is true, there is a high probability that the observed z will qualify as a common outcome under the hypothesized sampling distribution and that the true H0 will be retained. (In Figure 11.1, this probability equals the proportion of white area (.95) in the hypothesized sampling distribution.) On the other hand, if H0 is seriously false, because the hypothesized population mean differs considerably from the true population mean, there is also a high probability that the observed z will qualify as a rare outcome under the hypothesized distribution and that the false H0 will be rejected. (In Figure 11.1, this probability can't be determined since, in this case, the hypothesized sampling distribution does not actually reflect the true sampling distribution. More about this later in the chapter.)

Even though we never really know whether a particular decision is correct or incorrect, it is reassuring that in the long run, most decisions will be correct— assuming the null hypotheses are either true or seriously false

***STRONG OR WEAK DECISIONS;**



Proportions of area associated with common and rare outcomes (α = .05).

Retaining H0 Is a Weak Decision:

There are subtle but important differences in the interpretation of decisions to retain H0 and to reject H0. H0 is retained whenever the observed z qualifies as a common outcome on the assumption that H0 is true. Therefore, H0 could be true. However, the same observed result also would qualify as a common outcome when the original value in H0 (500) is replaced with a slightly different value. Thus, the retention of H0 must be viewed as a relatively weak decision. Because of this weakness, many statisticians prefer to describe this decision as simply a failure to reject H0 rather than as the retention of H0. In any event, the retention of H0 can't be interpreted as proving H0 to be true. If H0 had been retained in the present example, it would have been appropriate to conclude not that the mean SAT math score for all local freshmen equals the national average, but that the mean SAT math score could equal the national average, as well as many other possible values in the general vicinity of the national average

Rejecting H0 Is a Strong Decision:

On the other hand, H0 is rejected whenever the observed z qualifies as a rare outcome—one that could have occurred just by chance with a probability of .05 or less— on the assumption that H0 is true. This suspiciously rare outcome implies that H0 is probably false (and conversely, that H1 is

probably true). Therefore, the rejection of H0 can be viewed as a strong decision. When H0 was rejected in the present example, it was appropriate to report a definitive conclusion that the mean SAT math score for all local freshmen probably exceeds the national average. To summarize,

the decision to retain H0 implies not that H0 is probably true, but only that H0 could be true, whereas the decision to reject H0 implies that H0 is probably false (and that H1 is probably true).

Since most investigators hope to reject H0 in favor of H1, the relative weakness of the decision to retain H0 usually does not pose a serious problem

Why the Research Hypothesis Isn't Tested Directly

Even though H0, the null hypothesis, is the focus of a statistical test, it is usually of secondary concern to the investigator. Nevertheless, there are several reasons why, although of primary concern, the research hypothesis is identified with H1 and tested indirectly.

Lacks Necessary Precision

The research hypothesis, but not the null hypothesis, lacks the necessary precision to be tested directly.

To be tested, a hypothesis must specify a single number about which the hypothesized sampling distribution can be constructed. Because it specifies a single number, the null hypothesis, rather than the research hypothesis, is tested directly. In the SAT example, the null hypothesis specifies that a precise value (the national average of 500) describes the mean for the current population of interest (all local freshmen). Typically, the research hypothesis lacks the required precision. It merely specifies that some inequality exists between the hypothesized value (500) and the mean for the current population of interest (all local freshmen).

Supported by a Strong Decision to Reject

Logical considerations also argue for the indirect testing of the research hypothesis and the direct testing of the null hypothesis.

Because the research hypothesis is identified with the alternative hypothesis, the decision to reject the null hypothesis, should it be made, will provide strong support for the research hypothesis, while the decision to retain the null hypothesis, should it be made, will provide, at most, weak support for the null hypothesis

. As mentioned, the decision to reject the null hypothesis is stronger than the decision to retain it. Logically, a statement such as "All cows have four legs" can never be proven in spite of a steady stream of positive instances. It only takes one negative instance—one cow with three legs—to

disprove the statement. By the same token, one positive instance (common outcome) doesn't prove the null hypothesis, but one

***ONE-TAILED AND TWO-TAILED TESTS**

negative instance (rare outcome) disproves the null hypothesis. (Strictly speaking, however, since a rare outcome implies that the null hypothesis is probably but not definitely false, remember that there always is a very small possibility that the rare outcome reflects a true null hypothesis.) Logically, therefore, it makes sense to identify the research hypothesis with the alternative hypothesis. If, as hoped, the data favor the research hypothesis, the test will generate strong support for your hunch: It's probably true. If the data do not favor the research hypothesis, the hypothesis test will generate, at most, weak support for the null hypothesis: It could be true. Weak support for the null hypothesis is of little consequence, as this hypothesis—that nothing special is happening in the population—usually serves only as a convenient testing device.

Let's consider some techniques that make the hypothesis test more responsive to special condition

Two-Tailed Test:

Generally, the alternative hypothesis, H1, is the complement of the null hypothesis, H0. Under typical conditions, the form of H1 resembles that shown for the SAT example, namely,

$$H1: \mu \neq 500$$

This alternative hypothesis says that the null hypothesis should be rejected if the mean reading score for the population of local freshmen differs in either direction from the national average of 500. An observed z will qualify as a rare outcome if it deviates too far either below or above the national average. Panel A of Figure 11.2 shows rejection regions that are associated with both tails of the hypothesized sampling distribution. The corresponding decision rule, with its pair of critical z scores of ±1.96, is referred to as a two-tailed or nondirectional test.

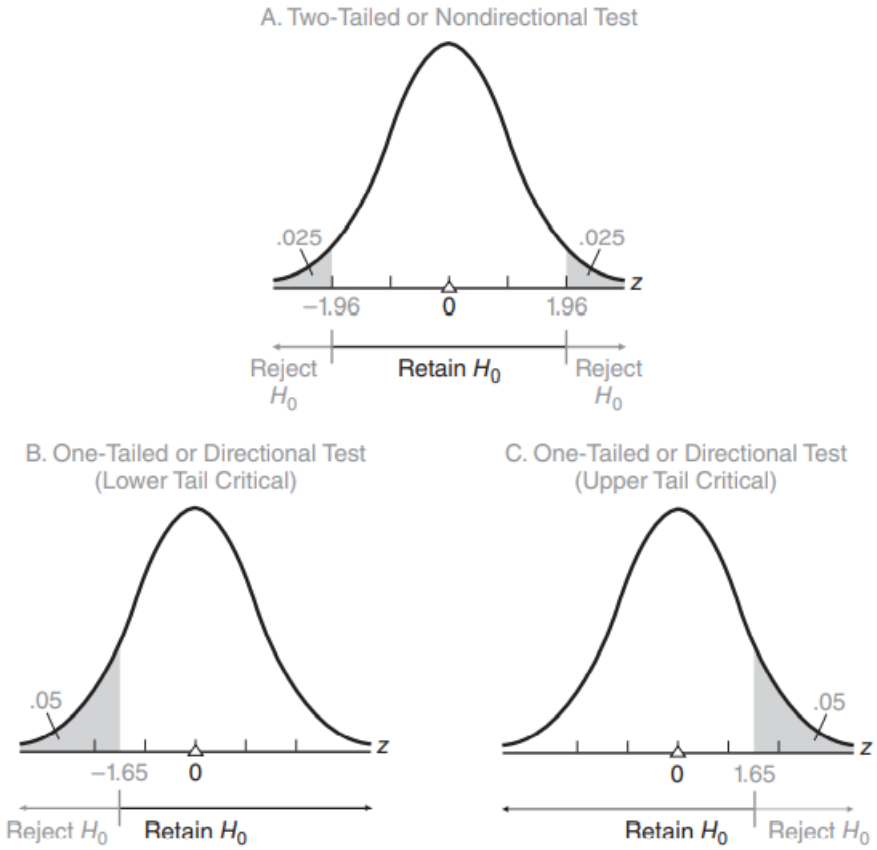
One-Tailed Test (Lower Tail Critical)

Now let's assume that the research hypothesis for the investigation of SAT math scores was based on complaints from instructors about the poor preparation of local freshmen. Assume also that if the investigation supports these complaints, a remedial program will be instituted. Under these circumstances, the investigator might prefer a hypothesis test that is specially designed to detect only whether the population mean math score for all local freshmen is less than the national average. This alternative hypothesis reads:

$$H1: \mu \leq 500$$

It reflects a concern that the null hypothesis should be rejected only if the population mean math score for all local freshmen is less than the national average of 500. Accordingly, an observed z triggers the decision to reject H0 only if z deviates too far below the national average. Panel B of Figure 11.2 illustrates a rejection region that is associated with only the lower tail of the hypothesized sampling distribution. The corresponding decision rule, with its critical z of –1.65, is referred to as a one-tailed or directional test with the lower tail critical. Use Table A in Appendix C to verify that if the critical z equals –1.65; then .05 of the total area under the distribution of z has been allocated to the lower rejection region. Notice that the level of significance, α, equals .05 for this one-tailed test and also for the original two-tailed test.

MORE ABOUT HYPOTHESIS TESTING



Extra Sensitivity of One-Tailed Test

This new one-tailed test is extra sensitive to any drop in the population mean for the local freshmen below the national average. If H0 is false because a drop has occurred, then the observed z will be more likely to deviate below the national average. As can be seen in panels A and B of Figure 11.2, an observed deviation in the direction of concern—below the national average—is more likely to penetrate the broader rejection region for the one-tailed test than that for the two-tailed test. Therefore, the decision to reject a false H0 (in favor of the research hypothesis) is more likely to occur in the one-tailed test than in the two-tailed test.

One-Tailed Test (Upper Tail Critical)

Panel C of Figure 11.2 illustrates a one-tailed or directional test with the upper tail critical. This one-tailed test is the mirror image of the previous test. Now the alternative hypothesis reads:

$$H1: \mu > 500$$

and its critical z equals 1.65. This test is specially designed to detect only whether the population mean math score for all local freshmen exceeds the national average. For example, the research hypothesis for this investigation might have been inspired by the possibility of eliminating an existing remedial math program if it can be demonstrated that, on the average, the SAT math scores of all local freshmen exceed the national average

***ONE-TAILED AND TWO-TAILED TESTS:**

Before a hypothesis test, if there is a concern that the true population mean differs from the hypothesized population mean only in a particular direction, use the appropriate one-tailed or directional test for extra sensitivity. Otherwise, use the more customary two-tailed or nondirectional test

Having committed yourself to a one-tailed test with its single rejection region, you must retain H0, regardless of how far the observed z deviates from the hypothesized population mean in the direction of "no concern." For instance, if a one-tailed test with the lower tail critical had been used with the data for 100 freshmen from the SAT example, H0 would have been retained because, even though the observed z equals an impressive value of 3, it deviates in the direction of no concern—in this case, above the national average. Clearly, a one-tailed test should be adopted only when there is absolutely

no concern about deviations, even very large deviations, in one direction. If there is the slightest concern about these deviations, use a two-tailed test.

The selection of a one- or two-tailed test should be made before the data are collected. Never "peek" at the value of the observed z to determine whether to locate the rejection region for a one-tailed test in the upper or the lower tail of the distribution of z. To qualify as a one-tailed test, the location of the rejection region must reflect the investigator's concern only about deviations in a particular direction before any inspection of the data. Indeed, the investigator should be able to muster a compelling reason, based on an understanding of the research hypothesis, to support the direction of the one-tailed test.

New Null Hypothesis for One-Tailed Tests:

When tests are one-tailed, a complete statement of the null hypothesis also should include all possible values of the population mean in the direction of no concern. For example, given a one-tailed test with the lower tail critical, such as H1: $\mu < 500$, the complete null hypothesis should be stated as H0: $\mu \geq 500$ instead of H0: $\mu = 500$. By the same token, given a one-tailed test with the upper tail critical, such as H1: $\mu > 500$, the complete null hypothesis should be stated as H0: $\mu \leq 500$.

If you think about it, the complete H0 describes all of the population means that could be true if a one-tailed test results in the retention of the null hypothesis. For instance, if a one-tailed test with the lower tail critical results in the retention of H0: $\mu \geq 500$, the complete H0 accurately reflects the fact that not only $\mu = 500$ could be true, but also that any other value of the population mean in the direction of no concern, that is, $\mu > 500$, could be true. (Remember, when the test is one-tailed, even a very deviant result in the direction of no concern—possibly reflecting a mean much larger than 500—still would trigger the decision to retain H0.) Henceforth, whenever a one-tailed test is employed, write H0 to include values of the population mean in the direction of no concern—even though the single number in the complete H0 identified by the equality sign is the one value about which the hypothesized sampling distribution is centered and, therefore, the one value actually used in the hypothesis test.

***INFLUENCE OF SAMPLE SIZE:**

Ordinarily, the investigator might not be too concerned about the low detection rate of .33 for the relatively small three-point effect of vitamin C on IQ. Under special circumstances, however, this low detection rate might be unacceptable. For example, previous experimentation might have established that vitamin C has many positive effects, including the reduction in the duration and severity of common colds, and no apparent negative side effects * Furthermore, huge quantities of vitamin C might be available at no cost to the school district. The establishment of one more positive effect,

even a fairly mild one such as a small increase in the population mean IQ, might clinch the case for supplying vitamin C to all students in the district. The investigator, therefore, might wish to use a test procedure for which, if H0 really is false because of a small effect, the detection rate is appreciably higher than .33.

To increase the probability of detecting a false H0, increase the sample size.

Assuming that vitamin C still has only a small three-point effect on IQ, we can check the properties of the projected one-tailed test when the sample size is increased from 36 to 100 students. Recall the formula for the standard error of the mean, x, namely,

For the original experiment with its sample size of 36,

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

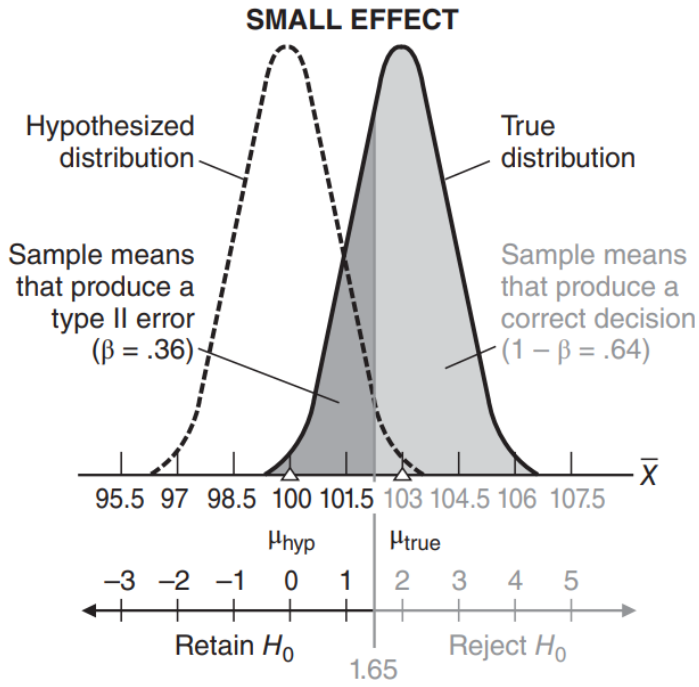
whereas for the new experiment with its sample size of 100,

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{36}} = \frac{15}{6} = 2.5$$

Clearly, any increase in sample size causes a reduction in the standard error of the mean.

$$\sigma_{\bar{x}} = \frac{15}{\sqrt{100}} = \frac{15}{10} = 1.5$$

MORE ABOUT HYPOTHESIS TESTING:



Consequences of Reducing Standard Error

As can be seen by comparing Figure 11.5 and Figure 11.6, the reduction of the standard error from 2.5 to 1.5 has two important consequences:

1. It shrinks the upper retention region back toward the hypothesized population mean of 100.
2. It shrinks the entire true sampling distribution toward the true population mean of 103.

The net result is that, among randomly selected sample means for 100 students, fewer sample means (.36) produce a type II error because they originate from the black sector, and more sample means (.64) produce a correct decision—that is, more lead to the detection of a false H0—because they originate from the shaded sector.

An obvious implication is that the standard error can be reduced to any desired value merely by increasing the sample size. To cite an extreme case, when the sample size equals 10,000 students (!), the standard error drops to 0.15. In this case, the upper retention region shrinks to the immediate vicinity of the hypothesized population mean of 100, and the entire true sampling distribution of the mean shrinks to the immediate vicinity of the true population mean of 103. The net result is that a type II error hardly ever is committed, and the small three-point effect virtually always is detected

***POWER AND SAMPLE SIZE**

The power of a hypothesis test equals the probability (1 − ß) of detecting a particular effect when the null hypothesis (H0) is false. Power is simply the complement (1 − ß) of the probability (ß) of failing to detect the effect, that is, the complement of the probability of a type II error. The shaded sectors in Figures 11.4, 11.5, and 11.6 illustrate varying degrees of power.

In Figures 11.5 and 11.6, sample sizes of 36 and 100 were selected, with computational convenience in mind, to dramatize different degrees of power for a small three-point effect of vitamin C on IQ. Preferably, the selection of sample size should reflect—as much as circumstances permit—your considered judgment about what constitutes (1) the smallest important effect and (2) a reasonable degree of power for detecting that effect. For example, the following considerations might influence the selection of a new sample size for the vitamin C study.

1. The smallest effect that merits detection, we might conclude, equals seven points. This might reflect our judgment, possibly supported by educational consultants, that only a mean IQ of at least 107 for all students in the school district justifies the effort and expense of upgrading the entire curriculum. Another possible reason for focusing on a seven-point effect—in the absence of any compelling reason to the contrary—might be that, since 7 is about one-half the size of the standard deviation of 15, it avoids extreme effect sizes by qualifying as a "medium" effect size, according to Jacob Cohen's widely adopted guidelines described in Section 14.9.
2. . A reasonable degree of power for this seven-point effect, we might conclude, equals .80. This degree of power will detect the specified effect with a tolerable rate of eighty times out of one hundred. In the absence of special concerns about the type II error, many investigators would choose .80 as a default value for power—along with .05 as the default value for the level of significance—to avoid the large sample sizes required by high degrees of power, such as .95 or .99.

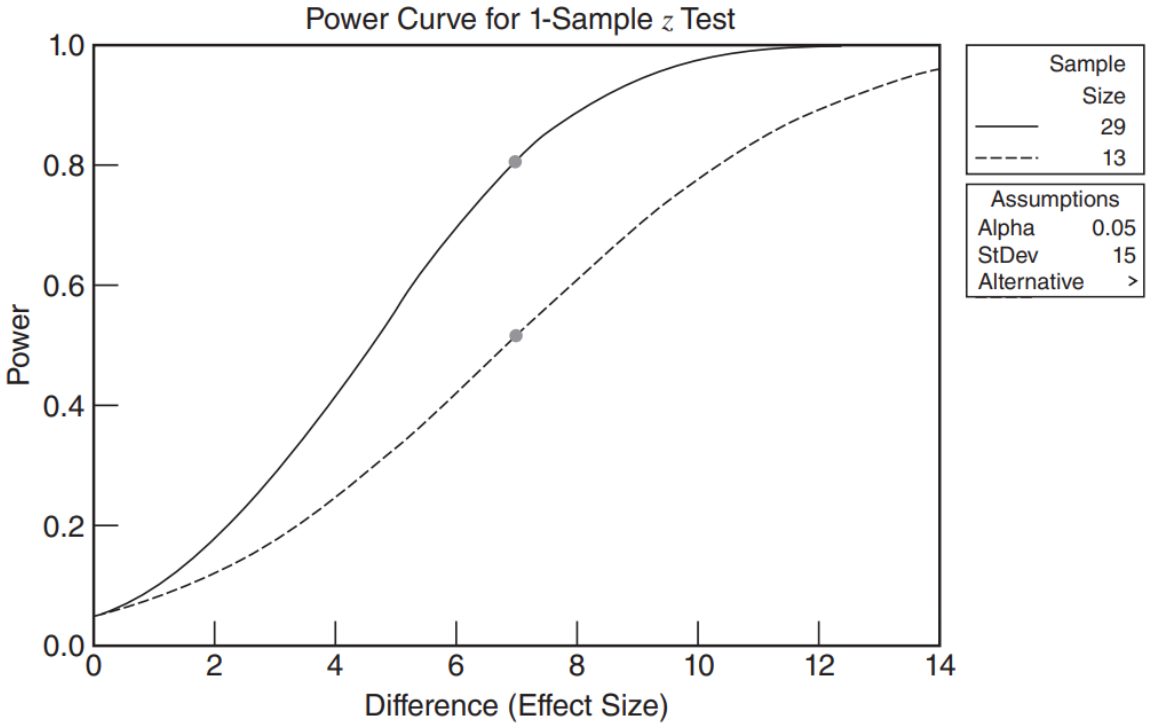
Power Curves

Basically, a power curve shows how the likelihood of detecting any possible effect—ranging from very small to very large—varies for a fixed sample size * With just a few key strokes, Minitab's Power and Sample Size software calculates that a sample size of 29 will satisfy the original specifications to detect a seven-point effect with power .80. The upper (solid line) power curve in Figure 11.7 is based on a sample size of 29, and it features a dot whose coordinates are a seven-point effect (difference) and a power of .80.

The S-shaped power curve for a sample of 29 also shows the growth in power with increases in effect size. Verify that power equals only about .40 for a smaller four-point effect and about .95 for a larger

ten-point effect. A four-point effect will be detected only about forty times in one hundred, while a ten-point effect will be detected about ninety-five times in one hundred

Practical considerations, such as limitations in money or facilities, might force a reduction (always painful) in the prescribed sample size of 29. Although the original specifications represent our best judgment about an appropriate sample size, there usually is latitude for compromise. For example, referring to Figure 11.7, we could consider the properties of the lower (broken line) power curve for a smaller sample size of 13. (Ordinarily, to minimize the loss of power, we probably would have considered power



The sample size of 29 also could be reduced indirectly by compromising other properties of the original specifications. We could reduce the prescribed sample size by enlarging the smallest important effect (preferably by small increments above the seven-point effect); by lowering the degree of power (preferably not much below .80); by increasing the level of significance (preferably not

above .10); by selecting, if appropriate, a one-tailed rather than a two-tailed test (if this had not been done already in the vitamin C investigation); or by taking some combination of these actions. For instance, we could enlarge the smallest important effect from seven to eight points. Although not shown in Figure 11.7, Minitab calculates that a smaller sample of 22 detects the larger eight-point effect with power equal to .80

Since a power analysis depends on a number factors, including the investigator's subjective judgment about what constitutes a reasonable detection rate for the smallest important effect—as well as the availability of local resources and any subsequent compromises—two equally competent investigators might select different sample sizes for the same study. Nonetheless, in the hands of a judicious investigator

the use of power curves represents a distinct improvement over the arbitrary selection of sample size, for power curves help identify a sample size that, being neither unduly small nor excessively large, produces a hypothesis test with the proper sensitivity

Power Analysis of Studies by Others

If you suspect that another investigator's reported failure to reject H0 might have been caused by an unduly small sample size, power curves can be consulted retroactively to evaluate the adequacy of the publicized results. For example, if the sample size reported for a vitamin C study had been only 13, you could have consulted the lower curve in Figure 11.7 to establish that your smallest important effect of seven points would have been detected with a very low power of approximately .50. You could have endorsed, therefore, the need for a replication or duplication of the original study with a more powerful, larger sample size.

Need Not Predict True Effect Size

The use of power curves does not require that you predict the true effect size—an impossible task—but merely that you specify the smallest effect that, if present, merits detection. If the true effect size actually is larger than the specified effect, the true power actually will exceed the specified power—since more of the true sampling distribution overlaps the rejection region for the false H0 than does the sampling distribution for the specified effect. (If this is not obvious, compare Figures 11.4 and 11.5.) Thus, a more important effect is even more likely to be detected. On the other hand, if the true effect size actually is smaller than the specified effect, the entire process works in reverse but still to your advantage, since an unimportant effect, which you would just as soon miss, is even less likely to be detected

Initiating a Power Analysis

It is beyond the scope of this book to provide detailed information about either manual or electronic calculations for a power analysis. Manual calculations are described in Chapter 8 of D. C. Howell, Statistical Methods for Psychology, 8th ed. (Belmont CA: Wadsworth, 2013). Electronic calculations are made by each of the three statistical packages—Minitab, SPSS, and SAS—featured in this book, as well a number of free websites, such as G*Power 3 at http://www.gpower.hhu.de/. Once you have decided what constitutes the smallest important effect that merits detection with a certain power, the step-by-step details of a power analysis, whether manual or electronic, usually are straightforward and amenable to any power analysis that you yourself might initiate.

For a one-tailed or directional test with the lower tail critical

H0: μ ≥ SOME NUMBERS

H1: μ < SOME NUMBERS

For a one-tailed or directional test with the upper tail critical,

H0: μ ≤ SOME NUMBERS

H1: μ > SOME NUMBERS

Unless there are obvious reasons for selecting either a larger or a smaller level of significance, use the customary .05 level.

There are four possible outcomes for any hypothesis test:

■ If H0 really is true, it is a correct decision to retain the true H0.

■ If H0 really is true, it is a type I error to reject the true H0.

■ If H0 really is false, it is a type II error to retain the false H0.

■ If H0 really is false, it is a correct decision to reject the false H0.

When generalizing beyond the existing data, there is always the possibility of a type I or type II error. At best, a hypothesis test tends to produce a correct decision when either H0 really is true or H0 really is false because of a large effect.

***Estimation**

***POINT ESTIMATE**

A point estimate for μ uses a single value to represent the unknown population mean.

This is the most straightforward type of estimate. If a random sample of 100 local freshmen reveals a sample mean SAT score of 533, then 533 will be the point estimate of the unknown population mean for all local freshmen. The best single point estimate for the unknown population mean is simply the observed value of the sample mean.

A Basic Deficiency

Although straightforward, simple, and precise, point estimates suffer from a basic deficiency. They tend to be inaccurate. Because of sampling variability, it's unlikely that a single sample mean, such as 533, will coincide with the population mean. Since point estimates convey no information about the degree of inaccuracy due to sampling variability, statisticians supplement point estimates with another, more realistic type of estimate, known as interval estimates or confidence intervals.

***CONIDENCE INTERVAL (CI) FOR μ**

A confidence interval for μ uses a range of values that, with a known degree of certainty, includes the unknown population mean.

For instance, the SAT investigator might use a confidence interval to claim, with 95 percent confidence, that the interval between 511.44 and 554.56 includes the population mean math score for all local freshmen. To be 95 percent confident signifies that if many of these intervals were constructed for a long series of samples, approximately 95 percent would include the population mean for all local freshmen. In the long run, 95 percent of these confidence intervals are true because they include the unknown population mean. The remaining 5 percent are false because they fail to include the unknown population mean.

Why Confidence Intervals WorK

To understand confidence intervals, you must view them in the context of three important properties of the sampling distribution of the mean described in Chapter 10.

■ The mean of the sampling distribution equals the unknown population mean for all local freshmen, whatever its value, because the mean of this sampling distribution always equals the population mean.

■ The standard error of the sampling distribution equals the value (11) obtained from dividing the population standard deviation (110) by the square root of the sample size (100).

■ The shape of the sampling distribution approximates a normal distribution because the sample size of 100 satisfies the requirements of the central limit theorem.

A Series of Confidence Intervals

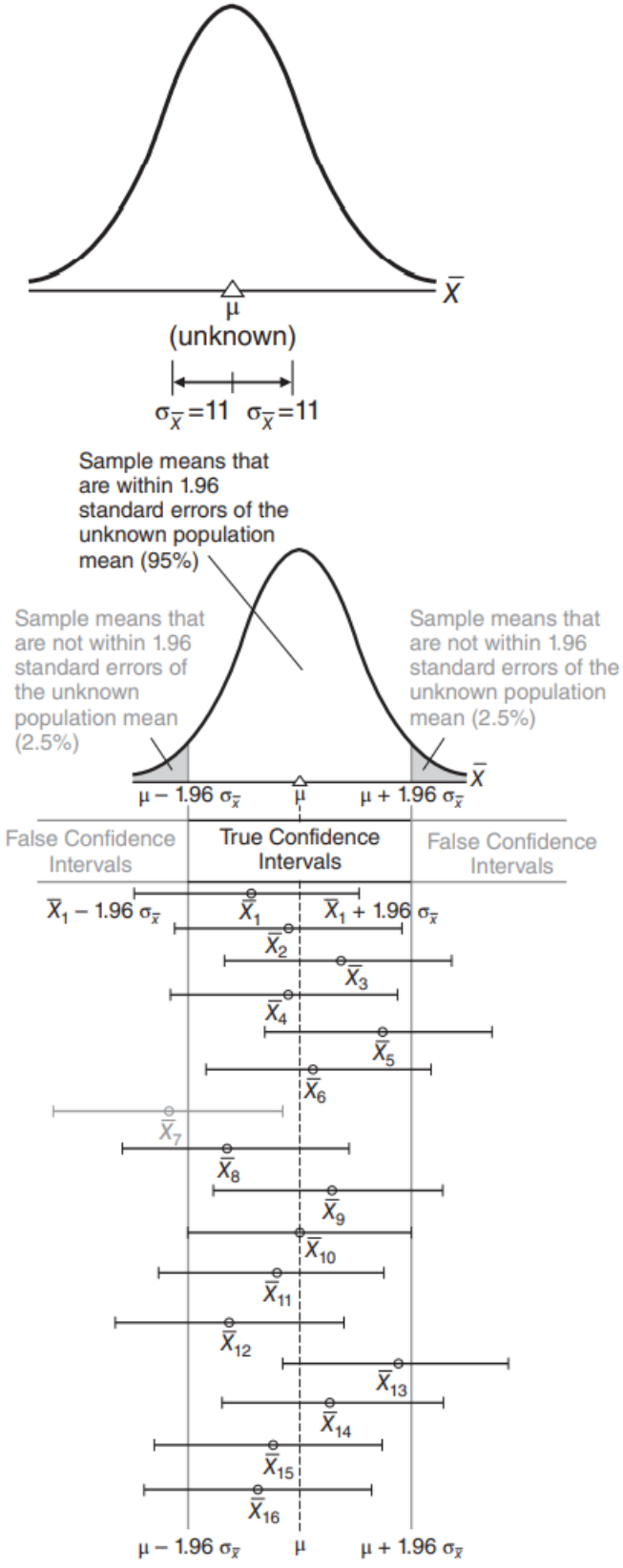
In practice, only one sample mean is actually taken from this sampling distribution and used to construct a single 95 percent confidence interval. However, imagine taking not just one but a series of randomly selected sample means from this sampling distribution. Because of sampling variability, these sample means tend to differ among themselves. For each sample mean, construct a 95 percent confidence interval by adding 1.96 standard errors to the sample mean and subtracting 1.96 standard errors from the sample mean; that is, use the expression

$$\bar{X} \pm 1.96\sigma_{\bar{X}},$$

to obtain a 95 percent confidence interval for each sample mean

True Confidence Interval

Why, according to statistical theory, do 95 percent of these confidence intervals include the unknown population mean? As indicated in Figure 12.2, because the sampling distribution is normal, 95 percent of all sample means are within 1.96 standard errors of the unknown population mean, that is, 95 percent of all sample means deviate less than 1.96 standard errors from the unknown population mean. Therefore, and this is the key point, when sample means are expanded into confidence intervals—by adding and subtracting 1.96 standard errors—95 percent of all possible confidence intervals are true because they include the unknown population mean. To illustrate



False Confidence Intervals

Five percent of all confidence intervals fail to include the unknown population mean. As indicated in Figure 12.2, 5 percent of all sample means (2.5 percent in each tail) deviate more than 1.96 standard errors from the unknown population mean. Therefore, when sample means are expanded into confidence intervals—by adding and subtracting 1.96 standard errors—5 percent of all possible confidence intervals

*** LEVEL OF CONFIDENCE**

The level of confidence indicates the percent of time that a series of confidence intervals includes the unknown population characteristic, such as the population mean. Any level of confidence may be assigned to a confidence interval merely by substituting an appropriate value for z.conf in Formula 12.1. For instance, to construct a 99 percent confidence interval from the data for SAT math scores, first consult Table A in Appendix C to verify that z.conf values of ±2.58 define the middle 99 percent of the total area under the normal curve. Then substitute numbers for symbols in Formula 12.1 to obtain

$$533 \pm (2.58)(11) = 533 \pm 28.38 = \begin{cases} 561.38 \\ 504.62 \end{cases}$$

It can be claimed, with 99 percent confidence, that the interval between 504.62 and 561.38 includes the value of the unknown mean math score for all local freshmen. This simplies that, in the long run, 99 percent of these confidence intervals will include the unknown population mean.

Effect on Width of Interval

Notice that the 99 percent confidence interval of 504.62 to 561.38 is wider and, therefore, less precise than the corresponding 95 percent confidence interval of 511.44 to 554.56. The shift from a 95 percent to a 99 percent level of confidence requires an increase in the value of z.conf from 1.96 to 2.58. This increase, in turn, causes a wider, less precise confidence interval. Any shift to a higher level of confidence always produces a wider, less precise confidence interval unless offset by an increase in sample size, as mentioned in the next section.

Choosing a Level of Confidence

Although many different levels of confidence have been used, 95 percent and 99 percent are the most prevalent. Generally, a larger level of confidence, such as 99 percent, should be reserved for situations

in which a false interval might have particularly serious consequences, such as the failure of a national opinion pollster to predict the winner of a presidential election.

*** EFFECT OF SAMPLE SIZE**

The larger the sample size, the smaller the standard error and, hence, the more precise (narrower) the confidence interval will be. Indeed, as the sample size grows larger, the standard error will approach zero and the confidence interval will shrink to a point estimate. Given this perspective, the sample size for a confidence interval, unlike that for a hypothesis test, never can be too large.

Selection of Sample Size

As with hypothesis tests, sample size can be selected according to specifications established before the investigation. To generate a confidence interval that possesses the desired precision (width), yet complies with the desired level of confidence, refer to formulas for sample size in other statistics books.* Valid use of these formulas requires that before the investigation, the population standard deviation be either known or estimated.

XXXXX

UNIT III

T-TEST

t-test for one sample – sampling distribution of t – t-test procedure – degrees of freedom – estimating the standard error – case studies t-test for two independent samples – statistical hypotheses – sampling distribution – test procedure – p-value – statistical significance – estimating effect size – meta analysis t-test for two related samples

***T-TEST FOR ONE SAMPLE**

What is the one-sample *t*-test?

The one-sample t-test is a statistical hypothesis test used to determine whether an unknown population mean is different from a specific value.

When can I use the test?

You can use the test for continuous data. Your data should be a random sample from a normal population.

What if my data isn't nearly normally distributed?

If your sample sizes are very small, you might not be able to test for normality. You might need to rely on your understanding of the data. When you cannot safely assume normality, you can perform a *nonparametric* test that doesn't assume normality.

Using the one-sample *t*-test

The sections below discuss what we need for the test, checking our data, performing the test, understanding test results and statistical details.

What do we need?

For the one-sample *t*-test, we need one variable.

We also have an idea, or hypothesis, that the mean of the population has some value. Here are two examples:

A hospital has a random sample of cholesterol measurements for men. These patients were seen for issues other than cholesterol. They were not taking any

medications for high cholesterol. The hospital wants to know if the unknown mean cholesterol for patients is different from a goal level of 200 mg.

We measure the grams of protein for a sample of energy bars. The label claims that the bars have 20 grams of protein. We want to know if the labels are correct or not.

For a valid test, we need data values that are:

Independent (values are not related to one another).

Continuous.

Obtained via a simple random sample from the population.

Also, the population is assumed to be normally distributed.

One-sample *t*-test example

Imagine we have collected a random sample of 31 energy bars from a number of different stores to represent the population of energy bars available to the general consumer. The labels on the bars claim that each bar contains 20 grams of protein.

Table 1: Grams of protein in random sample of energy bars

Energy Bar - Grams of Protein						
20.70	27.46	22.15	19.85	21.29	24.75	
20.75	22.91	25.34	20.33	21.54	21.08	
22.14	19.56	21.10	18.04	24.12	19.95	
19.72	18.28	16.26	17.46	20.53	22.12	
25.06	22.44	19.08	19.88	21.39	22.33	25.79

If you look at the table above, you see that some bars have less than 20 grams of protein. Other bars have more. You might think that the data support the idea that the labels are correct. Others might disagree. The statistical test provides a sound method to make a decision, so that everyone makes the same decision on the same set of data values.

Checking the data

Let's start by answering: Is the *t*-test an appropriate method to test that the energy bars have 20 grams of protein ? The list below checks the requirements for the test.

The data values are independent. The grams of protein in one energy bar do not depend on the grams in any other energy bar. An example of dependent values would be if you collected energy bars from a single production lot. A sample from a single lot is representative of that lot, not energy bars in general.

The data values are grams of protein. The measurements are continuous.

We assume the energy bars are a simple random sample from the population of energy bars available to the general consumer (i.e., a mix of lots of bars).

We assume the population from which we are collecting our sample is normally distributed, and for large samples, we can check this assumption.

We decide that the *t*-test is an appropriate method.

Before jumping into analysis, we should take a quick look at the data. The figure below shows a histogram and summary statistics for the energy bars.

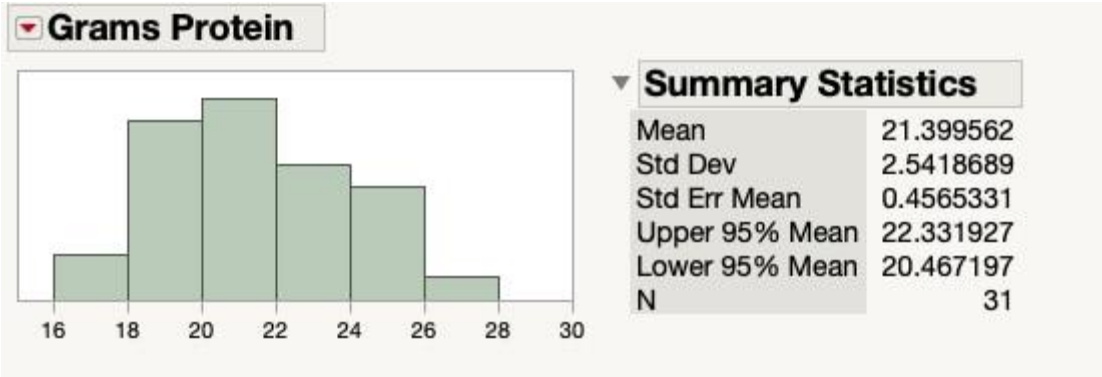


Figure 1: Histogram and summary statistics for the grams of protein in energy bars

From a quick look at the histogram, we see that there are no unusual points, or *outliers*. The data look roughly bell-shaped, so our assumption of a normal distribution seems reasonable.

From a quick look at the statistics, we see that the average is 21.40, above 20. Does this average from our sample of 31 bars invalidate the label's claim of 20 grams of protein for the unknown entire population mean? Or not?

How to perform the one-sample *t*-test

For the *t*-test calculations we need the average, standard deviation and sample size. These are shown in the summary statistics section of Figure 1 above.

We round the statistics to two decimal places. Software will show more decimal places, and use them in calculations. (Note that Table 1 shows only two decimal places; the actual data used to calculate the summary statistics has more.)

We start by finding the difference between the sample average and 20:

$$21.40-20 = 1.4021.40-20 = 1.40$$

Next, we calculate the standard error for the mean. The calculation is:

$$\text{Standard Error for the mean} = s\sqrt{n} = 2.54\sqrt{31} = 0.456sn = 2.5431 = 0.456$$

This matches the value in Figure 1 above.

We now have the pieces for our test statistic. We calculate our test statistic as:

$$t = \text{DifferenceStandard Error} = 1.400.456 = 3.07t = \text{DifferenceStandard Error} = 1.400.456 = 3.07$$

To make our decision, we compare the test statistic to a value from the *t*-distribution. This activity involves four steps.

We calculate a test statistic. Our test statistic is 3.07.

We decide on the risk we are willing to take for declaring a difference when there is not a difference. For the energy bar data, we decide that we are willing to take a 5% risk of saying that the unknown population mean is different from 20 when in fact it is not. In statistics-speak, we set $\alpha = 0.05$. In practice, setting your risk level (α) should be made before collecting the data.

We find the value from the *t*-distribution based on our decision. For a *t*-test, we need the *degrees of freedom* to find this value. The *degrees of freedom* are based on the sample size. For the energy bar data:

$$\text{degrees of freedom} = n-1 = 31-1 = 30n-1 = 31-1 = 30$$

The critical value of *t* with $\alpha = 0.05$ and 30 degrees of freedom is +/- 2.043. Most statistics books have look-up tables for the distribution. You can also find tables

online. The most likely situation is that you will use software and will not use printed tables.

We compare the value of our statistic (3.07) to the *t* value. Since 3.07 > 2.043, we reject the null hypothesis that the mean grams of protein is equal to 20. We make a practical conclusion that the labels are incorrect, and the population mean grams of protein is greater than 20.

Statistical details

Let's look at the energy bar data and the 1-sample *t*-test using statistical terms.

Our null hypothesis is that the underlying population mean is equal to 20. The null hypothesis is written as:

$$H_o:\mu=20 Ho:\mu=20$$

The alternative hypothesis is that the underlying population mean is not equal to 20. The labels claiming 20 grams of protein would be incorrect. This is written as:

$$H_a:\mu\neq20 Ha:\mu\neq20$$

This is a two-sided test. We are testing if the population mean is different from 20 grams in either direction. If we can reject the null hypothesis that the mean is equal to 20 grams, then we make a practical conclusion that the labels for the bars are incorrect. If we cannot reject the null hypothesis, then we make a practical conclusion that the labels for the bars may be correct.

We calculate the average for the sample and then calculate the difference with the population mean, mu:

$$\bar{x}-\mu x¯−\mu$$

We calculate the standard error as:

$$s\sqrt{n} sn$$

The formula shows the sample standard deviation as *s* and the sample size as *n*.

The test statistic uses the formula shown below:

$$\bar{x}-\mu s/\sqrt{n} x¯−\mu s/n$$

We compare the test statistic to a *t* value with our chosen alpha value and the degrees of freedom for our data. Using the energy bar data as an example, we set $\alpha = 0.05$. The degrees of freedom (*df*) are based on the sample size and are calculated as:

$$df=n-1=31-1=30df=n-1=31-1=30$$

Statisticians write the *t* value with $\alpha = 0.05$ and 30 degrees of freedom as:

$$t_{0.05,30}t0.05,30$$

The *t* value for a two-sided test with $\alpha = 0.05$ and 30 degrees of freedom is +/- 2.042. There are two possible results from our comparison:

The test statistic is less extreme than the critical *t* values; in other words, the test statistic is not less than -2.042, or is not greater than +2.042. You fail to reject the null hypothesis that the mean is equal to the specified value. In our example, you would be unable to conclude that the label for the protein bars should be changed.

The test statistic is more extreme than the critical *t* values; in other words, the test statistic is less than -2.042, or is greater than +2.042. You reject the null hypothesis that the mean is equal to the specified value. In our example, you conclude that either the label should be updated or the production process should be improved to produce, on average, bars with 20 grams of protein.

Testing for normality

The normality assumption is more important for small sample sizes than for larger sample sizes.

Normal distributions are symmetric, which means they are “even” on both sides of the center. Normal distributions do not have extreme values, or outliers. You can check these two features of a normal distribution with graphs. Earlier, we decided that the energy bar data was “close enough” to normal to go ahead with the assumption of normality. The figure below shows a normal quantile plot for the data, and supports our decision.

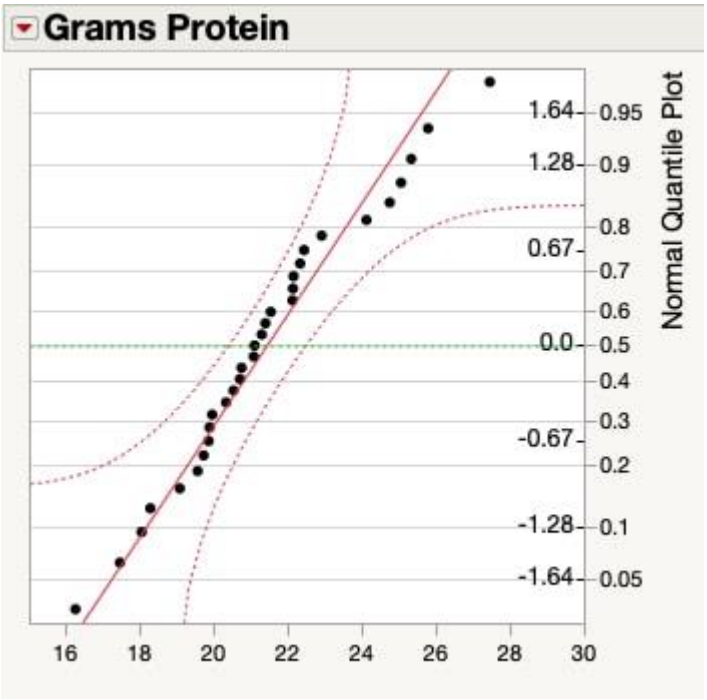


Figure 4: Normal quantile plot for energy bar data

You can also perform a formal test for normality using software. The figure below shows results of testing for normality with JMP software. We cannot reject the hypothesis of a normal distribution.

Goodness-of-Fit Test

	A2	Prob > A2
Anderson-Darling	0.3557569	0.4330

Figure 5: Testing for normality using JMP software

We can go ahead with the assumption that the energy bar data is normally distributed.

What if my data are not from a Normal distribution?

If your sample size is very small, it is hard to test for normality. In this situation, you might need to use your understanding of the measurements. For example, for the energy bar data, the company knows that the underlying distribution of grams of protein is normally distributed. Even for a very small sample, the company would likely go ahead with the *t*-test and assume normality.

What if you know the underlying measurements are not normally distributed? Or what if your sample size is large and the test for normality is rejected? In this situation, you can use a nonparametric test. Nonparametric analyses do not depend

on an assumption that the data values are from a specific distribution. For the one-sample *t*-test, the one possible nonparametric test is the Wilcoxon Signed Rank test.

Understanding p-values

Using a visual, you can check to see if your test statistic is more extreme than a specified value in the distribution. The figure below shows a *t*-distribution with 30 degrees of freedom.

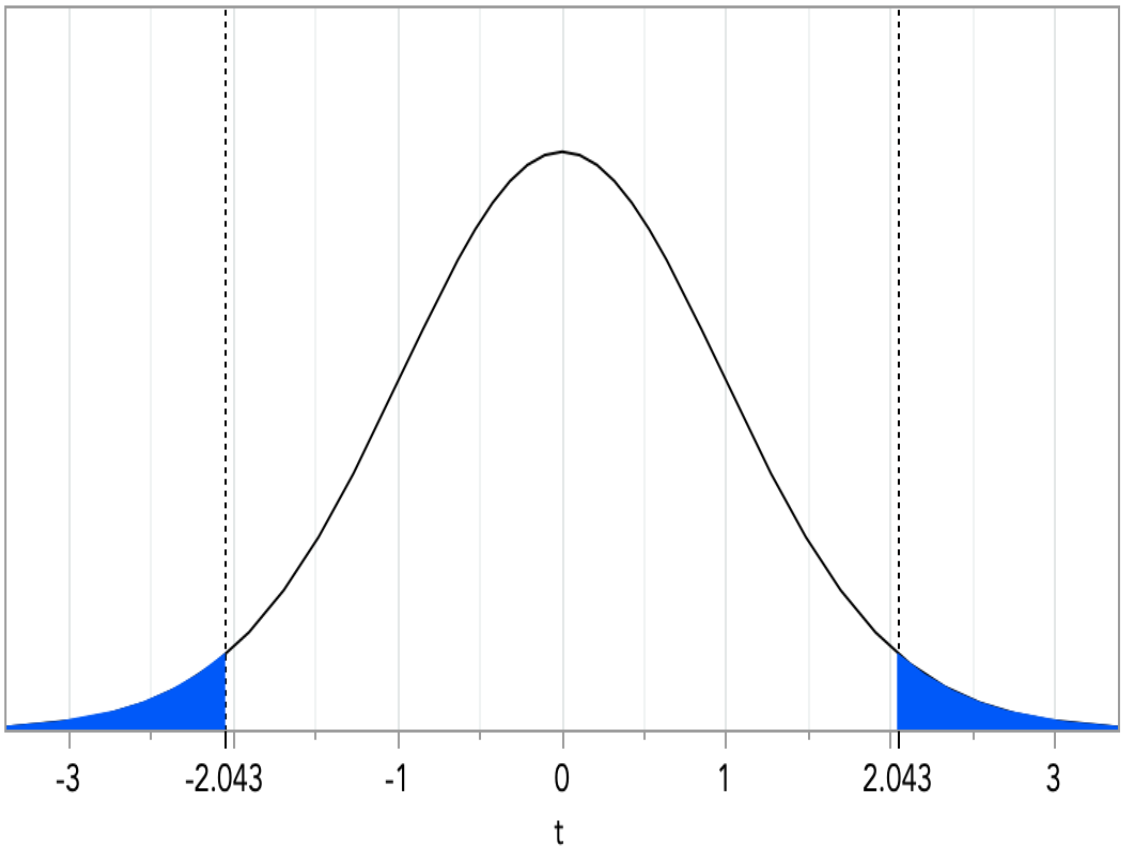


Figure 6: t-distribution with 30 degrees of freedom and α = 0.05

Since our test is two-sided and we set α = 0.05, the figure shows that the value of 2.042 "cuts off" 5% of the data in the tails combined.

The next figure shows our results. You can see the test statistic falls above the specified critical value. It is far enough "out in the tail" to reject the hypothesis that the mean is equal to 20.

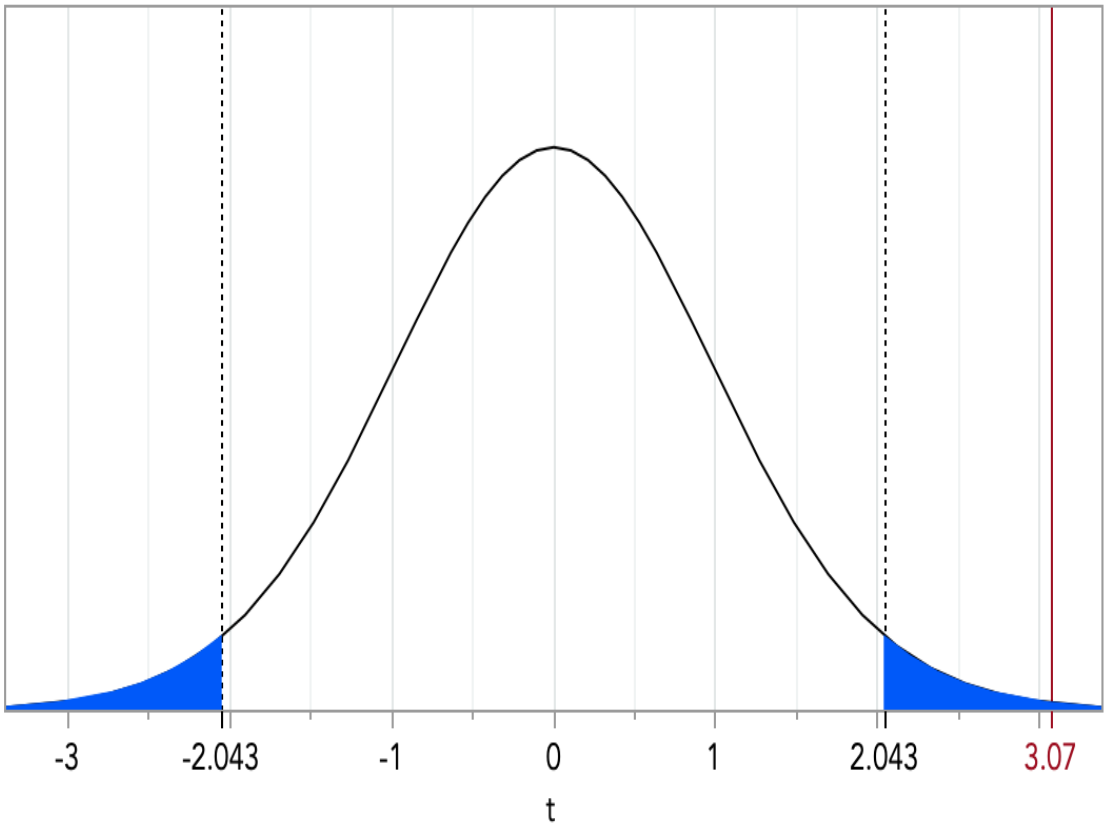


Figure 7: Our results displayed in a t-distribution with 30 degrees of freedom

Putting it all together with Software

You are likely to use software to perform a *t*-test. The figure below shows results for the 1-sample *t*-test for the energy bar data from JMP software.

Test Mean	
Hypothesized Value	20
Actual Estimate	21.3996
DF	30
Std Dev	2.54187

	t Test
Test Statistic	3.0656
Prob > |t|	0.0046*
Prob > t	0.0023*
Prob < t	0.9977

Figure 8: One-sample t-test results for energy bar data using JMP software

The software shows the null hypothesis value of 20 and the average and standard deviation from the data. The test statistic is 3.07. This matches the calculations above.

The software shows results for a two-sided test and for one-sided tests. We want the two-sided test. Our null hypothesis is that the mean grams of protein is equal to 20. Our alternative hypothesis is that the mean grams of protein is not equal to 20. The software shows a *p*-value of 0.0046 for the two-sided test. This *p*-value describes the likelihood of seeing a sample average as extreme as 21.4, or more extreme, when the underlying population mean is actually 20; in other words, the probability of observing a sample mean as different, or even more different from 20, than the mean we observed in our sample. A *p*-value of 0.0046 means there is about 46 chances out of 10,000. We feel confident in rejecting the null hypothesis that the population mean is equal to 20

*sampling distribution of t:

Like the sampling distribution of *z*, the sampling distribution of *t represents the distribution that would be obtained if a value of t were calculated for each sample mean for all possible random samples of a given size from some population.* In the early 1900s, William Gosset discovered the sampling distribution of *t* and subsequently reported his achievement under the pen name of "Student." Actually, Gosset discovered not just one but an entire family of *t* sampling distributions (or "Student's" distributions). Each *t* distribution is associated with a special number referred to as *degrees of freedom*, first discussed in Section 4.6. The concept of degrees of freedom is introduced because we're using variability in a sample to estimate the unknown variability in the population. Recall that when the *n deviations about the sample mean* are used to estimate variability in the population, only $n - 1$ are free to vary because of the restriction that the sum of these deviations must always equal zero. Since one degree of freedom is lost because of the zero-sum restriction, there are only $n - 1$ degrees of freedom, that is, symbolically,

DEGREES OF FREEDOM (ONE SAMPLE)

$$df = n - 1$$

where *df* represents degrees of freedom and *n* equals the sample size. Since the gas

mileage investigation involves six cars, the corresponding *t* test is based on a sampling distribution with five degrees of freedom (from $df = 6 - 1$).

HYPOTHESIS TEST SUMMARY: *t* TEST FOR A POPULATION MEAN (GAS MILEAGE INVESTIGATION)

Research Problem

Does the mean gas mileage for some population of cars drop below the legally required minimum of 45 mpg?

Statistical Hypotheses

$$H_0 : \mu \geq 45$$
$$H_1 : \mu < 45$$

Decision Rule

Reject H_0 at the .01 level of significance if $t \leq -3.365$ (from Table B, Appendix C, given $df = n - 1 = 6 - 1 = 5$).

Calculations

Given $\overline{X} = 43$, $s_{\overline{X}} = 0.89$
(See Table 13.1 on page 240 for computations.),

$$t = \frac{43 - 45}{0.89} = -2.25$$

Decision

Retain H_0 at the .01 level of significance because $t = -2.25$ is less negative than -3.365.

Interpretation

The population mean gas mileage *could* equal the required 45 mpg or more. The manufacturer shouldn't be penalized.

Compared to the Standard Normal Distribution

Figure 13.1 shows three *t* distributions. When there is an infinite (∞) number of degrees of freedom (and, therefore, the sample standard deviation becomes the same as the population standard deviation), the distribution of *t* is the same as the standard normal distribution of *z*. Notice that even with only four or ten degrees of freedom, a *t* distribution shares a number of properties with the normal distribution. All *t* distributions are symmetrical, unimodal, and bell-shaped, with a dense concentration that peaks in the middle (when *t* equals 0) and tapers off both to the right and left of the middle (as *t*

becomes more positive or negative, respectively). *The inflated tails of the t distribution, particularly apparent with small values of df, constitute the most important difference between t and z distributions.*

Table for *t* Distributions

To save space, tables for *t* distributions concentrate only on the critical values of *t* that correspond to the more common levels of significance. Table B of Appendix C lists the critical *t* values for either one- or two-tailed hypothesis tests at the .05, .01, and .001 levels of significance. All listed critical *t* values are positive and originate

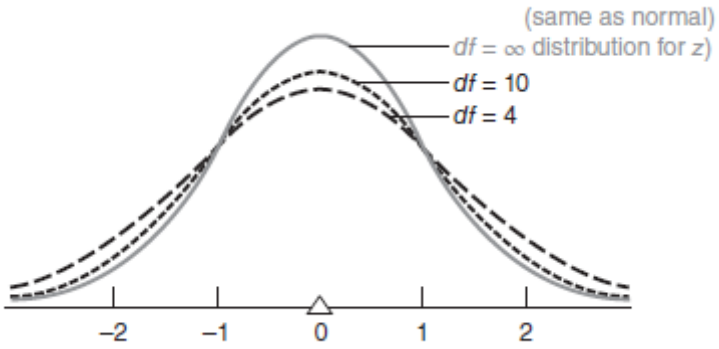


FIGURE 13.1

Various t distributions

from the upper half of each distribution. Because of the symmetry of the *t* distribution, you can obtain the corresponding critical *t* values for the lower half of each distribution merely by placing a negative sign in front of any entry in the table.

Finding Critical *t* Values

To find a critical *t* in Table B, read the entry in the cell intersected by the row for the correct number of degrees of freedom and the column for the test specifications. For example, to find the critical *t* for the gas mileage investigation, first go to the righthand panel for a one-tailed test, then locate both the row corresponding to five degrees of freedom and the column for a one-tailed test at the .01 level of significance. The intersected cell specifies 3.365. A negative sign must be placed in front of 3.365, since the hypothesis test requires the lower tail to be critical. Thus, –3.365 is the critical *t* for the gas mileage investigation, and the corresponding decision rule is illustrated in

Figure 13.2, where the distribution of *t* is centered about zero (the equivalent value of *t* for the original null hypothesized value of 45 mpg).

If the gas mileage investigation had involved a two-tailed test (still at the .01 level with five degrees of freedom), then the left-hand panel for a two-tailed test would have been appropriate, and the intersected cell would have specified 4.032. Both positive and negative signs would have to be placed in front of 4.032, since both tails are critical. In this case, 4.032 would have been the pair of critical *t* values.

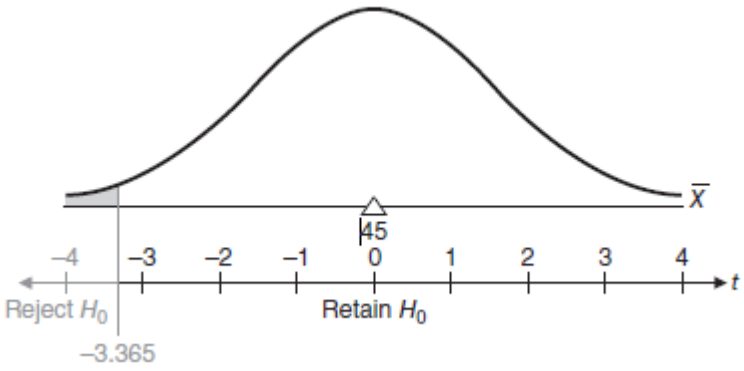


FIGURE 13.2

Hypothesized sampling distribution of t (gas mileage investigation).

Missing *df* in Table B of Appendix C

If the desired number of degrees of freedom doesn't appear in the *df* column of Table B, use the row in the table with the next smallest number of degrees of freedom. For example, if 36 degrees of freedom are specified, use the information from the row for 30 degrees of freedom. Always rounding off to the next smallest *df* produces a slightly larger critical *t*, making the null hypothesis slightly more difficult to reject. This procedure defuses potential disputes about borderline decisions by investigators with a stake in rejecting the null hypothesis

*t-test procedure

The TTEST Procedure Overview

The TTEST procedure performs t tests for one sample, two samples, and paired observations. The one-sample t test compares the mean of the sample to a given number. The two-sample t test compares the mean of the first sample minus the mean of the second sample to a given number. The paired observations t test compares the mean of the differences in the observations to a given number.

For one-sample tests, PROC TTEST computes the sample mean of the variable and compares it with a given number. Paired comparisons use the one sample process on the differences between the observations. Paired comparisons can be made between many pairs of variables with one call to PROC TTEST. For group comparisons, PROC TTEST computes sample means for each of two groups of observations and tests the hypothesis that the population means differ by a given amount. This latter analysis can be considered a special case of a one-way analysis of variance with two levels of classification.

The underlying assumption of the t test in all three cases is that the observations are random samples drawn from normally distributed populations. This assumption can be checked using the UNIVARIATE procedure; if the normality assumptions for the t test are not satisfied, you should analyze your data using the NPAR1WAY procedure. The two populations of a group comparison must also be independent. If they are not independent, you should question the validity of a paired comparison.

PROC TTEST computes the group comparison t statistic based on the assumption that the variances of the two groups are equal. It also computes an approximate t based on the assumption that the variances are unequal (the Behrens-Fisher problem). The degrees of freedom and probability level are given for each; Satterthwaite's (1946) approximation is used to compute the degrees of freedom associated with the approximate t. In addition, you can request the Cochran and Cox (1950) approximation of the

probability level for the approximate t. The folded form of the F statistic is computed to test for equality of the two variances (Steel and Torrie 1980).

FREQ and WEIGHT statements are available. Data can be input in the form of observations or summary statistics. Summary statistics and their confidence intervals, and differences of means are output. For two-sample tests, the pooled-variance and a test for equality of variances are also produced.

Getting Started

One-Sample t Test

A one-sample t test can be used to compare a sample mean to a given value. This example, taken from Huntsberger and Billingsley (1989, p. 290), tests whether the mean length of a certain type of court case is 80 days using 20 randomly chosen cases. The data are read by the following DATA step:

```
title 'One-Sample t Test';

data time; input time @@;

datalines;

43 90 84 87 116 95 86 99 93 92 121 71 66 98 79 102 60 112 105 98
;

run;
```

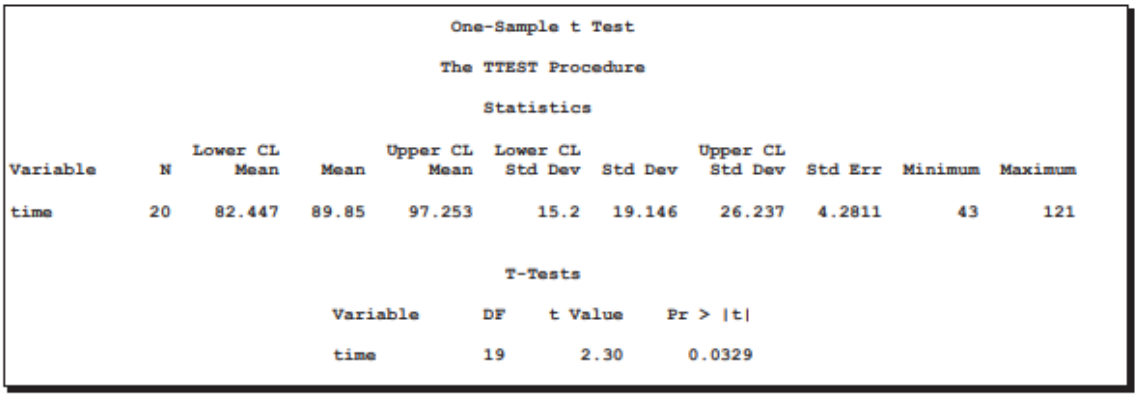
The only variable in the data set, time, is assumed to be normally distributed. The trailing at signs (@@) indicate that there is more than one observation on a line. The following code invokes PROC TTEST for a one-sample t test:

```
proc ttest h0=80 alpha=0.1;
```

var time;

run;

The VAR statement indicates that the time variable is being studied, while the H0= option specifies that the mean of the time variable should be compared to the value 80 rather than the default null hypothesis of 0. This ALPHA= option requests 10% confidence intervals rather than the default 5% confidence intervals. The output is displayed in Figure 67.1.



One-Sample t Test

The TTEST Procedure

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
time	20	82.447	89.85	97.253	15.2	19.146	26.237	4.2811	43	121

T-Tests

Variable	DF	t Value	Pr > |t|
time	19	2.30	0.0329

Figure 67.1. One-Sample t Test Results

Summary statistics appear at the top of the output. The sample size (N), the mean and its confidence bounds (Lower CL Mean and Upper CL Mean), the standard deviation and its confidence bounds (Lower CL Std Dev and Upper CL Std Dev), and the standard error are displayed with the minimum and maximum values of the time variable. The test statistic, the degrees of freedom, and the p-value for the t test are displayed next; at the 10% -level, this test indicates that the mean length of the court cases are significantly different from 80 days (t = 2:30; p = 0:0329).

Comparing Group Means

If you want to compare values obtained from two different groups, and if the groups are independent of each other and the data are

normally distributed in each group, then a group t test can be used. Examples of such group comparisons include

test scores for two third-grade classes, where one of the classes receives tutoring

fuel efficiency readings of two automobile nameplates, where each nameplate uses the same fuel

sunburn scores for two sunblock lotions, each applied to a different group of people

political attitude scores of males and females

In the following example, the golf scores for males and females in a physical education class are compared. The sample sizes from each population are equal, but this is not required for further analysis. The data are read by the following statements:

```
title ’Comparing Group Means’;

data scores; input Gender $ Score @@;

datalines;

 f 75 f 76 f 80 f 77 f 80 f 77 f 73 m 82 m 80 m 85 m 85 m 78 m 87 m 82 ;

run;
```

The dollar sign ($) following Gender in the INPUT statement indicates that Gender is a character variable. The trailing at signs (@@) enable the procedure to read more than one observation per line.

You can use a group t test to determine if the mean golf score for the men in the class differs significantly from the mean score for the women. If you also suspect that the distributions of the golf

scores of males and females have unequal variances, then submitting the following statements invokes PROC TTEST with options to deal with the unequal variance case.

```
proc ttest cochran ci=equal umpu;

class Gender;

var Score;

 run;
```

The CLASS statement contains the variable that distinguishes the groups being compared, and the VAR statement specifies the response variable to be used in calculations. The COCHRAN option produces p-values for the unequal variance situation using the Cochran and Cox(1950) approximation. Equal tailed and uniformly most powerful unbiased (UMPU) confidence intervals for are requested by the CI= option. Output from these statements is displayed in Figure 67.2 through Figure 67.4.

Comparing Group Means

The TTEST Procedure

Statistics

Variable	Class	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	UMPU Lower CL Std Dev	Std Dev
Score	f	7	74.504	76.857	79.211	1.6399	1.5634	2.5448
Score	m	7	79.804	82.714	85.625	2.028	1.9335	3.1472
Score	Diff (1-2)		-9.19	-5.857	-2.524	2.0522	2.0019	2.8619

Statistics

Variable	Class	UMPU Upper CL Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
Score	f	5.2219	5.6039	0.9619	73	80
Score	m	6.4579	6.9303	1.1895	78	87
Score	Diff (1-2)	4.5727	4.7242	1.5298		

Figure 67.2. Simple Statistics

Simple statistics for the two populations being compared, as well as for the difference of the means between the populations, are displayed in Figure 67.2. The Variable column denotes the response variable, while the Class column indicates the population corresponding to the statistics in that row. The sample size (N) for each population, the sample means (Mean), and lower and upper confidence bounds for the means (Lower CL Mean and Upper CL Mean) are displayed next. The standard deviations (Std Dev) are displayed as well, with equal tailed confidence bounds in the Lower CL Std Dev and Upper CL Std Dev columns and UMPU confidence bounds in the UMPU Upper CL Std Dev and UMPU Lower CL Std Dev columns. In addition, standard error of the mean and the minimum and maximum data values are displayed.

Comparing Group Means

The TTEST Procedure

T-Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
Score	Pooled	Equal	12	-3.83	0.0024
Score	Satterthwaite	Unequal	11.5	-3.83	0.0026
Score	Cochran	Unequal	6	-3.83	0.0087

Figure 67.3. t Tests

The test statistics, associated degrees of freedom, and p-values are displayed in Figure 67.3. The Method column denotes which t test is being used for that row, and the Variances column indicates what assumption about variances is being made. The pooled test assumes that the two populations have equal variances and uses degrees of freedom n1 +n22, where n1 and n2 are the sample sizes for the two populations. The remaining two tests do not assume that the populations have equal variances. The Satterthwaite test uses the Satterthwaite approximation for degrees of freedom, while the

Cochran test uses the Cochran and Cox approximation for the p-value

Comparing Group Means

The TTEST Procedure

Equality of Variances

Variable	Method	Num DF	Den DF	F Value	Pr > F
Score	Folded F	6	6	1.53	0.6189

Figure 67.4. Tests of Equality of Variance

Examine the output in Figure 67.4 to determine which t test is appropriate. The “Equality of Variances” test results show that the assumption of equal variances is reasonable for these data (the Folded F statistic F 0 = 1:53, with p = 0:6189). If the assumption of normality is also reasonable, the appropriate test is the usual pooled t test, which shows that the average golf scores for men and women are significantly different (t = 3:83; p = 0:0024). If the assumption of equality of variances is not reasonable, then either the Satterthwaite or the Cochran test should be used. The assumption of normality can be checked using PROC UNIVARIATE; if the assumption of normality is not reasonable, you should analyze the data with the nonparametric Wilcoxon Rank Sum test using PROC NPAR1WAY

Syntax

The following statements are available in PROC TTEST.

PROC TTEST < options > ;

CLASS variable ;

PAIRED variables ;

BY variables ;

VAR variables ;

FREQ variable ;

WEIGHT variable ;

No statement can be used more than once. There is no restriction on the order of the statements after the PROC statement.

PROC TTEST

Statement PROC TTEST < options > ;

The following options can appear in the PROC TTEST statement.

ALPHA=p

specifies that confidence intervals are to be 100(1 p)% confidence intervals, where 0 <p< 1. By default, PROC TTEST uses ALPHA=0.05. If p is 0 or less, or 1 or more, an error message is printed.

CI=EQUAL

CI=UMPU

CI=NONE

specifies whether a confidence interval is displayed for and, if so, what kind. The CI=EQUAL option specifies an equal tailed confidence interval, and it is the default. The CI=UMPU option specifies an interval based on the uniformly most powerful unbiased test of H0: = 0. The CI=NONE option requests that no confidence interval be displayed for . The values EQUAL and UMPU together request t that both types of confidence intervals be displayed. If the value NONE is specified with one or both of the values EQUAL and UMPU, NONE takes precedencthat both types of confidence

intervals be displayed. If the value NONE is specified with one or both of the values EQUAL and UMPU, NONE takes precedence.

COCHRAN

requests the Cochran and Cox (1950) approximation of the probability level of the approximate t statistic for the unequal variances situation.

DATA=SAS-data-set

names the SAS data set for the procedure to use. By default, PROC TTEST uses the most recently created SAS data set. The input data set can contain summary statistics of the observations instead of the observations themselves. The number, mean, and standard deviation of the observations are required for each BY group (one sample and paired differences) or for each class within each BY group (two samples). For SAS OnlineDoc□: Version CLASS Statement 3575 more information on the DATA= option, see the “Input Data Set of Statistics”

H0=m

requests tests against m instead of 0 in all three situations (one-sample, two-sample, and paired observation t tests). By default, PROC TTEST uses H0=0.

BY Statement

BY variables ;

You can specify a BY statement with PROC TTEST to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables.

If your input data set is not sorted in ascending order, use one of the following alternatives:

Sort the data using the SORT procedure with a similar BY statement.

Specify the BY statement option NOTSORTED or DESCENDING in the BY statement for the TTEST procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (ac- cording to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.

Create an index on the BY variables using the DATASETS procedure (in base SAS software).

For more information on the BY statement, refer to the discussion in *SAS Language Reference: Concepts*. For more information on the DATASETS procedure, refer to the *SAS Procedures Guide*.

CLASS Statement

CLASS *variable* ;

A CLASS statement giving the name of the classification (or grouping) variable must accompany the PROC TTEST statement in the two independent sample cases. It should be omitted for the one sample or paired comparison situations. If it is used without the VAR statement, all numeric variables in the input data set (except

those appearing in the CLASS, BY, FREQ, or WEIGHT statement) are included in the analysis.

The class variable must have two, and only two, levels. PROC TTEST divides the observations into the two groups for the *t* test using the levels of this variable. You can use either a numeric or a character variable in the CLASS statement.

Class levels are determined from the formatted values of the CLASS variable. Thus, you can use formats to define group levels. Refer to the discussions of the FOR- MAT procedure, the FORMAT statement, formats, and informats in *SAS Language Reference: Dictionary*.

FREQ Statement

FREQ *variable* ;

The *variable* in the FREQ statement identifies a variable that contains the frequency of occurrence of each observation. PROC TTEST treats each observation as if it appears *n* times, where *n* is the value of the FREQ variable for the observation. If the value is not an integer, only the integer portion is used. If the frequency value is less than 1 or is missing, the observation is not used in the analysis. When the FREQ statement is not specified, each observation is assigned a frequency of 1. The FREQ statement cannot be used if the DATA= data set contains statistics instead of the original observations.

PAIRED Statement

PAIRED *PairLists* ;

The *PairLists* in the PAIRED statement identifies the variables to be compared in paired comparisons. You can use one or more *PairLists*. Variables or lists of variables are separated by an asterisk (*) or a colon (:). The asterisk requests comparisons between each variable on the left with each variable on the right. The colon requests comparisons between the first variable on the left and the first on the right, the second on the left and the second on the right, and so forth. The number of variables on the left must equal the number on the right when the colon is used. The differences are calculated by taking the variable on the left minus the variable on the right for both the asterisk and colon. A pair formed by a variable with itself is ignored. Use the PAIRED statement only for paired comparisons. The CLASS and VAR statements cannot be used with the PAIRED statement.

Examples of the use of the asterisk and the colon are shown in the following table.

These PAIRED statements...	yield these comparisons
PAIR A*B; ED	A-B
PAIR A*B C*D; ED	A-B and C-D
PAIR (A B)*(C D); ED	A-C, A-D, B-C, and B-D
PAIR (A B)*(C B); ED	A-C, A-B, and B-C
PAIR (A1-A2)*(B1-	A1-B1, A1-B2, A2-

ED	B2);	B1, and A2-B2
PAIR ED	(A1-A2):(B1-B2);	A1-B1 and A2-B2

VAR Statement

VAR *variables* ;

The VAR statement names the variables to be used in the analyses. One-sample comparisons are conducted when the VAR statement is used without the CLASS statement, while group comparisons are conducted when the VAR statement is used with a CLASS statement. If the VAR statement is omitted, all numeric variables in the input data set (except a numeric variable appearing in the BY, CLASS, FREQ, or WEIGHT statement) are included in the analysis. The VAR statement can be used with one- and two-sample *t* tests and cannot be used with the PAIRED statement.

WEIGHT Statement

WEIGHT *variable* ;

The WEIGHT statement weights each observation in the input data set by the value of the WEIGHT variable. The values of the WEIGHT variable can be nonintegral, and they are not truncated. Observations with negative, zero, or missing values for the WEIGHT variable are not used in the analyses. Each observation is assigned a

weight of 1 when the WEIGHT statement is not used. The WEIGHT statement cannot be used with an input data set of summary statistics.

Details

Input Data Set of Statistics

PROC TTEST accepts data containing either observation values or summary statis- tics. It assumes that the DATA= data set contains statistics if it contains a charac- ter variable with name _TYPE_ or _STAT_. The TTEST procedure expects this character variable to contain the names of statistics. If both _TYPE_ and _STAT_ variables exist and are of type character, PROC TTEST expects _TYPE_ to contain the names of statistics including 'N', 'MEAN', and 'STD' for each BY group (or for each class within each BY group for two-sample *t* tests). If no 'N', 'MEAN', or 'STD' statistics exist, an error message is printed.

FREQ, WEIGHT, and PAIRED statements cannot be used with input data sets of statistics. BY, CLASS, and VAR statements are the same regardless of data set type. For paired comparisons, see the _DIF_ values for the _TYPE_=T observations in

output produced by the OUTSTATS= option in the PROC COMPARE statement (re-

fer to the *SAS Procedures Guide*).

Missing Values

An observation is omitted from the calculations if it has a missing value for either the CLASS variable, a PAIRED variable, or the variable to be tested. If more than

one variable is listed in the VAR statement, a missing value in one variable does not eliminate the observation from the analysis of other nonmissing variables.

Computational Methods

The t Statistic

The form of the *t* statistic used varies with the type of test being performed.

To compare an individual mean with a sample of size *n* to a value *m*, use

$$t = \frac{\bar{x} - m}{s / \sqrt{n}}$$

where x is the sample mean of the observations and s^2 is the sample variance of the observations.

To compare *n* paired differences to a value *m*, use

$$t = \frac{\bar{d} - m_d}{s / \sqrt{n}}$$

where d is the sample mean of the paired differences and s^2 is the sample

variance of the paired differences.

To compare means from two independent samples with n_1 and n_2 observations to a value m, use

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - m}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where s^2 is the pooled variance

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and s_1^2 and s_2^2 are the sample variances of the two groups. The use of this t statistic depends on the assumption that $\sigma_1^2 = \sigma_2^2$, where σ_1^2 and σ_2^2 are the population variances of the two groups.

The Folded Form F Statistic

The folded form of the F statistic, F_0, tests the hypothesis that the variances are equal, where

$$F_0 = \frac{\max(S_1^2; S_2^2)}{\min(S_1^2; S_2^2)}$$

A test of F is a two-tailed F test because you do not specify which variance you expect to be larger. The p-value gives the probability of a greater F value under the

null hypothesis that $^2_1 = {}^2_2$.

The Approximate t Statistic

Under the assumption of unequal variances, the approximate t statistic is computed as

$$t = \frac{x_1 - x_2}{\sqrt{\frac{1}{p_w}\left(\frac{s_1^2}{w_1} + \frac{s_2^2}{w_2}\right)}}$$

where

$$w_1 = \frac{s_1^2}{n_1}; \quad w_2 = \frac{s_2^2}{n_2}$$

The Cochran and Cox Approximation

The Cochran and Cox (1950) approximation of the probability level of the approxi- mate *t* statistic is the value of *p* such that

$$t = \frac{w_1 t_1 + w_2 t_2}{w_1 + w_2}$$

where t_1 and t_2 are the critical values of the *t* distribution corresponding to a signifi- cance level of *p* and sample sizes of n_1 and n_2, respectively. The number of degrees of freedom is undefined when $n_1 = n_2$. In general, the Cochran and Cox test tends to be conservative (Lee and Gurland 1975).

Satterthwaite's Approximation

The formula for Satterthwaite's (1946) approximation for the degrees of freedom for the approximate *t* statistic is:

$$df = \frac{(w_1 + w_2)^2}{\frac{w_1^2}{n_1 - 1}}$$

Refer to Steel and Torrie (1980) or Freund, Littell, and Spector (1986) for more in- formation.

Confidence Interval Estimation

The form of the confidence interval varies with the statistic for which it is computed. In the following confidence intervals involving means, $t_{1\ ;n\ 1}$ is the $100(1\ \ _2)$% quantile of the t distribution with $n\ \ 1$ degrees of freedom. The confidence interval for

an individual mean from a sample of size n compared to a value m is given by

$$s\ (x\quad m)\quad t_{1\quad ;n\ \ 1} p$$

$$_2 \qquad - \qquad \overline{n}$$

where x is the sample mean of the observations and s^2 is the sample variance of the observations

paired differences with a sample of size n differences compared to a value m

is given by

$$\frac{s_d}{(d - m) t_{1;n-1} \frac{p}{\sqrt{n}}}$$

where d and s^2 are the sample mean and sample variance of the paired differ- ences, respectively

the difference of two means from independent samples with n_1 and n_2 obser- vations compared to a value m is given by

$$\frac{}{((x_1 - x_2) - m) t}$$

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where s^2 is the pooled variance

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

and where s^2 and s^2 are the sample variances of the two groups. The use of

1 2

this confidence interval depends on the assumption that $^2 = ^2$, where ^{2} and

1 2 1

^{2} are the popu1lation variances of the two groups.

The distribution of the estimated standard deviation of a mean is not symmetric, so alternative methods of estimating confidence intervals are possible. PROC TTEST computes two estimates. For both methods, the data are assumed to have a normal distribution with mean and variance ^{2}, both unknown. The methods are as follows:

The default method, an equal-tails confidence interval, puts an equal amount of area (^{2}) in each tail of the chi-square distribution. An equal tails test of $H_0: = _0$ has acceptance region

$$^2 \quad \frac{(n \ 1)S}{^2} \quad ^2$$

$$^2 \qquad ^2$$

$$_{-;n\,1} \qquad ^2 \qquad ^1 _{-;n \ 1}$$

which can be algebraically manipulated to give the following $100(1-\alpha)\%$ confidence interval for σ^2:

$$\left(\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}} \; ; \; \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}} \right)$$

In order to obtain a confidence interval for σ, the square root of each side is taken, leading to the following $100(1-\alpha)\%$ confidence interval:

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2};n-1}}} \; ; \; \sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2};n-1}}} \right)$$

The second method yields a confidence interval derived from the uniformly most powerful unbiased test of $H_0: = {}_0$ (Lehmann 1986). This test has acceptance region

$$c \frac{(n\ 1)S}{\ } c \qquad {}_1 \qquad {}_2 \qquad {}^2 \qquad {}_2$$

where the critical values c_1 and c_2 satisfy

$$\int_{c_1}^{c_2} f_n(y)dy = 1$$

and

$$\int_{c_1}^{c_2} y f_n(y)dy = n(1\ )$$

where $f_n(y)$ is the chi-squared distribution with n degrees of freedom. This acceptance region can be algebraically manipulated to arrive at

$$P \frac{(n\ \ 1)S}{{}_2} \ {}_2 \frac{(n\ \ 1)S}{c_2} = 1 \qquad c_1$$

where c_1 and c_2 solve the preceding two integrals. To find the area in each tail of the chi-square distribution to which these two critical values correspond,

solve $c_1 =$ ^{2} $_{1\ ;n\ 1}$ and $c_2 =$ ^{2} $_{;n\ 1}$

^{2} sum to . Hence, a $100(1\ )\%$ confidence interval for ^{2} is given by

$$\frac{(n\ \ 1)S^2}{\ ^{2}\ _{1\ ;n\ 1}} \ ; \ \frac{(n\ \ 1)S^2}{\ ^{2}\ _{\ ;n\ 1}}$$

In order to obtain a $100(1\ )\%$ confidence interval for , the square root is taken of both terms, yielding

$$\sqrt{\frac{(n\ \ 1)S^2}{\ ^{2}\ _{1\ ;n\ 1}}} \ ; \ \sqrt{\frac{(n\ \ 1)S^2}{\ ^{2}\ _{\ ;n\ 1}}}$$

Displayed Output

For each variable in the analysis, the TTEST procedure displays the following sum- mary statistics for each group:

the name of the dependent variable

the levels of the classification variable

N, the number of nonmissing values

Lower CL Mean, the lower confidence bound for the mean

the Mean or average

Upper CL Mean, the upper confidence bound for the mean

Lower CL Std Dev, the lower confidence bound for the standard deviation

Std Dev, the standard deviation

Upper CL Std Dev, the upper confidence bound for the standard deviation

Std Err, the standard error of the mean

the Minimum value, if the line size allows

the Maximum value, if the line size allows

upper and lower UMPU confidence bounds for the standard deviation, dis- played if the CI=UMPU option is specified in the PROC TTEST statement

Next, the results of several *t* tests are given. For one-sample and paired observations

t tests, the TTEST procedure displays

t Value, the *t* statistic for testing the null hypothesis that the mean of the group is zero

DF, the degrees of freedom

Pr > |t|, the probability of a greater absolute value of *t* under the null hypothesis. This is the two-tailed significance probability. For a one-tailed test, halve this probability.

For two-sample *t* tests, the TTEST procedure displays all the items in the following list. You need to decide whether equal or unequal variances are appropriate for your data.

Under the assumption of unequal variances, the TTEST procedure displays results using Satterthwaite's method. If the COCHRAN option is specified, the results for the Cochran and Cox approximation are also displayed.

t Value, an approximate *t* statistic for testing the null hypothesis that the means of the two groups are equal

DF, the approximate degrees of freedom

Pr > |t|, the probability of a greater absolute value of *t* under the null hypothesis. This is the two-tailed significance probability. For a one- tailed test, halve this probability.

Under the assumption of equal variances, the TTEST procedure displays results obtained by pooling the group variances.

t Value, the *t* statistic for testing the null hypothesis that the means of the two groups are equal

* DF, the degrees of freedom

Pr > |t|, the probability of a greater absolute value of *t* under the null hypothesis. This is the two-tailed significance probability. For a one- tailed test, halve this probability.

PROC TTEST then gives the results of the test of equality of variances:

the *F* ′ (folded) statistic (see the "The Folded Form F Statistic" section on page 3578)

Num DF and Den DF, the numerator and denominator degrees of freedom in each group

Pr > F, the probability of a greater *F* ′ value. This is the two-tailed signif- icance probability.

ODS Table Names

PROC TTEST assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 67.1. ODS Tables Produced in PROC TTEST

ODS Table Name	Description	Statement
Equality Statistics	Tests for equality of variance	CLASS statement
TTests	Univariate summary statistics *t*-tests	by default by default

Examples

Example 67.1. Comparing Group Means Using Input Data Set of Summary Statistics

The following example, taken from Huntsberger and Billingsley (1989), compares two grazing methods using 32 steer. Half of the steer are allowed to graze continu- ously while the other half are subjected to controlled grazing time. The researchers want to know if these two grazing methods impact weight gain differently. The data are read by the following DATA step.

```
title 'Group Comparison Using Input Data Set of Summary Statistics';

data graze;

length GrazeType $ 10; input GrazeType $ WtGain @@; datalines;

controlled   45   controlled   62
controlled   96   controlled   128
controlled   120  controlled   99
controlled   28   controlled   50
controlled   109  controlled   115
controlled   39   controlled   96
controlled   87   controlled   100
controlled   76   controlled   80
continuous   94   continuous   12
continuous   26   continuous   89
continuous   88   continuous   96
continuous   85   continuous   130
continuous   75   continuous   54
continuous   112  continuous   69
continuous   104  continuous   95
continuous   53   continuous   21
;
run;
```

The variable GrazeType denotes the grazing method: ‘controlled’ is controlled graz- ing and ‘continuous’ is continuous grazing. The dollar sign ($) following GrazeType makes it a character variable, and the trailing at signs (@@) tell the procedure that there is more than one observation per line. The MEANS procedure is invoked to create a data set of summary statistics with the following statements:

```
proc sort;

by GrazeType;

proc means data=graze noprint; var WtGain;

by GrazeType;

output out=newgraze; run;
```

The NOPRINT option eliminates all output from the MEANS procedure. The VAR statement tells PROC MEANS to compute summary statistics for the WtGain vari- able, and the BY statement requests a separate set of summary statistics for each level of GrazeType. The OUTPUT OUT= statement tells PROC MEANS to put the sum- mary statistics into a data set called newgraze so that it may be used in subsequent procedures. This new data set is displayed in Output 67.1.1 by using PROC PRINT as follows:

```
proc print data=newgraze; run;
```

The _STAT_ variable contains the names of the statistics, and the GrazeType vari- able indicates which group the statistic is from.

Output 67.1.1. Output Data Set of Summary Statistics

Group Comparison Using Input Data Set of Summary Statistics

Obs	GrazeType	_TYPE_	_FREQ_	_STAT_	WtGain
1	continuous	0	16	N	16.000
2	continuous	0	16	MIN	12.000
3	continuous	0	16	MAX	130.000
4	continuous	0	16	MEAN	75.188
5	continuous	0	16	STD	33.812
6	controlled	0	16	N	16.000
7	controlled	0	16	MIN	28.000
8	controlled	0	16	MAX	128.000
9	controlled	0	16	MEAN	83.125
10	controlled	0	16	STD	30.535

The following code invokes PROC TTEST using the newgraze data set, as denoted by the DATA= option.

```
proc ttest data=newgraze; class GrazeType;

var WtGain; run;
```

The CLASS statement contains the variable that distinguishes between the groups being compared, in this case GrazeType. The summary statistics and confidence intervals are displayed first, as shown in Output 67.1.2.

Output 67.1.2. Summary Statistics

Group Comparison Using Input Data Set of Summary Statistics The TTEST Procedure

Statistics

Variable	Class	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
WtGain	continuous	16	57.171	75.188	93.204	.	33.812	.	8.4529	12	130
WtGain	controlled	16	66.854	83.125	99.396	.	30.535	.	7.6337	28	128
WtGain	Diff (1-2)		-31.2	-7.938	15.323	25.743	32.215	43.061	11.39		

In Output 67.1.2, the Variable column states the variable used in computations and the Class column specifies the group for which the statistics are computed. For each class, the sample size, mean, standard deviation and standard error, and maximum and minimum values are displayed. The confidence bounds for the mean are also displayed; however, since summary statistics are used as input, the confidence bounds for the standard deviation of the groups are not calculated.

Output 67.1.3. t Tests

Variable	Method	Variances	DF	t Value	Pr > |t|
WtGain	Pooled	Equal	30	-0.70	0.4912
WtGain	Satterthwaite	Unequal	29.7	-0.70	0.4913

Group Comparison Using Input Data Set of Summary Statistics

Variable	Method	Num DF	Den DF	F Value	Pr > F
WtGain	Folded F	15	15	1.23	0.6981

Equality of Variances

Output 67.1.3 shows the results of tests for equal group means and equal variances. A group test statistic for the equality of means is reported for equal and unequal variances. Before deciding which test is appropriate, you should look at the test for equality of variances; this test does not indicate a significant difference in the two variances $(F' = 1.23; p = 0.6981)$, so the pooled t statistic should be used. Based on the pooled statistic, the two grazing methods are not significantly different $(t = 0.70; p = 0.4912)$. Note that this test assumes that the observations in both data sets are normally distributed; this assumption can be checked in PROC UNIVARIATE using the raw data.

Example 67.2. One-Sample Comparison Using the FREQ Statement

This example examines children's reading skills. The data consist of Degree of Read- ing Power (DRP) test scores from 44 third-grade children and are taken from Moore (1995, p. 337). Their scores are given in the following DATA step.

```
title 'One-Mean Comparison Using FREQ Statement'; data read;

input score count @@; datalines;

40 2    47 2    52 2    26 1    19 2
25 2    35 4    39 1    26 1    48 1
14 2    22 1    42 1    34 2    33 2
18 1    15 1    29 1    41 2    44 1
51 1    43 1    27 2    46 2    28 1
49 1    31 1    28 1    54 1    45 1
;
run;
```

The following statements invoke the TTEST procedure to test if the mean test score is equal to 30. The count variable contains the frequency of occurrence of each test score; this is specified in the FREQ statement.

```
proc ttest data=read h0=30; var score;

freq count; run;
```

The output, shown in Output 67.2.1, contains the results.

Output 67.2.1. TTEST Results

One-Mean Comparison Using FREQ Statement

Frequency: count

Statistics

Variable	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
score	44	31.449	34.864	38.278	9.2788	11.23	14.229	1.693	14	54

Variable	DF	t Value	Pr > |t|
score	43	2.87	0.0063

T-Tests

The SAS log states that 30 observations and two variables have been read. However, the sample size given in the TTEST output is N=44. This is due to specifying the count variable in the FREQ statement. The test is significant $(t = 2{:}87,\ p = 0{:}0063)$ at the 5% level, thus you can conclude that the mean test score is different from 30.

Example 67.3. Paired Comparisons

When it is not feasible to assume that two groups of data are independent, and a natural pairing of the data exists, it is advantageous to use an analysis that takes the correlation into account. Utilizing this correlation results in higher power to detect existing differences between the means. The differences between paired observations are assumed to be normally distributed. Some examples of this natural pairing are

pre- and post-test scores for a student receiving tutoring

fuel efficiency readings of two fuel types observed on the same automobile

sunburn scores for two sunblock lotions, one applied to the individual's right arm, one to the left arm

political attitude scores of husbands and wives

In this example, taken from *SUGI Supplemental Library User's Guide, Version 5 Edi- tion,* a stimulus is being examined to determine its effect on systolic blood pressure. Twelve men participate in the study. Their systolic blood pressure is measured both before and after the stimulus is applied. The following statements input the data:

```
title 'Paired Comparison'; data pressure;

input SBPbefore SBPafter @@; datalines;

        120 128    124 131    130 131    118 127
        140 132    128 125    140 141    135 137
        126 118    130 132    126 129    127 135

        ;

        run;
```

The variables SBPbefore and SBPafter denote the systolic blood pressure before and after the stimulus, respectively.

The statements to perform the test follow.

```
proc ttest;

paired SBPbefore*SBPafter; run;
```

The PAIRED statement is used to test whether the mean change in systolic blood pressure is significantly different from zero. The output is displayed in Output 67.3.1.

Output 67.3.1. TTEST Results

Paired Comparison The TTEST Procedure

Statistics

Difference	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum	
SBPbefore - SBPafter	12	-5.536	-1.833	1.8698	4.1288	5.8284	9.8958	1.6825	-9	8	

T-Tests

Difference	DF	t Value	Pr > |t|
SBPbefore - SBPafter	11	-1.09	0.2992

The variables SBPbefore and SBPafter are the paired variables with a sample size of 12. The summary statistics of the difference are displayed (mean, standard de- viation, and standard error) along with their confidence limits. The minimum and maximum differences are also displayed. The *t* test is not significant $(t = 1{:}09;\ p = 0{:}2992)$, indicating that the stimuli did not significantly affect systolic blood pressure.

Note that this test of hypothesis assumes that the differences are normally distributed. This assumption can be investigated using PROC UNIVARIATE with the NORMAL option. If the assumption is not satisfied, PROC NPAR1WAY should be used.

Degrees of freedom:

The notion of degrees of freedom is used throughout the remainder of this book. Typically, when it is used to estimate some unknown population characteristic, not all observed values within the sample are free to vary. For example, the gas mileage data consist of six values: 40, 44, 46, 41, 43, and 44. Nevertheless, the *t* test for these data has only five degrees of freedom because of the zero-sum restriction. Only five of these six observed values are free to vary about their mean of 43 and, therefore, provide valid information for purposes of estimation. The *concept of degrees of freedom is introduced only because we are using observations in a sample to estimate some unknown characteristic of the population.*

In subsequent sections, we'll encounter other mathematical restrictions, and sometimes several degrees of freedom will be lost. In any event, however, the degrees of freedom always indicate the number of values free to vary, given one or more mathematical

restrictions on a set of values used to *estimate* some unknown population characteristic.

*estimating the standard error

If the population standard deviation is unknown, it must be estimated from the sample. This seemingly minor complication has important implications for hypothesis testing—indeed, it is the reason why the *z* test must be replaced by the *t* test. Now *s* replaces σ in the formula for the standard error of the mean. Instead of

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

*Essentially, the inflated tails are caused by the extra variability of the estimated standard error in the denominator of *t*.

ESTIMATED STANDARD ERROR OF THE MEAN

$$s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

where *x s* represents the estimated standard error of the mean; *n* equals the sample size; and *s* has been defined as

$$s = \sqrt{\frac{SS}{n-1}} = \sqrt{\frac{SS}{df}}$$

where *s* is the sample standard deviation; *df* refers to the degrees of freedom; and *SS* has been defined as

$$SS = \Sigma(X - \bar{X})^2 = \Sigma X^2 - \frac{(\Sigma X)^2}{n}$$

This new version of the standard error, the estimated standard error of the mean, *is used whenever the unknown population standard deviation must be estimated*.

t-test for two independent samples:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In statistics, the two sample t-test for independent samples is a type of hypothesis test that can be used to determine whether the means of two populations are statistically different given the two samples are independent and have normal distributions. As data scientists, it is important to understand how to use the two sample t-test for independent samples so that you can correctly analyze your data. In this blog post, we will discuss the two sample t-test for independent samples in detail, including the formula and examples.

Table of Contents

*What is two-sample T-test?

A two-sample T-test is defined as statistical hypothesis testing technique in which two independent samples are compared to determine if the means of two populations are statistically different. The two-sample T-test is used when the standard deviations of the populations to be compared are unknown and the sample size is small. The size of sample 30 or less is considered as small sample. That said, the size of the sample is not a strict condition for using T-test. The two-sample T-test is used when the two samples are independent and have normal distributions. In order to use a two-sample T-test as desccribed in this blog, you need to have two independent samples. The independent samples mean that the ***two samples cannot be from the same group of people and they cannot be related in any way***. However, two-sample T-test can also be used for pairwise comparisons when the "two" samples represent the same items tested in different scenarios. The pairwise t-test will be dealt with in different blog.

Let's say you want to know if two different brands of batteries have the same average life. You could take a battery from each brand, use them until they die, and record the results. This would be an extremely time-consuming process, and it's not very likely that you'd get a large enough sample size to draw any conclusions. Another option is to use a two-sample T-test. This test

allows you to compare the averages of two groups without having to measure the batteries' life spans yourself.

The following are a few real-life examples where two-sample T-test for independent samples can be used:

Comparing the average test scores of two classes from two different schools

Comparing the average weights of two different oor independent groups of people

Determining whether the medication have the same efficacy on two different or independent groups of people

Compare whether the effect of vaccination on two different groups

T-statistics when population variances or standard deviations are unequal

The formula for T-statistics is different based on whether the populations' standard deviation are same / equal or different. When the standard deviations of populations are not equal, the following formula is used to calculate the T-statistics and degrees of freedom.

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Where $\bar{X}1$ is mean of first sample, $\bar{X}2$ is mean of second sample, $\mu 1$ is the mean of first population, $\mu 2$ is the mean of second population, s1 is the standard deviation of first sample, s2 is the standard deviation of second sample, n1 is the size of the first sample, n2 is the size of the second sample.

The degrees of freedom can be calculated as the sum of two sample sizes minus two.

Degrees of freedom, df = n1 + n2 – 2

A confidence interval for the difference between two means specifies a range of values within which the difference between the means of the two populations may lie. The difference between the means of two populations can be estimated based on the following formula:

$$(\overline{x_1} - \overline{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Difference in population means = Difference in sample means +/- T*standard error

T-statistics when population variances or standard deviations are equal

In case, the two populations' standard deviations are equal, the formula termed as pooled t-statistics is used based on the usage of pooled standard deviations of the two samples. The following is the formula for the pooled t-statistics:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

In the above formula, Sp is termed as pooled standard deviation. The formula for pooled variance can be calculated based on the following:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

The value for the degree of freedom can be calculated as the sum of two sample sizes minus two.

Degrees of freedom, df = n1 + n2 – 2

When two-sample T-test instead of two-sample Z-test?

When the population standard deviations are known and the sample size is large, we go for two-sample Z-test for comparing the two different populations. The sample size greater than 30 is considered to be large sample size. Otherwise, a two-sample T-test is known with T-distribution and a given degrees of freedom.

Two-sample T-test: Examples

Lets say we need to compare the performance of two call centers in terms of average call lengths and find out if the difference is statistically significant or the difference is a chance occurrence. To start with, we will need to formulate the null and alternate hypothesis.

Null hypothesis, H0: There is no difference between the average call length between two call centers.

Alternate hypothesis, Ha: There is a difference between the average call length and hence the performance.

We randomly select 20 calls from each call center and calculate the average call lengths. The two call centers seem to have different average call lengths. Is this difference statistically significant?

First, we need to calculate the two sample means and standard deviations:

Call Center A: Sample mean, $\bar{X}1 = 122$ seconds, SD, S1 = 15 seconds, n1 = 20

Call Center B: Sample mean, $\bar{X}2 = 135$ seconds, SD, S2 = 20 seconds, n2 = 20

Next, we use a two-sample t-test to determine if the difference between two sample means is statistically significant. We will use a 95% confidence level and $\alpha = 0.05$.

The two-sample t-statistic is calculated as the following assuming that the standard deviations of the population is not same and the population mean is same.

t = ((135 – 122) – 0)/SQRT((20*20/20) + ((15*15)/20))

t = 13/SQRT(20 + 11.25)

t = 13/SQRT(31.25)

t = 2.3256

The value of degrees of freedom can be calculated as the following:

Degree of freedom, df = n1 + n2 -2 = 20 + 20 – 2 = 38

The critical value of a two-tailed T-test with degrees of freedom as 38 and level of significance as 0.05 comes out to be 2.0244. Since the current t-value of 2.3256 is greater than the critical value of 2.0244, one can reject the null hypothesis that there is no difference between the performance in terms of the call length time. Thus, based on the given evidence, the alternate hypothesis stands as true.

Summary

The two-sample t-test for independent samples is a statistical method for comparing two different populations. The t-test can be used when the population standard deviations are not known and the sample size is smaller (less than 30). The two sample t-statistic calculation depends on given degrees of freedom, df = n1 + n2 – 2. If the value of two samples t-test for independent samples exceeds critical T at alpha level, then you can reject null hypothesis that there is no difference between two data sets (H0). Otherwise if two sample T-statistics is less than or equal to critical T at alpha level, then one cannot reject H0; this means both values could have come from same distribution in which case any observed difference would be due to chance alone. Different formulas are required to be used for performing t-test for two independent samples based on whether the variances of two populations are equal or otherwise

*statistical hypotheses:

Null Hypothesis

According to the null hypothesis, nothing special is happening because EPO does

not facilitate endurance. In other words, either there is no difference between the means

for the two populations (because EPO has no effect on endurance) or the difference

between population means is negative (because EPO hinders endurance). An equivalent

statement in symbols reads:

$H_0 : \mu_1 - \mu_2 \leq 0$

where H_0 represents the null hypothesis and μ_1 and μ_2 represent the mean endurance

scores for the treatment and control populations, respectively.

Alternative (or Research) Hypothesis

The investigator wants to reject the null hypothesis only if the treatment increases

endurance scores. Given this perspective, the alternative (or research) hypothesis

should specify that the difference between population means is positive because EPO

facilitates endurance. An equivalent statement in symbols reads:

$H_1 : {}_1 {}_2 0$

where H_1 represents the alternative hypothesis and, as above, μ_1 and μ_2 represent the mean endurance scores for the treatment and control populations, respectively. This directional alternative hypothesis translates into a one-tailed test with the upper tail critical. As emphasized in Section 11.3, a directional alternative hypothesis should be used when there's a concern *only* about differences in a particular direction.

Two Other Possible Alternative Hypotheses

Although not appropriate for the current experiment, there are two other possible alternative hypotheses:

1. Another directional hypothesis, expressed as

$H_1 : \mu_1 - \mu_2 < 0$

translates into a one-tailed test with the lower tail critical.

2. A nondirectional hypothesis, expressed as

$H_1 : \mu_1 - \mu_2 \neq 0$

translates into a two-tailed test.

*sampling distribution:

SAMPLING DISTRIBUTION OF $X_1 - X_2$

Because of the inevitable variability associated with any difference between the sample mean endurance scores for the treatment and control groups, $X_1 -$

X_2, we can't

interpret a single observed mean difference at face value. The new mean difference for a repeat experiment would most likely differ from that for the original experiment

The sampling distribution of $X_1 -$

X_2 is a concept introduced to account for the

variability associated with differences between sample means. *It represents the entire*

spectrum of differences between sample means based on all possible pairs of random samples from the two underlying populations. Once the sampling distribution has been centered about the value of the null hypothesis, we can determine whether the one observed sample mean difference qualifies as a common or a rare outcome. (A common outcome signifies that the observed sample mean difference could be due to variability or chance and, therefore, shouldn't be taken seriously. On the other hand, a rare outcome signifies that the observed sample mean difference probably reflects a real difference and, therefore, should be taken seriously.) Since all the possible pairs of random samples usually translate into a huge number of possibilities—often of astronomical proportions—the sampling distribution of $X_1 - X_2$ isn't constructed from scratch. As with the sampling distribution of X described in Chapter 9, statistical theory must be relied on for information about the mean and standard error for this new sampling distribution.

Mean of the Sampling Distribution, $\mu_{X_1 X_2}$

Recall from Chapter 9 that the mean of the sampling distribution of X equals the population mean, that is,

$$\mu_X = \mu$$

where μ_X is the mean of the sampling distribution and μ is the population mean. Similarly, the mean of the new sampling distribution of $X_1 - X_2$ equals the difference between population means, that is,

$$\mu_{X_1 - X_2} = \mu_1 - \mu_2$$

where

$\mu_{X_1 - X_2}$ is the mean of the new sampling distribution and $\mu_1 - \mu_2$ is the difference between population means. This conclusion is not particularly startling. Because of sampling variability, it's unlikely that the one observed difference between sample

means equals the difference between population means. Instead, it's likely that, just by chance, the one observed difference is either larger or smaller than the difference between population means. However, because not just one but all possible differences between sample means contribute to the mean of the sampling distribution, $\mu_{X_1 X_2}$, the effects of sampling variability are neutralized, and the mean of the sampling distribution equals the difference between population means. Accordingly, these two terms are used interchangeably. Any claims about the difference between population means, including the null hypothesized claim that this difference equals zero, can be transferred directly to the mean of the sampling distribution.

Standard Error of the Sampling Distribution, $\sigma_{X_1 X_2}$

Also recall from Chapter 9 that the standard deviation of the sampling distribution (or standard error) of X equals

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sigma^2}{n}}$$

where $_{\bar{x}}$ is the standard error, is the population standard deviation, and n is the sample size. To highlight the similarity between this expression and that for the new sampling distribution, the population variance, ^{2}, is introduced in the above equation by placing both the numerator and denominator under a common square root sign.

The standard deviation of the new sampling distribution of $X_1 X_2$ equals

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where $\sigma_{\bar{X}_1-\bar{X}_2}$ is the new standard error, σ_1^2 and σ_2^2 are the two population variances, and n_1 and n_2 are the two sample sizes.

The new standard error emerges directly from the original standard error with the addition of a second term, σ_2^2 divided by n_2, reflecting extra variability due to the shift from a single sample mean to differences between two sample means. Therefore, the value of the new standard error always will be larger than that of the original one. The original standard error reflects only the variability of single sample means about the mean of their sampling distribution. But the new standard error reflects extra variability when, as a result of random pairings, large differences between pairs of sample means occur, just by chance, because they happen to deviate in opposite directions.

You might find it helpful to view the **standard error of the difference between means,** $\sigma_{\bar{X}_1-\bar{X}_2}$, as a rough measure of the average amount by which any sample mean difference deviates from the difference between population means. Viewed in this fashion, if the observed difference between sample means is smaller than the standard error, it qualifies as a common outcome—well within the average expected by chance, and the null hypothesis, H_0, is retained. On the other hand, if the observed difference is sufficiently larger than the standard error, it qualifies as a rare outcome—well beyond the average expected by chance, and H_0 is rejected.

The size of the standard error for two samples, $\sigma_{\bar{X}_1-\bar{X}_2}$—much like that of the standard error for one sample described earlier—becomes smaller with increases in sample sizes. With larger sample sizes, the values of $\bar{X}_1 - \bar{X}_2$ tend to cluster closer to the difference between population means, $\mu_1 - \mu_2$, allowing more precise generalizations.

*test procedure – p

Most investigators adopt a less-structured approach to hypothesis testing than that described in this book. The null hypothesis is neither retained nor rejected, but viewed with *degrees of suspicion*, depending on the degree of rarity of the observed value of *t* or, more generally, the test result. Instead of subscribing to a single *predetermined* level of significance, the investigator waits until *after* the test result has been observed and then assigns a probability, known as a *p*-value, representing the degree of rarity attained by the test result.

The *p*-value for a test result represents the degree of rarity of that result, given that the null hypothesis is true. Smaller *p*-values tend to discredit the null hypothesis and to support the research hypothesis.

Strictly speaking, the *p*-value indicates the degree of rarity of the observed test result when combined with all potentially *more deviant* test results. In other words, the *p*-value represents the proportion of area, beyond the observed result, in the tail of the sampling distribution, as shown in Figure 14.3 by the shaded sectors for two different

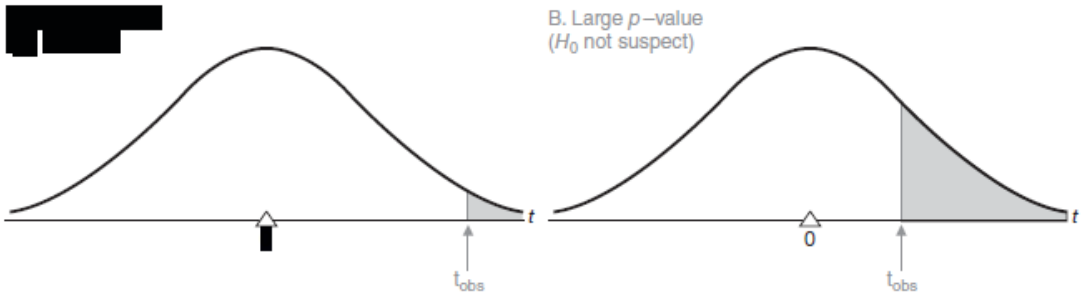
test results. In the left panel of Figure 14.3, a relatively deviant (from zero) observed *t* is associated with a small *p*-value that makes the null hypothesis suspect, while in the right panel, a relatively non-deviant observed *t* is associated with a large *p*-value that does not make the null hypothesis suspect.

Figure 14.3 illustrates *one-tailed p*-values that are appropriate whenever the investigator has an interest only in deviations in a particular direction, as with a one-tailed hypothesis test. Otherwise, *two-tailed p*-values are appropriate. Although not shown in Figure 14.3, two-tailed *p*-values would require equivalent shaded areas to be located in *both* tails of the sampling distribution, and the resulting two-tailed *p*-value would be twice as large as its corresponding one-tailed *p*-value.

Finding Approximate *p*-Values

Table B in Appendix C can be used to find *approximate p*-values, that is, *p*-values involving an inequality, such as $p < .05$ or $p > .05$. To aid in the identification of these approximate *p*-values, a shaded outline has been superimposed over the entries for *t* in Table B. Once you've located the observed *t* relative to the tabular entries, simply follow the vertical line upward to identify the correct approximate *p*-value.

To find the approximate *p*-value for the *t* of 2.16 for the EPO experiment, first identify the row in Table B for a one-tailed test with 10 degrees of freedom. The three entries in this row, 1.812, 2.764, and 4.144, serve as benchmarks for degrees of rarity corresponding to *p*-values of .05, .01, and .001, respectively. Since the observed *t* of 2.16 is located between the first entry of 1.812 and the second entry of 2.764, follow the vertical line between the two entries upward to $p < .05$. From most perspectives, this is a small *p*-value: The test result is rare—it could have occurred just by chance with a probability less than .05, given that H_0 is true. Therefore, support has been mustered for the research hypothesis. This conclusion is consistent with the decision to reject H_0 when a more structured hypothesis test at the .05 level of significance was conducted for the same data.



Shaded sectors showing small and large p*-values.*

Reading *p*-Values Reported by Others

A single research report might describe a batch of tests with a variety of approximate *p*-values, such as $p < .05$, $p < .01$, and $p < .001$, and, if the test failed to support the research hypothesis, $p > .05$. You must attend carefully to the direction of the inequality symbol. For example, the test result supports the research hypothesis when $p < .05$ but not when $p > .05$.

As illustrated in many of the computer outputs in this book, when statistical tests are performed by computers, with their capacity to obtain *exact p*-values (or values of *Sig.* in the case of SPSS), reports contain many different *p*-values, such as $p = .03$, $p = .27$, and $p = .009$. Even though more precise equalities replace inequalities, exact *p*-values listed on computer printouts are interpreted the same way as approximate *p*-values read from tables. For example, it's still true that $p = .03$ describes a rare test result, while $p = .27$ describes a result that is not particularly rare. Sometimes you'll see even a very rare $p = .000$, which, however, does not signify that *p* actually equals zero—an impossibility, since the *t* sampling distribution extends outward to infinity—but merely that rounding off causes the disappearance of non-zero digits from the reported *p*-value.

Evaluation of the *p*-Value Approach

This less-structured approach does have merit. Having eliminated the requirement that the null hypothesis be either retained or rejected, you can postpone a decision until sufficient evidence has been mustered, possibly from a series of investigations. This perspective is very attractive when test results are borderline. For instance, imagine

a hypothesis test in which the null hypothesis is retained, even though an observed t of 1.70 is only slightly less deviant than the critical t of 1.812 for the .05 level of significance. Given the less-structured approach, an investigator might, with the aid of a computer, establish that $p = .06$ for the observed t. Reporting the borderline result, with $p = .06$, implies at least some support for the research hypothesis.

One weakness of this less-structured approach is that, in the absence of a firm commitment to either retain or reject the null hypothesis according to some predetermined level of significance, it's difficult to deal with the important notions of type I and type II errors. For this reason, a more-structured approach to hypothesis testing will continue to be featured in this book, although not to the exclusion of the important approach involving p-values.

Level of Significance or p-Value?

A final word of caution. Do not confuse the level of significance with a p-value, even though both originate from the same column headings of Table B in Appendix C. Specified *before* the test result has been observed, the level of significance describes a degree of rarity that, if attained subsequently by the test result, triggers the decision to reject H_0. Specified *after* the test result has been observed, a p-value describes the most impressive degree of rarity actually attained by the test result.

You need not drop a personal preference for a more structured hypothesis test, with a predetermined level of significance, just because a research report contains only p-values. For instance, any p-value less than .05, such as $p < .05$, $p = .03$, $p < .01$, or $p < .001$, implies that, with the same data, H_0 would have been rejected at the .05 level of significance. By the same token, any p-value greater than .05, such as $p > .05$, $p < .10$, $p < .20$, or $p = .18$ implies that, with the same data, H_0 would have been retained at the .05 level of significance.

*statistical significance

14.7 STATISTICALLY SIGNIFICANT RESULTS

It's important that you accurately interpret the findings of others—often reported as "having statistical significance." Tests of hypotheses often are referred to as *tests of significance*, and test results are described as being *statistically significant* (if the null hypothesis has been rejected) or as not being statistically significant (if the null hypothesis has been retained). *Rejecting the null hypothesis* and *statistically significant* both signify that the test result can't be attributed to chance. However, correct usage dictates that *rejecting the null hypothesis* always refers to the population, such as rejecting the hypothesized zero difference between two population means, while *statistically significant* always refers to the sample, such as assigning statistical significance to the observed difference between two sample means. Either phrase can be used. However, assigning *statistical significance* to a population mean difference would be misleading, since a population mean difference equals a fixed value controlled by "nature," not something controlled by the results of a statistical test. *Rejecting* a sample mean difference also would be misleading, since a sample mean difference is an observed result that serves as the basis for statistical tests, not something to be rejected.

Statistical significance doesn't imply that the underlying effect is important. Statistical significance between pairs of sample means *implies only that the null hypothesis is probably false, and not whether it's false because of a large or small difference between population means.*

Beware of Excessively Large Sample Sizes

Using excessively large sample sizes can produce statistically significant results that lack importance. For instance, assume a new EPO experiment with the same amount of variability among endurance scores as in the original experiment, that is, with a pooled variance, s_p^2, equal to 16.2 (from Table 14.1). But assume that the new experiment has a *much smaller* mean difference, $X_1 X_2$, equal to only 0.50 minutes

(instead of 5 minutes in the original experiment) and much *larger* sample sizes each equal to 500 patients (instead of 6). Because of these much larger sample sizes, the new standard error would equal only 0.25 (instead of 2.32) and the new *t* would equal 2.00. Now we would have rejected the null hypothesis at the .05 level, even though the new difference between sample means is only one-tenth the size of the original difference. With large sample sizes and, therefore, with a small standard error, even a very small and unimportant *effect* (*difference between population means*) will be detected, and the test will be reported as statistically significant.

Statistical significance merely indicates that an observed effect, such as an observed difference between the sample means, is sufficiently large, relative to the standard error, to be viewed as a rare outcome. (Statistical significance also implies that the observed outcome is *reliable,* that is, it would reappear as a similarly rare outcome in a repeat experiment.) It's very desirable, therefore, that we go beyond reports of statistical significance by estimating the size of the effect and, if possible, judging its importance.

Avoid an Erroneous Conditional Probability

Rejecting H_0 at, for instance, the .05 level of significance, signifies that the probability of the observed, or a more extreme, result is less than or equal to .05 *assuming H_0 is true*. This is a conditional probability that takes the form:

Pr (the observed result, given H_0 is true) .05.

The probability of .05 depends entirely on the *assumption* that H_0 is true since that probability of .05 originates from the hypothesized sampling distribution centered about H_0.

This statement often is confused with another enticing but erroneous statement, namely H_0 itself is true with probability .05 or less, that reverses the order of events in the conditional probability. The new, erroneous conditional probability takes the form:

Pr (H_0 is true, given the observed result) .05.

At issue is the question of what the probability of .05 refers to. Our hypothesis testing procedure only supports the first, not the second conditional probability. Having rejected H_0 at the .05 level of significance, we can conclude, without indicating a specific probability, that H_0 is *probably false,* but we can't reverse the original conditional probability and conclude that it's true with only probability .05 or less. We have not tested the truth of H_0 on the basis of the observed result. To do so goes beyond the scope of our statistical test and makes an unwarranted claim regarding the probability that the null hypothesis actually is true.

*estimating effect size:

ESTIMATING EFFECT SIZE: POINT ESTIMATES

AND CONFIDENCE INTERVALS

It would make sense to estimate the effect for the EPO experiment featured in this chapter since the results are statistically significant. (But, strictly speaking, *only* if the results are statistically significant. Otherwise, we would be estimating an "effect" that could be merely transitory and attributed to chance.)

Po i n t E s t ima t e (X_1 X_2)

As you probably recall from Chapter 12, a point estimate is the most straightforward type of estimate. It identifies the observed difference for X_1 X_2, in this case,
5 minutes, as an estimate of the unknown effect, that is, the unknown difference between population means, $\mu_1 - \mu_2$. On average, the treatment patients stay on the treadmill for 11 minutes, which is almost twice as long as the 6 minutes for the control patients. If you think about it, this impressive estimate of effect size isn't surprising. With the very small groups of only 6 patients, we had to create a large, fictitious mean difference of 5 minutes in order to claim a statistically significant result. If this result had occurred in a real experiment, it would have signified a powerful effect of EPO on endurance that could be detected even with very small samples.

Confidence Interval

Although simple, straightforward, and precise, point estimates tend to be inaccurate because they ignore sampling variability. Confidence intervals do not because, as noted in Chapter 12, they are based on the variability in the sampling distribution of $X_1 X_2$.

To estimate the range of possible effects of EPO on endurance, a confidence interval can be constructed for the difference between population means, $\mu_1 - \mu_2$.

Confidence intervals for $\mu_1 - \mu_2$ specify ranges of values that, in the long run, include the unknown effect (difference between population means) a certain percent of the time.

Given two independent samples, a confidence interval for $\mu_1 - \mu_2$ can be constructed from the following expression:

CONFIDENCE INTERVAL (CI) FOR $\mu_1 - \mu_2$ (TWO INDEPENDENT SAMPLES)

$$\bar{X}_1 - \bar{X}_2 \pm \left(t_{conf}\right)\left(s_{\bar{X}_1 - \bar{X}_2}\right) \quad ($$

where $X_1 X_2$ represents the difference between sample means; t_{conf} represents a number, distributed with $n_1 + n_2 - 2$ degrees of freedom, from the *t* tables, which satisfies the confidence specifications; and

$X_1 X_2 s$ represents the estimated standard error defined in Formula 14.3.

To find the appropriate value of t_{conf} in Formula 14.4, refer to Table B in Appendix C and follow essentially the same procedure described earlier. For example, if a 95 percent confidence interval is desired for the EPO experiment, first locate the row corresponding to 10 degrees of freedom (from $df = n_1 + n_2 - 2 = 6 + 6 - 2 = 10$) and then locate the column for the 95 percent level of confidence, that is, the column heading identified with a single asterisk. The intersected cell specifies a value of 2.228 to be entered for t_{conf} in Formula 14.4. Given this value for t_{conf}, and values of 5 for the difference between sample means, $X_1 X_2$, and of 2.32 for the estimated standard error, $X_1 X_2 s$ (from Table 14.1), Formula 14.4 becomes

$$5 \pm (2.228)(2.32) = 5 \pm 5.17 = \begin{cases} 10.17 \\ -0.17 \end{cases}$$

Now it can be claimed, with 95 percent confidence, that the interval between –0.17 minutes and 10.17 minutes includes the true effect size, that is, the true difference between population means for endurance scores.

Interpreting Confidence Intervals for $\mu_1 - \mu_2$

The numbers in this confidence interval refer to *differences* between population means, and the signs are particularly important since they indicate the *direction* of these differences. Otherwise, the interpretation of a confidence interval for $\mu_1 - \mu_2$ is the same as that for μ. In the long run, 95 percent of all confidence intervals, similar to the one just stated, will include the unknown difference between population means. Although we never really know whether this particular confidence interval is true or false, we can be *reasonably confident* that the true effect (or true difference between population means) is neither less than –0.17 minutes nor more than 10.17 minutes If only positive differences had appeared in this confidence interval, a single interpretation would have been possible. However, the appearance of a negative difference in the lower limit indicates that EPO might hinder endurance, and therefore, no single interpretation is possible. Furthermore, the automatic inclusion of a zero difference in an interval with dissimilar signs indicates that EPO may have had no effect whatsoever on endurance.*

The range of possible differences (from a low of –0.17 minute to a high of 10.17 minutes) is very large and imprecise—as you would expect, given the very small sample sizes and, therefore, the relatively large standard error. A repeat experiment should use larger sample sizes in order to produce a narrower, more precise confidence interval that would reduce the range of possible population mean differences and effect sizes.

ESTIMATING EFFECT SIZE: COHEN'S *d*

Using a variation of the *z* score formula in Chapter 5, Cohen's *d* describes effect size by expressing *the observed mean difference in standard deviation units*. To calculate *d*, divide the observed mean difference by the standard deviation, that is,

STANDARDIZED EFFECT SIZE, COHEN'S *d* (TWO INDEPENDENT SAMPLES)

$$d = \frac{\text{mean difference}}{\text{standard deviation}} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2}} \quad (1$$

where, according to current usage, *d* refers to a standardized *estimate* of the effect size; X_1 and X_2 are the two sample means; and 2

s_p is the sample standard deviation

obtained from the square root of the pooled variance estimate.

Division of the mean difference by the standard deviation has several desirable consequences:

- The standard deviation supplies a *stable* frame of reference not influenced by increases in sample size. Unlike the standard error, whose value decreases as sample size increases, the value of the standard deviation remains the same, except for chance, as sample size increases. Therefore, straightforward comparisons can be made between *d* values based on studies with appreciably different sample sizes.
- The original units of measurement cancel out because of their appearance in both the numerator and denominator. Subsequently, *d* always emerges as an estimate in standard deviation units, regardless of whether the original mean difference is based on, for example, reaction times in *milliseconds* of pilots to two different cockpit alarms or weight losses in *pounds* of overweight subjects to two types of dietary restrictions. Except for chance, comparisons are straightforward between values of *d*—with larger values of *d* reflecting larger effect sizes—even though the original mean differences are based on very different units of measurement, such as milliseconds and pounds.

Cohen's Guidelines for *d*

After surveying the research literature, Jacob Cohen suggested a number of general

guidelines (see Table 14.2) for interpreting values of *d*:

- Effect size is *small if d* is less than or in the vicinity of 0.20, that is, one-fifth of a standard deviation.
- Effect size is *medium* if *d* is in the vicinity of 0.50, that is, one-half of a standard deviation.
- Effect size is *large* if *d* is more than or in the vicinity of 0.80, that is, four-fifths of a standard deviation.*

Although widely adopted, Cohen's abstract guidelines for small, medium, and large effects can be difficult to interpret. You might find these guidelines more comprehensible by referring to Table 14.3, where Cohen's guidelines for *d* are converted into more

Table 14.3
COHEN'S GUIDELINES FOR *d* AND MEAN DIFFERENCES FOR GPA, IQ, AND SAT SCORES

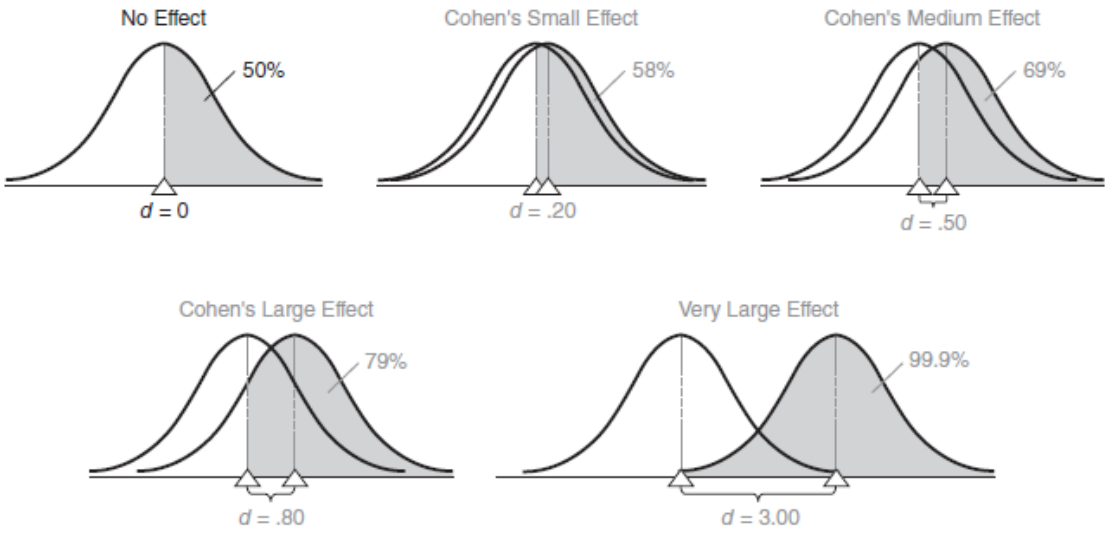
		MEAN DIFFERENCE		
		GPA	IQ	SAT
d	EFFECT SIZE	$s_p = 0.50$	$s_p = 15$	$s_p = 100$
.20	Small	0.10	3	20
.50	Medium	0.25	7.5	50
.80	Large	0.40	12	80

concrete mean differences involving GPAs, IQs, and SAT scores. Notice that Cohen's medium effect, a *d* value of .50, translates into mean differences of .25 for GPAs, 7.5 for IQs, and 50 for SAT scores. To qualify as medium effects, the average GPA would have to increase, for example, from 3.00 to 3.25; the average IQ from 100 to 107.5; and the average SAT score from 500 to 550.

Furthermore, for a particular measure, such as SAT scores, a 20-point mean difference corresponds to Cohen's small effect, while an 80-point mean difference corresponds to his large effect. However, *do not interpret Cohen's guidelines without regard to special circumstances.* A "small" 20-point increase in SAT scores might be viewed as virtually worthless if it occurred after a lengthy series of workshops on

taking SAT tests, but viewed as worthwhile if it occurred after a brief study session.

You might also find it helpful to visualize the impact of each of Cohen's guidelines on the degree of separation between pairs of normal curves. Although, of course, not every distribution is normal, these curves serve as a convenient frame of reference to render values of *d* more meaningful. As shown in Figure 14.4, separation between pairs of normal curves is nonexistent (and overlap is complete) when $d = 0$. Separation becomes progressively more conspicuous as the values of *d*, corresponding to Cohen's



small, medium, and large effects, increase from .20 to .50 and then to .80. Separation becomes very conspicuous, with relatively little overlap, given a *d* value of 3.00, equivalent to three standard deviations, for a very large effect.

To dramatize further the differences between selected *d* values, the percents (and shaded sectors) in Figure 14.4 reflect scores in the higher curve that exceed the mean of the lower curve. When $d = 0$, the two curves coincide, and it's a tossup, 50%, whether or not the scores in one curve exceed the mean of the other curve. As values of *d* increase, the percent of scores in the higher curve that exceed the mean of the lower curve varies from a modest 58% (six out of ten) when $d = .20$ to a more impressive 79% (eight out of ten) when $d = .80$ to a most impressive 99.9% (ten out of ten) when $d = 3.00$.

We can use *d* to estimate the standardized effect size for the statistically significant

results in the EPO experiment described in this chapter. When the mean difference of 5 is divided by the standard deviation of 4.02 (from the square root of the pooled variance estimate of 16.2 in Table 14.1), the value of *d* equals a large 1.24, that is, a mean difference equivalent to one and one-quarter standard deviations. (Being itself a product of chance sampling variability, this value of *d*—even if based on real data—would be highly speculative because of the instability of *d* when sample sizes are small.)

*meta analysis t-test for two related samples

The most recent *Publication Manual of the American Psychological Association* recommends that reports of statistical significance tests include some estimate of effect size. Beginning in the next section, we'll adopt this recommendation by including the standardized estimate of effect size, *d*, in reports of statistically significant mean differences. (A slightly more complicated estimate will be used in later chapters when effect size can't be conceptualized as a simple mean difference.) The routine reporting of effect sizes will greatly facilitate efforts to summarize research findings.

Because of the inevitable variability, attributable to differences in design, subject populations, measurements, etc., as well as chance, the size of effects differs among similar studies. Traditional literature reviews attempt to make sense out of these differences on the basis of expert judgment. Within the last couple of decades, literature reviews have been supplemented by more systematic reviews, referred to as "metaanalysis." A meta-analysis begins with an intensive review of all relevant studies. This includes small and even unpublished studies, to try to limit potential "publication bias" arising from only reporting statistically significant results. Typically, extensive details are recorded for each study, such as estimates of effect, design (for example, experimental versus observational), subject population, variability, sample size, etc. Then the collection of previous findings are combined using statistical procedures to obtain either a composite estimate (for example, a standardized mean difference, such as Cohen's *d*) of the overall effect and its confidence interval, or estimates of subsets of

similar effects, if required by the excessive variability among the original effects.*

UNIT -4

ANALYSIS OF VARIANCE

F-test – ANOVA – estimating effect size – multiple comparisons – case studies Analysis of variance with repeated measures Two-factor experiments – three f-tests – two-factor ANOVA – other types of ANOVA Introduction to chi-square tests

F TEST

In previous chapters, the null hypothesis has been tested with a t ratio. In the two-sample case, t reflects the ratio between the observed difference between the two sample means in the numerator and the estimated standard error in the denominator. For three or more samples, the null hypothesis is tested with a new ratio, the F ratio. Essentially, F reflects the ratio of the observed differences between all sample means (measured as variability

between groups) in the numerator and the estimated error term or pooled variance estimate (measured as variability within groups) in the denominator term, that is,

F RATIO

$$F = \frac{variability\ between\ groups}{variability\ within\ groups} \qquad (16.1)$$

Like t, F has its own family of sampling distributions that can be consulted, as described in Section 16.6, to test the null hypothesis. The resulting test is known as an F test.

An F test of the null hypothesis is based on the notion that if the null hypothesis is true, both the numerator and the denominator of the F ratio would tend to be about the same, but if the null hypothesis is false, the numerator would tend to be larger than the denominator.

If Null Hypothesis Is True

If the null hypothesis is true (because there is no treatment effect due to different

sleep deprivation periods), the two estimates of variability (between and within groups) would reflect only random error. In this case,

$$F = \frac{random\ error}{random\ error}$$

Except for chance, estimates in both the numerator and the denominator are similar, and generally, F varies about a value of 1.

If Null Hypothesis Is False

If the null hypothesis is false (because there is a treatment effect due to different

sleep deprivation periods), both estimates still would reflect random error, but the estimate for between groups would also reflect the treatment effect. In this case,

$$F = \frac{random\ error + treatment\ effect}{random\ error}$$

When the null hypothesis is false, the presence of a treatment effect tends to cause a chain reaction: The observed differences between group means tend to be large, as does the variability between groups. Accordingly, the numerator term tends to exceed the denominator term, producing an F whose value is larger than 1. When the null

hypothesis is false because of a large treatment effect, there is an even more pronounced chain reaction, beginning with very large observed differences between group

means and ending with an F whose value tends to be considerably larger than 1.

When Status of Null Hypothesis Is Unknown

In practice, of course, we never really know whether the null hypothesis is true or false. Following the usual procedure, we assume the null hypothesis to be true and view the observed F within the context of its hypothesized sampling distribution, as

shown in Figure 16.1. If, because the differences between group means are relatively

small, the observed F appears to emerge from the dense concentration of possible F

ratios smaller than the critical F, the experimental outcome would be viewed as a common occurrence. Therefore, the null hypothesis would be retained. On the other hand,

if, because the differences between group means are relatively large, the observed F

appears to emerge from the sparse concentration of possible F ratios equal to or greater

than the critical F, the experimental outcome would be viewed as a rare occurrence,

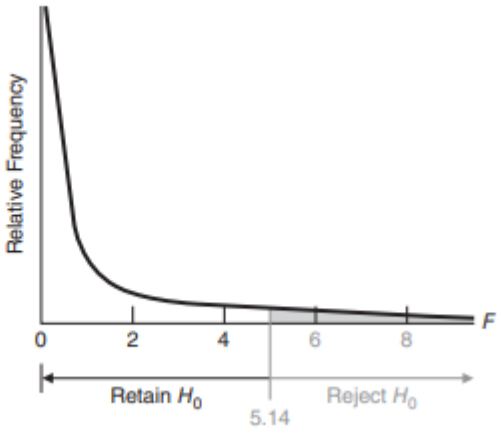
and the null hypothesis would be rejected. In the latter case, the value of the observed

F is presumed to be inflated by a treatment effect.

Test Results for Outcomes A and B

Full-fledged F tests for Outcomes A and B agree with the earlier intuitive decisions.

Given the .05 level of significance, the null hypothesis should be retained for Outcome



FIGURE

Hypothesized sampling distribution of F (for 2 and 6 degrees of freedom)

A, since the observed F of 0.75 is smaller than the critical F of 5.14. However, the null hypothesis should be rejected for Outcome B, since the observed F of 7.36 exceeds the critical F. The hypothesis test for Outcome B, as summarized in the accompanying box, will be discussed later in more detail.

HYPOTHESIS TEST SUMMARY

One-Factor *F* Test (Sleep Deprivation Experiment, Outcome B)

Research Problem
On average, are subjects' aggression scores in a controlled social situation affected by sleep deprivation periods of 0, 24, or 48 hours?

Statistical Hypotheses

H_0: $\mu_0 = \mu_{24} = \mu_{48}$
H_1: H_0 is false.

Decision Rule
Reject H_0 at the .05 level of significance if $F \geq 5.14$ (from Table C, Appendix C, given $df_{between} = 2$ and $df_{within} = 6$).

Calculations

$F = 7.36$ (See Tables 16.3 and 16.6 for additional details.)

Decision
Reject H_0 at the .05 level of significance because $F = 7.36$ exceeds 5.14.

Interpretation
Hours of sleep deprivation affect the subjects' mean aggression scores in a controlled social situation.

ANOVA

Traditionally, both in statistics textbooks and the literature, ANOVA results have been summarized as shown in Table 16.6. "Source" refers to the source of variability, that is, between groups, within groups, and total. Notice the arrangement of column headings from SS and df to MS and F. Also, notice that the bottom row for total variability contains entries only for SS and df. Ordinarily, the shaded numbers in parentheses don't appear in ANOVA tables, but in Table 16.6 they show the origin of each MS and of F. The asterisk in Table 16.6 emphasizes that the observed F of 7.36 exceeds the critical F of 5.14 and therefore causes the null hypothesis to be rejected at the .05 level of significance. Other Labels Some ANOVA summary tables use labels other than those shown in Table 16.6. For instance, "Between" might be replaced with "Treatment," since the variability between groups reflects any treatment effect. Or "Between" might be replaced by a description of the actual experimental treatment, such as "Hours of Sleep Deprivation" or "Sleep Deprivation." Likewise, "Within" might be replaced with "Error," since variability within groups reflects only the presence of random error

Table 16.6
ANOVA TABLE (SLEEP DEPRIVATION EXPERIMENT)

SOURCE	*SS*	*df*	*MS*	*F*
Between	54	2	$\left(\frac{54}{2}=\right)$ 27	$\left(\frac{27}{3.67}=\right)$ 7.36*
Within	22	6	$\left(\frac{22}{6}=\right)$ 3.67	
Total	76	8		

** Significant at the .05 level.*

ESTIMATING EFFECT SIZE

a standardized estimate of effect size, which has the desirable property of being independent of sample sizes. Like the t test, a statistically significant F indicates merely that the null hypothesis is probably false; otherwise, it fails to provide an accurate estimate of effect size. A new estimate of effect size must both reflect the overall effect associated with the null hypothesis in ANOVA and be independent of sample sizes.* A most straightforward estimate, denoted as η2 , capitalizes on existing information in the ANOVA summary table by specifying that SSbetween be divided by SStotal, that is,

PROPORTION OF EXPLAINED VARIANCE, η^2 (ONE-FACTOR ANOVA)

$$\eta^2 = \frac{SS_{between}}{SS_{total}} \qquad (16.8)$$

where η2 represents the proportion of explained variance and SSbetween and SStotal represent the between group and total sum of squares, respectively. This ratio estimates not population mean differences, but the proportion (from 0 to 1) of the total variance for all scores, as reflected in SStotal, that can be explained by or attributed to the variance of treatment groups, as reflected in SSbetween. Speaking very generally, η2 indicates the proportion of differences among all scores attributable to differences among treatment groups. The larger this proportion, the larger the estimated size of the overall effect of the treatment on the dependent variable. The Greek symbol η2 , pronounced eta-squared, is often referred to as the squared curvilinear correlation coefficient. This terminology reinforces the notion that η2 is just a square root away from a number describing the nonlinear correlation between values of the independent and dependent variables. Notice also that this interpretation of η2 is very similar to that for r2 , the squared linear correlation coefficient described in Chapter 7. Refer to Figure 16.3 to gain some appreciation of how the squared curvilinear correlation, η2 , reflects the proportion of variance explained by the independent variable. This figure shows values of η2 for three different outcomes, reflecting no effect, a maximum effect, and a partial effect, for the sleep deprivation experiment. Panel I There is no apparent visual separation between the scores for each of the three groups, since each group mean, Xgroup, equals the grand mean, Xgrand. Therefore, there is no variability between groups, SS between = 0, and

$$\eta^2 = \frac{SS_{between}}{SS_{total}} = \frac{SS_{between}}{SS_{between} + SS_{within}} = \frac{0}{0 + SS_{within}} = \frac{0}{SS_{within}} = 0$$

The value of 0 for η2 implies that none of the variance among scores can be attributed to variance between treatment groups. The treatment variable has no effect whatsoever on the dependent variable

*Independence of sample size is an important property of estimates of effect size. Essentially, large sample sizes in ANOVA automatically inflates the numerator term, MSbetween, relative to the denominator term, MSwithin, of the F test. For instance, the F of 0.75 for Outcome A in Table 16.1, was not significant at the .05 level. If, however, the sample size for each of the three groups were increased from n = 3 to n = 30, the new F of 7.50 would have been significant at the .01 level even though the differences between group means remain the same.

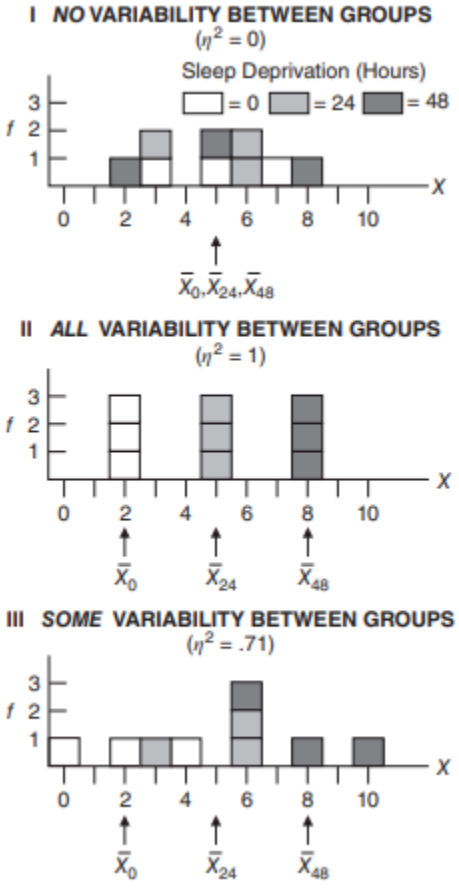


FIGURE 16.3 Values of η2 for three possible outcomes of the sleep deprivation experiment.

Panel II

There is complete visual separation between the scores for each of the three groups, since each score, X, coincides with its own distinctive group mean, . Xgroup Therefore, there is no variability within groups, SSwithin = 0, and 2 1

$$\eta^2 = \frac{SS_{between}}{SS_{total}} = \frac{SS_{between}}{SS_{between} + SS_{within}} = \frac{SS_{between}}{SS_{between} + 0} = \frac{SS_{between}}{SS_{between}} = 1$$

The value of 1 for η2 implies that all of the variance among scores can be attributed to the variance between treatment groups. The treatment variable has a maximum (perfect) effect on the dependent variable.

Panel III

In spite of some overlap, there is an apparent visual separation between scores for the three groups. Now there is variability both within and between groups, as for Outcome B of the sleep deprivation experiment (Table 16.6), SSbetween = 54 and SSwithin = 22, and

$$\eta^2 = \frac{SS_{between}}{SS_{total}} = \frac{SS_{between}}{SS_{between} + SS_{within}} = \frac{54}{54 + 22} = \frac{54}{76} = .71$$

1 6 . 1 0 M U LT I P L E C O M PA R I S O N S

Rejection of the overall null hypothesis indicates only that all population means are not equal. In the case of the original sleep deprivation experiment, the rejection of H0 signals the presence of one or more inequalities between the mean aggression scores for populations of subjects exposed to 0, 24, or 48 hours of sleep deprivation, that is, between μ0, μ24, and μ48. To pinpoint the one or more differences between pairs of population means that contribute to the rejection of the overall H0, we must use a test of multiple comparisons. A test of multiple comparisons is designed to evaluate not just one but a series of differences between population means, such as those for each of the three possible differences between pairs of population means for the present experiment, namely, μ0 − μ24, μ0 − μ48, and μ24 − μ48.

t Test Not Appropriate

These differences can’t be evaluated with a series of regular t tests, except under special circumstances alluded to later in this section. The regular t test is designed to evaluate a single comparison for a pair of observed means, not multiple comparisons for all possible pairs of observed means. Among other complications, the use of mulTable 16.7 GUIDELINES FOR η2 η2 EFFECT .01 Small .09 Medium .25 Large *Cohen, J. (1988). Statistical Power Analysis in the Behavioral Sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

tiple t tests increases the probability of a type I error (rejecting a true null hypothesis) beyond the value specified by the level of significance. Coin-Tossing Example A coin-tossing example might clarify this problem. When a fair coin is tossed only once, the probability of heads equals .50—just as, when a single t test is to be conducted at the .05 level of significance, the probability of a type I error equals .05. When a

fair coin is tossed three times, however, heads can appear not only on the first toss but also on the second or third toss, and hence the probability of heads on at least one of the three tosses exceeds .50. By the same token, for a series of three t tests, each conducted at the .05 level of significance, a type I error can be committed not only on the first test but also on the second or third test, and hence the probability of committing a type I error on at least one of the three tests exceeds .05. In fact, the cumulative probability of at least one type I error can be as large as .15 for a series of three t tests and even larger for a more extended series of t tests. Tukey's HSD Test The above shortcoming does not apply to a number of specially designed multiple comparison tests, including Tukey's HSD or "honestly significant difference" test. Once the overall null hypothesis has been rejected in ANOVA, Tukey's HSD test can be used to test all possible differences between pairs of means, and yet the cumulative probability of a type I error never exceeds the specified level of significance. Finding the Critical Value Given a significant F for the overall null hypothesis, as in the sleep deprivation experiment, Tukey's test supplies a single critical value, HSD, for evaluating the significance of each difference for every possible pair of means, that is, 0 24 0 48 XXXX , , and X X 24 48. Essentially, the critical value for HSD is adjusted upward for the number of group means, k, being compared to compensate for the increased cumulative probability of incurring at least one type I error. The net effect of this upward adjustment is to make it more difficult to reject the null hypothesis for any particular pair of population means—and to increase the likelihood of detecting only honestly significant (or real) differences. If the absolute difference between any pair of means equals or exceeds the critical value for HSD, the null hypothesis for that particular pair of population means can be rejected. To determine HSD, use the following expression:

TUKEY'S *HSD* TEST

$$HSD = q\sqrt{\frac{MS_{within}}{n}} \qquad (16.9)$$

where HSD is the positive critical value for any difference between two means; q is a value, technically referred to as the Studentized Range Statistic, obtained from Table G in Appendix C; MSwithin is the customary mean square for within-group variability for the overall ANOVA; and n is the sample size in each group.*

*Equation 16.9 assumes equal sample sizes. Otherwise, if sample sizes are not equal and you lack access to an automatically adjusting computer program, such as Minitab, SAS, or SPSS, replace n in Equation 16.9 with the mean of all sample sizes, n.

To obtain a value for q at the .05 level (light numbers) or the .01 level (dark numbers) in Table G, find the cell intersected by k, the number of groups, and dfwithin, the degrees of freedom for within-group (or error) variability in the original ANOVA. Given values of k = 3 and dfwithin = 6 for the sleep deprivation experiment, the intersected cell shows a value of 4.34 for q at the .05 level. Substituting q = 4.34, MSwithin = 3.67, and n = 3 in Equation 16.9, we can solve for HSD as follows:

$$HSD = q\sqrt{\frac{MS_{within}}{n}} = 4.34\sqrt{\frac{3.67}{3}} = 4.34\,(1.10) = 4.77$$

Interpretation for Sleep Deprivation Experiment Table 16.8 shows absolute differences of either 3, 6, or 3 for the three pairs of means in the current experiment. (This table serves as a good model for evaluating the significance of differences between all possible pairs of sample means.) Since only the difference of 6 for the comparison involving X 0 and X 48 exceeds the critical HSD value of 4.77, only the null hypothesis for μ0 − μ48 can be rejected at the .05 level. We can conclude that, when compared with 0 hours of sleep deprivation, 48 hours of sleep deprivation tends to produce, on average, more aggressive behavior in a controlled social situation. There is no evidence, however, that subjects deprived of sleep for 24 hours are either more aggressive than those deprived for 0 hours or less aggressive than those deprived for 48 hours. Estimating Effect Size The effect size for any significant difference between pairs of means can be estimated with Cohen's d, as adapted from Equation 14.5 on page 262 that is,

STANDARDIZED EFFECT SIZE, COHEN'S *d* (ADAPTED FOR ANOVA)

$$d = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{MS_{within}}} \qquad (16.10)$$

where d is an estimate of the standardized effect size, X1and X2 are the pair of significantly different means, and MSwithin , the square root of the within-group mean square for the one-factor ANOVA, represents the sample standard deviation.

Table 16.8
ALL POSSIBLE ABSOLUTE DIFFERENCES BETWEEN PAIRS OF MEANS (FOR THE SLEEP DEPRIVATION EXPERIMENT)

	$\overline{X}_0 = 2$	$\overline{X}_{24} = 5$	$\overline{X}_{48} = 8$
$\overline{X}_0 = 2$	—	3	6*
$\overline{X}_{24} = 5$		—	3
$\overline{X}_{48} = 8$			—

**Significant at the .05 level.*

To estimate the standardized effect size for the one significant difference between means for 0 and 48 hours of sleep deprivation, enter X48 – X0 = 6 and MSwithin = 3.67 in Equation 16.10 and solve for d:

$$d\left(\overline{X}_{48}, \overline{X}_0\right) = \frac{6}{\sqrt{3.67}} = \frac{6}{19.2} = 3.13$$

which is a very large effect, equivalent to more than three standard deviations. (According to Cohen's guidelines for d, described on page 262, the effect size is large if d is more than 0.8.) This result isn't

surprising given the very large effect size of η2 = .71 for the proportion of explained variance attributable to the differences between all three groups in the sleep deprivation experiment

TWO FACTOR EXPERIMENT

Often referred to as the "bystander effect," do crowds affect our willingness to assume responsibility for the welfare of others and ourselves? For instance, does the presence of bystanders inhibit our reaction to potentially dangerous smoke seeping from a wall vent? Hoping to answer this question, a social psychologist measures any delay in a subject's alarm reaction (the dependent variable) as smoke fills a waiting room occupied only by the subject, plus "crowds" of either zero, two, or four associates of the experimenter—the first independent variable or factor—who act as regular subjects but, in fact, ignore the smoke. As a second independent variable or factor, the social psychologist randomly assigns subjects to one of two "degrees of danger," that is, the rate at which the smoke enters the room, either non dangerous (slow rate) or dangerous (rapid rate). Using this two factor ANOVA design, the psychologist can test not just two but three null hypotheses, namely, the effect on subjects' reaction times of (1) crowd size, (2) degree of danger and, as a bonus, (3) the combination or interaction of crowd size and degree of danger. For computational simplicity, assume that the social psychologist randomly assigns two subjects to be tested (one at a time) with crowds of either zero, two, or four people and either the non-dangerous or dangerous conditions. The resulting six groups, each consisting of two subjects, represent all possible combinations of the two factors.* Tables for Main Effects and Interaction Table 18.1 shows one set of possible outcomes for the two-factor study. Although, as indicated in Chapter 16, the actual computations in ANOVA usually are based on totals, preliminary interpretations can be based on either totals or means. In Table 18.1, the shaded numbers represent four different types of means: 1. The three column means (9, 12, 15) represent the mean reaction times for each crowd size when degree of danger is ignored. Any differences among these column means not attributable to chance are referred to as the main effect of crowd size on reaction time. In ANOVA, main effect always refers to the effect of a single factor, such as crowd size, when any other factor, such as degree of danger, is ignored.

Table 18.1
OUTCOME OF TWO-FACTOR EXPERIMENT
(REACTION TIMES IN MINUTES)

	CROWD SIZE						
DEGREE OF DANGER	**ZERO**		**TWO**		**FOUR**		**ROW MEAN**
Dangerous	8	8	8	7	10	9	8
	8		6		8		
Nondangerous	9	10	15	17	24	21	16
	11		19		18		
Column mean		9		12		15	Grand mean = 12

Note: Shaded numbers are means.

*The current example simulates some of the main findings from an extensive meta- analytic review by Fischer, P., et al. (2011). The Bystander Effect: A Meta-Analysis Review on Bystander Intervention in Dangerous and Non-Dangerous Emergencies. Psychological Bulletin, 137, 517–537.

18.1 A TWO-FACTOR EXPERIMENT:

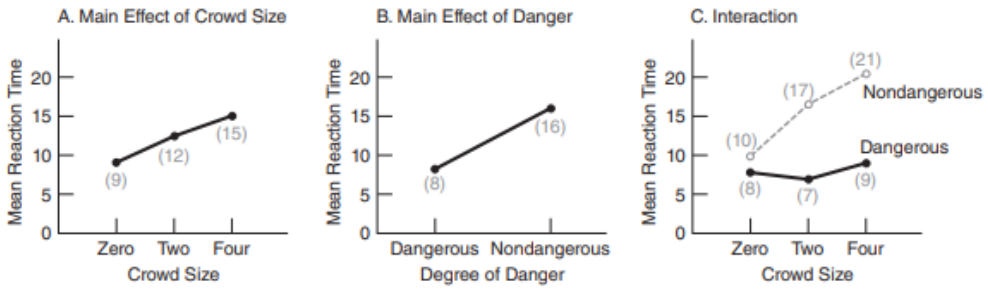
RESPONSIBILITY IN CROWDS 341

2. The two row means (8, 16) represent the mean reaction times for degree of danger when crowd size is ignored. Any difference between these row means not attributable to chance is referred to as the main effect of degree of danger on reaction time.

3. The mean of the reaction times for each group of two subjects yields the six means (8, 7, 9, 10, 17, 21) for each combination of the two factors. Often referred to as cell means or treatment-combination means, these means reflect not only the main effects for crowd size and degree of danger described earlier but, more importantly, any effect due to the interaction between crowd size and degree of danger, as described below.

4. Finally, the one mean for all three column means—or for both row means— yields the overall or grand mean (12) for all subjects in the study.

Graphs for Main Effects To preview the experimental outcomes, let's look for obvious trends in a series of graphs based on Table 18.1. The slanted line in panel A of Figure 18.1 depicts the large differences between column means, that is, between mean reaction times for subjects, regardless of degree of danger, with crowds of zero, two, and four people. The relatively steep slant of this line suggests that the null hypothesis for crowd size might be rejected. The steeper the slant is, the larger the observed differences between column means and the greater the suspected main effect of crowd size. On the other hand, a fairly level line in panel A of Figure 18.1 would have reflected the relative absence of any main effect due to crowd size. The slanted line in panel B of Figure 18.1 depicts the large difference between row means, that is, between mean reaction times for dangerous and non-dangerous conditions, regardless of crowd size. The relatively steep slope of this line suggests that the null hypothesis for degree of danger also might be rejected; that is, there might be a main effect due to degree of danger. Graph for Interaction These preliminary conclusions about main effects must be qualified because of a complication due to the combined effect or interaction of crowd size and degree of danger on reaction time. Interaction occurs whenever the effects of one factor on the dependent variable are not consistent for all values (or levels) of the second factor.



ANALYSIS OF VARIANCE (TWO FACTORS) Panel C of Figure 18.1 depicts the interaction between crowd size and degree of danger. The two nonparallel lines in panel C depict differences between the

three cell means in the first row and the three cell means in the second row—that is, between the mean reaction times for the dangerous condition for different crowd sizes and the mean reaction times for the non-dangerous condition for different crowd sizes. Although the line for the dangerous conditions remains fairly level, that for the no dangerous conditions is slanted, suggesting that the reaction times for the non-dangerous conditions, but not those for the dangerous conditions, are influenced by crowd size. Because the effect of crowd size is not consistent for the no dangerous and dangerous conditions—portrayed by the apparent non parallelism between the two lines in panel C of Figure 18.1—the null hypothesis (that there is no interaction between the two factors) might be rejected. Section 18.3 contains additional comments about interaction, as well as a more preferred definition of interaction. Summary of Preliminary Interpretations To summarize, a non-statistical evaluation of the graphs of data for the two-factor experiment suggests a number of preliminary interpretations. Each of the three null hypotheses regarding the effects of crowd size, degree of danger, and the interaction of these factors might be rejected. Because of the suspected interaction, however, any generalizations about the main effects of one factor must be qualified in terms of specific levels of the second factor. Pending the outcome of the statistical analysis, you can speculate that the crowd size probably influences the reaction times for the nondangerous but not the dangerous conditions.

THREE F TESTS

As suggested in Figure 18.2, F ratios in both a one- and a two-factor ANOVA always consist of a numerator (shaded) that measures some aspect of variability between

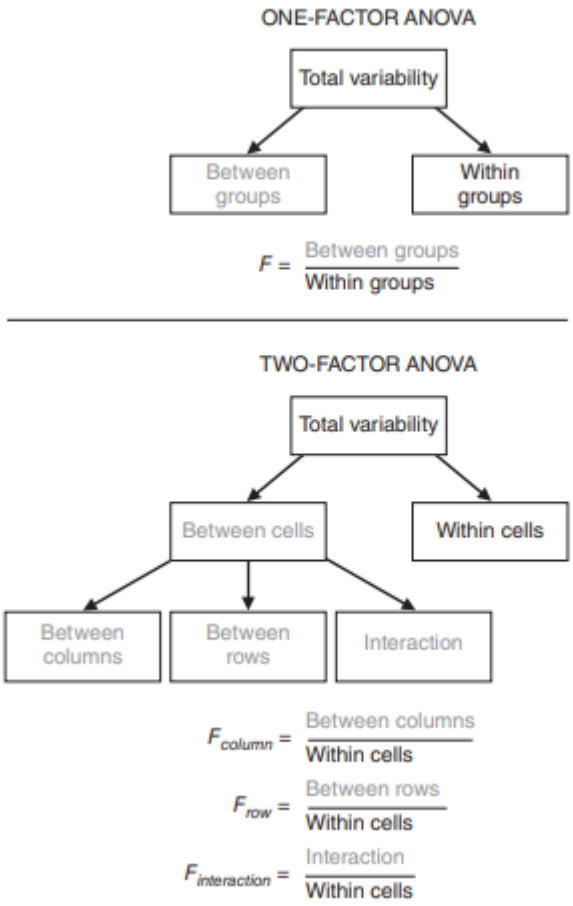


FIGURE 18.2 Sources of variability and F ratios in one- and two-factor ANOVAs.

Interaction Between rows Between columns F = Between groups Within groups F column = Between columns Within cells F row = Between rows Within cells F interaction = Interaction Within cells Between cells Within cells Total variability Between groups Within groups Total variability ONE-FACTOR ANOVA TWO-FACTOR ANOVA groups or cells and a denominator that measures variability within groups or cells. In a one-factor ANOVA, a single null hypothesis is tested with one F ratio. In two-factor ANOVA, three different null hypotheses are tested, one at a time, with three F ratios: F column, F row, and F interaction. The numerator of each of these three F ratios reflects a different aspect of variability between cells: variability between columns (crowd size), variability between rows (degree of danger), and interaction—any remaining variability between cells not attributable to either variability between columns (crowd size) or rows (degree of danger). The shaded numerator terms for the three F ratios in the bottom panel of Figure 18.2 estimate random error and, if present, a treatment effect (for subjects treated differently by the investigator). The denominator term always estimates only random error (for subjects treated similarly in the same cell). In practice, a sufficiently large F value is viewed as rare, given that the null hypothesis is true, and therefore, it leads to the rejection of the null hypothesis. Otherwise, the null hypothesis is retained

Two-way ANOVA

A **two-way ANOVA** ("analysis of variance") is used to determine whether or not there is a statistically significant difference between the means of three or more independent groups that have been split on two variables (sometimes called "factors").

When to Use a Two-Way ANOVA

You should use a two-way ANOVA when you'd like to know how two factors affect a response variable and whether or not there is an interaction effect between the two factors on the response variable

Two-Way ANOVA Assumptions

For the results of a two-way ANOVA to be valid, the following assumptions should be met:

1. Normality – The response variable is approximately normally distributed for each group.

2. Equal Variances – The variances for each group should be roughly equal.

3. Independence – The observations in each group are independent of each other and the observations within groups were obtained by a random sample.

OTHER TYPES OF ANOVA

One- and two-factor studies do not exhaust the possibilities for ANOVA. For instance, you could use ANOVA to analyze the results of a three-factor study with three independent variables, three 2-way interactions, and one 3-way interaction. Furthermore, regardless of the number of factors, each subject might be measured repeatedly along all levels of one or more factors. Although the basic concepts described in this book transfer almost intact to a wide assortment of more intricate research designs, computational procedures grow more complex, and the interpretation of results often is more difficult. Intricate research designs, requiring the use of complex types of ANOVA, provide the skilled investigator with powerful tools for evaluating complicated situations. Under no circumstances, however, should a study be valued simply because of the complexity of its design and statistical analysis. Use the least complex design and analysis that will answer your research questions

Introduction To An Chi-Square Test

The Chi-Square test is a statistical procedure used by researchers to examine the differences between categorical variables in the same population.

For example, imagine that a research group is interested in whether or not education level and marital status are related for all people in the U.S.

After collecting a simple random sample of 500 U.S. citizens, and administering a survey to this sample, the researchers could first manually observe the frequency distribution of marital status and education category within their sample.

The researchers could then perform a Chi-Square test to validate or provide additional context for these observed frequencies.

$$\chi^2 = \Sigma \frac{(oij - eij)^2}{eij}$$

First, Chi-Square *only* tests whether two individual variables are independent in a binary, “yes” or “no” format.

Chi-Square testing does not provide any insight into the *degree* of difference between the respondent categories, meaning that researchers are not able to tell which statistic (result of the Chi-Square test) is greater or less than the other.

Second, Chi-Square requires researchers to use numerical values, also known as frequency counts, instead of using percentages or ratios. This can limit the flexibility that researchers have in terms of the processes that they use.

UNIT V

PREDICTIVE ANALYTICS

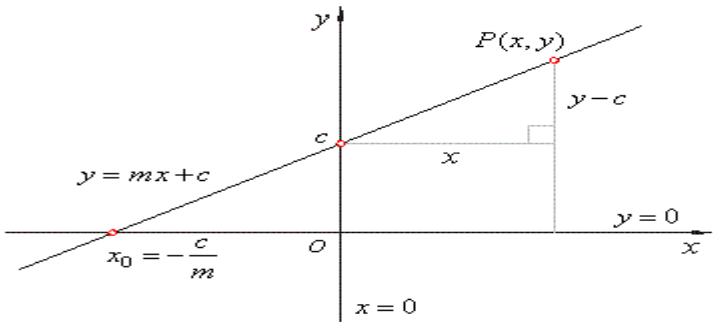
Linear least squares – implementation – goodness of fit – testing a linear model – weighted resampling Regression using StatsModels – multiple regression – nonlinear relationships – logistic regression – estimating parameters – accuracy Time series analysis – moving averages – missing values – serial correlation – autocorrelation Introduction to survival analysis

Linear least squares:

In statistics, linear regression is a linear approach to modelling the relationship between a dependent variable and one or more independent variables. In the case of one independent variable it is called simple linear regression. For more than one independent variable, the process is called mulitple linear regression. We will be dealing with simple linear regression in this tutorial.

Let **X** be the independent variable and **Y** be the dependent variable. We will define a linear relationship between these two variables as follows:

$$Y = mX + c$$



This is the equation for a line that you studied in high school. **m** is the slope of the line and **c** is the y intercept. Today we will use this equation to

train our model with a given dataset and predict the value of **Y** for any given value of X

So to minimize the error we need a way to calculate the error in the first place. A **loss function** in machine learning is simply a measure of how different the predicted value is from the actual value.
Today we will be using the **Quadratic Loss Function** to calculate the loss or error in our model. It can be defined as:

$$L(x) = \sum_{i=1}^{n} (y_i - p_i)^2$$

Least Squares method

Now that we have determined the loss function, the only thing left to do is minimize it. This is done by finding the partial derivative of **L**, equating it to 0 and then finding an expression for **m** and **c**. After we do the math, we are left with these equations:

$$m = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$c = \bar{y} - m\bar{x}$$

Here $\bar{x}$ is the mean of all the values in the input **X** and $\bar{y}$ is the mean of all the values in the desired output **Y**. This is the Least Squares method. Now we will implement this in python and make predictions.

***Implementation**

There are Seven stages of predictive analytics implementation

- 01 Project definition. It is essential to be specific about what you hope to achieve by implementing predictive analytics methodology. ...
- 02 Data collection.
- 03 Data analysis.
- 04 Statistics.
- 05 Modeling.
- 06 Deployment
- 07 Monitoring.

***GODNESS OF FIT**

Having fit a linear model to the data, we might want to know how good it is. Well, that depends on what it's for. One way to evaluate a model is its predictive power.

In the context of prediction, the quantity we are trying to guess is called a **dependent variable** and the quantity we are using to make the guess is called an **explanatory** or **independent variable**.

To measure the predictive power of a model, we can compute the **coefficient of determination**, more commonly known as "R-squared":

$$R^2 = 1 - \frac{Var(\varepsilon)}{Var(Y)}$$

To understand what R^2 means, suppose (again) that you are trying to guess someone's weight. If you didn't know anything about them, your best strategy would be to guess $\bar{y}$; in that case the MSE of your guesses would be Var(Y):

$$MSE = \frac{1}{n}\sum(\bar{y} - y_i)^2 = Var(Y)$$

But if I told you their height, you would guess $\hat{\alpha} + \hat{\beta}\, x_i$; in that case your MSE would be Var(ε).

$$MSE = \frac{1}{n}\sum(\hat{\alpha} + \hat{\beta}x_i - y_i)^2 = Var(\varepsilon)$$

So the term Var(ε)/Var(Y) is the ratio of mean squared error with and without the explanatory variable, which is the fraction of variability left unexplained by the model. The complement, R^2, is the fraction of variability explained by the model.

If a model yields $R^2 = 0.64$, you could say that the model explains 64% of the variability, or it might be more precise to say that it reduces the MSE of your predictions by 64%.

In the context of a linear least squares model, it turns out that there is a simple relationship between the coefficient of determination and Pearson's correlation coefficient, ρ:

$$R^2 = \rho^2$$

See `http://wikipedia.org/wiki/Howzzat!`

Exercise 9.8 The Wechsler Adult Intelligence Scale (WAIS) is meant to be a measure of intelligence; scores are calibrated so that the mean and standard deviation in the general population are 100 and 15.

Suppose that you wanted to predict someone's WAIS score based on their SAT scores. According to one study, there is a Pearson correlation of 0.72 between total SAT scores and WAIS scores.

If you applied your predictor to a large sample, what would you expect to be the mean squared error (MSE) of your predictions?

Hint: What is the MSE if you always guess 100?

Exercise 9.9 Write a function named `Residuals` that takes X, Y, $\hat{\alpha}$ and $\hat{\beta}$ and returns a list of ε_i.

Write a function named `CoefDetermination` that takes the ε_i and Y and returns R^2. To test your functions, confirm that $R^2 = \rho^2$. You can download a solution from `http://thinkstats.com/correlation.py`.

Exercise 9.10 Using the height and weight data from the BRFSS (one more time), compute $\hat{\alpha}$, $\hat{\beta}$ and R^2. If you were trying to guess someone's weight, how much would it help to know their height? You can download a solution

***MULTIPLE REGRESSION EQUATIONS**

Any serious predictive effort usually culminates in a more complex equation that contains not just one but several X, or predictor variables. For instance, a serious effort to predict college GPA might culminate in the following equation:

$$\dot{Y} = .410(X1) + .005(X2) + .001(X3) + 1.03$$

where Y′ represents predicted college GPA and X1 , X2 , and X3 refer to high school GPA, IQ score, and SAT score, respectively. By capitalizing on the combined predictive power of several predictor variables, these multiple regression equations supply more accurate predictions for Y′ (often referred to as the criterion variable) than could be obtained from a simple regression equation

COMMON FEAUTERS

Although more difficult to visualize, multiple regression equations possess many features in common with their simple counterparts. For instance, they still qualify as least squares equations, since they minimize the sum of the squared predictive errors. By the same token, they are accompanied by standard errors of estimate that roughly measure the average amounts of predictive error. Be assured, therefore, that this chapter will serve as a good point of departure if, sometime in the future, you must deal with multiple regression equations.

***WEIGHTED RESAMPLING**

For what follows, assume an array s of m input samples, and an output array s 0 that will hold the n output samples of the resampling. Assume further that associated with each sample si is a weight w(si), and that the weights have been normalized to sum to 1. This can of course be done in time O(m), but typical efficient implementations keep a running weight total during weight generation, and then normalize their sampling range rather than normalizing the weights themselves. We thus discount the normalization cost in our analysis.

The na¨ıve approach to resampling has been re-invented many times. A correct, if inefficient, way to resample is via the pseudocode of Figure 1. The sample procedure selects the first sample such that the sum of weights in the input up to and including this sample is greater than some index value µ. The index value is chosen in resample by uniform ran

to sample(µ):

t ← 0

for i from 1 to m do

t ← t + w(si)

if t > µ then

return si

to resample:

for i from 1 to n do

µ ← random-real([0..1])

s

0

i ← sample(µ)

dom sampling from the distribution [0..1], with each output position being filled in turn.

Despite its poor performance, the na¨ıve algorithm has its advantages. It is easy to verify that it is a perfect sampling algorithm. It is easy to implement, and easy to parallelize. The expected running time is o(1 2mn).

To derive a O(m + n log m) algorithm from the na¨ıve algorithm, note that the linear scan of input samples in Figure 1 can be replaced with a binary search. One way to do this would be to treat the array of input samples as a heap. This heap-based algorithm, not shown here for space reasons, does dramatically improve on the performance of the na¨ıve algorithm without sacrificing correctness.

For some input particle distributions, a further constant factor improvement to both the na¨ıve and heap-based algorithms can be had by sorting or heapifying, respectively, the input particle array so that the largest particles are likely to be encountered first in the search. The amortized cost of these operations is small, but may be larger than the cost savings in typical distributions; the approach also adds a bit to the complexity of the implementation

. A Merge-based O(m + n log n) Resampling Algorithm

The real problem with the na¨ıve algorithm is not so much the cost per scan of the input as it is the fact that each scan is independent. It seems a shame not to try to do all the work in one scan. Let us generate an array u of n variates up-front, then sort it. At this point, a merge operation, as shown in Figure 2, can be used to generate all n outputs in a single pass over the m inputs. The merge operation is simple. Walk the input array once. Each time the sum of weights hits the current variate ui , output a sample and move to the next variate ui+1. The time complexity of the initial sort is O(n log n) and of the merge pass is O(m + n), for a total time complexity of O(m + n log n).

Complexity-wise, we seem to have simply moved the log factor of the heap-based algorithm from m to n, replacing an O(m + n log m) algorithm with an O(m + n log n) one.

to merge(u):

j ← 1

t ← u1

for i from 1 to n do

μ ← ui

while μ < t do

t ← t + w(sj)

j ← j + 1

s

0

i ← sj

However, the new algorithm has an important distinction. The log factor this time comes merely from sorting an array of uniform variates. If we could somehow generate the variates in sorted order (at amortized constant cost) we could make this approach run in time O(m + n). The next section shows how to achieve this.

An Optimal O(m + n) Resampling Algorithm

As discussed in the previous section, if we can generate the variates comprising a uniform sample of n values in increasing order, we can resample in time O(m + n). Assume without loss of generality that our goal is simply to generate the first variate in a uniform sample of n + 1 values. Call the first variate μ0, the set of remaining variates U and note that |U| = n. Now, for any given variate μi ∈ X, we have that

Pr(μ0 < μi) = 1 − μ0

Since this is independently true for each μi ,

we define p(μ0) = Pr(∀μi ∈ U . μ0 < μi) = (1 − μ0) n

Thus, if we successively generate n variates ui drawn from the distribution (1−μ0) n−i , those variates will be statistically indistinguishable from the set of variates produced by generating n uniform variates and then sorting them. To generate a variate from the target distribution, it is sufficient to observe that the likelihood of generating a variate μ is given by

μ =

R μ0

u=0 (1 − u)

ndu

R 1

u=0 (1 − u)

ndu

=

−(1−u)

n+1

n+1

μ0

u=0

−(1−u)n+1

n+1

1

u=0

=

−1

n+1

(1 − μ0)

n+1 − 1

0 −

−1

n+1

= 1 − (1 − μ0)

n+1

to randomize:

$u_1 \leftarrow (1 - \mu)$

1

n

for i from 2 to n do

$u_i \leftarrow u_{i-1} + (1 - u_{i-1})(1 - \mu)$

1

$n-i+1$

However, what we need is μ_0 in terms of μ, so we solve

$\mu = 1 - (1 - \mu_0)$

$n+1$

$(1 - \mu_0)$

$n+1 = 1 - \mu$

$\mu_0 = 1 - (1 - \mu)$

1

$n+1$

$\mu_0 = 1 - \mu$

1

$n+1$

(The last step is permissible because μ is a uniform deviate in the range 0..1, and therefore statistically equivalent to $(1 - \mu)$.)

We now have the formula we need for selecting the first deviate from a set of n in increasing order. To select the next deviate, we simply decrease n by 1, select a deviate from the whole range, and then scale and offset it to the remaining range. We repeat this process until n = 0. (Recall that $|U| = n$, so the last deviate will be selected when n = 0.) Figure 3 shows this process.

We now have the array u of deviates in sorted order that we need to feed the merge algorithm of the previous section. We thus have an $O(m + n)$ algorithm for random weighted selection.

***REGRESSION USING STATS MODELS**

Linear Regression

Linear models with independently and identically distributed errors, and for errors with heteroscedasticity or autocorrelation. This module allows estimation by ordinary least squares (OLS), weighted least squares (WLS), generalized least squares (GLS), and feasible generalized least squares with autocorrelated AR(p) errors.

See Module Reference for commands and arguments.

Examples

```
# Load modules and data
In [1]: import numpy as np

In [2]: import statsmodels.api as sm

In [3]: spector_data = sm.datasets.spector.load()

In [4]: spector_data.exog = sm.add_constant(spector_data.exog, prepend=False)

# Fit and summarize OLS model
In [5]: mod = sm.OLS(spector_data.endog, spector_data.exog)

In [6]: res = mod.fit()

In [7]: print(res.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  GRADE   R-squared:                       0.416
Model:                            OLS   Adj. R-squared:                  0.353
Method:                 Least Squares   F-statistic:                     6.646
Date:                Tue, 08 Feb 2022   Prob (F-statistic):            0.00157
Time:                        18:23:05   Log-Likelihood:                -12.978
No. Observations:                  32   AIC:                             33.96
Df Residuals:                      28   BIC:                             39.82
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
GPA            0.4639      0.162      2.864      0.008       0.132       0.796
TUCE           0.0105      0.019      0.539      0.594      -0.029       0.050
PSI            0.3786      0.139      2.720      0.011       0.093       0.664
const         -1.4980      0.524     -2.859      0.008      -2.571      -0.425
==============================================================================
Omnibus:                        0.176   Durbin-Watson:                   2.346
Prob(Omnibus):                  0.916   Jarque-Bera (JB):                0.167
Skew:                           0.141   Prob(JB):                        0.920
Kurtosis:                       2.786   Cond. No.                         176.
==============================================================================
```

Detailed examples can be found here:

- OLS
- WLS
- GLS
- Recursive LS
- Rolling LS

Technical Documentation

The statistical model is assumed to be

> $Y = X\beta + \mu$, where $\mu \sim N(0, \Sigma)$.

Depending on the properties of Σ, we have currently four classes available:

- GLS : generalized least squares for arbitrary covariance Σ
- OLS : ordinary least squares for i.i.d. errors $\Sigma = \mathbf{I}$
- WLS : weighted least squares for heteroskedastic errors $\mathbf{diag}(\Sigma)$
- GLSAR : feasible generalized least squares with autocorrelated AR(p) errors $\Sigma = \Sigma(\rho)$

All regression models define the same methods and follow the same structure, and can be used in a similar fashion. Some of them contain additional model specific methods and attributes.

GLS is the superclass of the other regression classes except for RecursiveLS, RollingWLS and RollingOLS.

General reference for regression models:

- D.C. Montgomery and E.A. Peck. “Introduction to Linear Regression Analysis.” 2nd. Ed., Wiley, 1992.

Econometrics references for regression models:

- R.Davidson and J.G. MacKinnon. “Econometric Theory and Methods,” Oxford, 2004.
- W.Green. “Econometric Analysis,” 5th ed., Pearson, 2003.

Attributes

The following is more verbose description of the attributes which is mostly common to all regression classes

pinv_wexog : array

The *p* x *n* Moore-Penrose pseudoinverse of the whitened design matrix. It is approximately equal to $\left(X^{T}\Sigma^{-1}X\right)^{-1}X^{T}\Psi$, where Ψ is defined such that $\Psi\Psi^{T}=\Sigma^{-1}$.

cholsimgainv : array

The *n* x *n* upper triangular matrix Ψ^{T} that satisfies $\Psi\Psi^{T}=\Sigma^{-1}$.

df_model : float

The model degrees of freedom. This is equal to *p* - 1, where *p* is the number of regressors. Note that the intercept is not counted as using a degree of freedom here.

df_resid : float

---

df_resid : float

The residual degrees of freedom. This is equal *n* - *p* where *n* is the number of observations and *p* is the number of parameters. Note that the intercept is counted as using a degree of freedom here.

llf : float

The value of the likelihood function of the fitted model.

nobs : float

The number of observations *n*

normalized_cov_params : array

A *p* x *p* array equal to $(X^{T}\Sigma^{-1}X)^{-1}$.

sigma : array

The *n* x *n* covariance matrix of the error terms: $\mu\sim N\left(0,\Sigma\right)$.

wexog : array

The whitened design matrix $\Psi^{T}X$.

wendog : array

The whitened response variable $\Psi^{T}Y$.

---

class statsmodels.regression.linear_model.OLS(*endog*, *exog=None*, *missing='none'*, *hasconst=None*, ***kwargs*)[source]¶

Ordinary Least Squares

Parameters
Endog: array_like

A 1-d endogenous response variable. The dependent variable.

Exog:array_like
A nobs x k array where *nobs* is the number of observations and *k* is the number of regressors. An intercept is not included by default and should be added by the user. See **statsmodels.tools.add_constant**.

Missing:str

Available options are ‘none’, ‘drop’, and ‘raise’. If ‘none’, no nan checking is done. If ‘drop’, any observations with nans are dropped. If ‘raise’, an error is raised. Default is ‘none’.

Hasconst: None or bool

Indicates whether the RHS includes a user-supplied constant. If True, a constant is not checked for and k_constant is set to 1 and all result statistics are calculated as if a constant is present. If False, a constant is not checked for and k_constant is set to 0.

****kwargs**

Extra arguments that are used to set model properties when using the formula interface.

See also

WLS

Fit a linear model using Weighted Least Squares.

GLS

Fit a linear model using Generalized Least Squares.

Notes

No constant is added by the model unless you are using formulas.

Examples

```
>>> import statsmodels.api as sm
>>> import numpy as np
>>> duncan_prestige = sm.datasets.get_rdataset("Duncan", "carData")
>>> Y = duncan_prestige.data['income']
>>> X = duncan_prestige.data['education']
>>> X = sm.add_constant(X)
>>> model = sm.OLS(Y,X)
>>> results = model.fit()
>>> results.params
const       10.603498
education    0.594859
dtype: float64
```

```
>>> results.tvalues
const        2.039813
education    6.892802
dtype: float64

>>> print(results.t_test([1, 0]))
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0            10.6035      5.198      2.040      0.048       0.120      21.087
==============================================================================

>>> print(results.f_test(np.identity(2)))
<F test: F=array([[159.63031026]]), p=1.2607168903696672e-20, df_denom=43,
df_num=2>
```

Attributes

Weights:scalar

Has an attribute weights = array(1.0) due to inheritance from WLS.

***NONLINEAR RELATIONSHIP**

Nonlinearity is a term used in statistics to describe **a situation where there is not a straight-line or direct relationship between an independent variable and a dependent variable**. In a nonlinear relationship, changes in the output do not change in direct proportion to changes in any of the inputs.

A nonlinear relationship is a type of relationship between two entities in which change in one entity does not correspond with constant change in the other entity. This might mean the relationship between the two entities seems unpredictable or virtually absent. However, nonlinear entities can be related to each other in ways that are fairly predictable, but simply more complex than in a linear relationship

UNDERSTANDING LINEAR RELATIONSHIP

A linear relationship exists when two quantities are proportional to each other. If you increase one of the quantities, the other quantity either increases or decreases at a constant rate. For example, if you get paid $10 an hour, there is a linear relationship between your hours worked and your pay. Working another hour always results in a $10 pay increase, regardless of how many hours you already worked.

Differentiating Linear and Nonlinear Relationships

Any relationship between two quantities that doesn't fit the definition of a linear relationship is called a nonlinear relationship. The easiest way to differentiate a linear relationship from a nonlinear relationship is by mapping them on a graph. Use the x-axis of the graph to represent one of the quantities and the y-axis to represent the other. Using the previous example, plot hours worked on the x-axis and money earned on the y-axis. Then plot some known data points on the graph, such as one hour worked = $10, two hours worked = $20, and three hours worked = $30. Since you can connect the points to form a straight line, you know you have a linear relationship.

Types of Nonlinear Relationships

Some nonlinear relationships are monotonic, meaning they always increase or decrease, but not both. Monotonic relationships differ from linear relationships because they do not increase or decrease at a constant rate. When graphed, they appear as curves. If a monotonic relationship occurs where increases in one entity cause a decrease in the other entity, this is called an inverse relationship. However, nonlinear relationships can also be too irregular to fit any of these categories.

Examples of Nonlinear Relationships

Nonlinear relationships, and often monotonic relationships, arise regularly when comparing geometrical measurements of a single shape. For example, there is a monotonic nonlinear relationship between the radius of a sphere and the volume of that same sphere. Nonlinear relationships also appear in real world situations, such as in the relationship between the value of a motorcycle and the amount of time you owned the motorcycle, or in the amount of time it takes to do a job in relation to the number of people there to help. If your boss raises your hourly rate to $15 per hour when you work overtime, the relationship of your hours worked to your pay acquired might become nonlinear

***LOGISTICS REGRESSION**

Logistic regression is **a statistical model that in its basic form uses a logistic function to model a binary dependent variable**, although many more complex extensions exist. In regression analysis, logistic regression (or logit regression) is estimating the parameters of a logistic model (a form of binary regression).

Logistic Regression was used in the biological sciences in early twentieth century. It was then used in many social science applications. Logistic Regression is used when the dependent variable(target) is categorical

For example,

- To predict whether an email is spam (1) or (0)
- Whether the tumor is malignant (1) or not (0)

Consider a scenario where we need to classify whether an email is spam or not. If we use linear regression for this problem, there is a need for setting up a threshold based on which classification can be done. Say if the actual class is malignant, predicted continuous value 0.4 and the threshold value is 0.5, the data point will be classified as not malignant which can lead to serious consequence in real time.

From this example, it can be inferred that linear regression is not suitable for classification problem. Linear regression is unbounded, and this brings logistic regression into picture. Their value strictly ranges from 0 to 1.

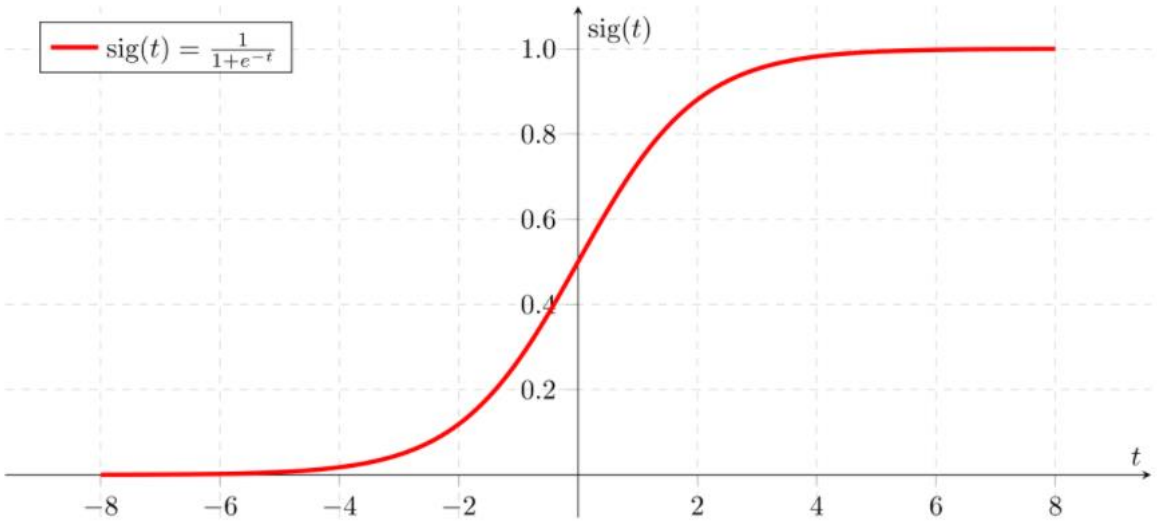
Simple Logistic Regression

Model

Output = 0 or 1

Hypothesis => Z = WX + B

hΘ(x) = sigmoid (Z)



f ‘Z’ goes to infinity, Y(predicted) will become 1 and if ‘Z’ goes to negative infinity, Y(predicted) will become 0

Analysis of the hypothesis

The output from the hypothesis is the estimated probability. This is used to infer how confident can predicted value be actual value when given an input X. Consider the below example,

X = [x0 x1] = [1 IP-Address]

Based on the x1 value, let’s say we obtained the estimated probability to be 0.8. This tells that there is 80% chance that an email will be spam.

Mathematically this can be written as,

$h_\Theta(x) = P(Y=1 | X; theta)$

Probability that Y=1 given X which is parameterized by 'theta'.

$P(Y=1 | X; theta) + P(Y=0 | X; theta) = 1$

$P(Y=0 | X; theta) = 1 - P(Y=1 | X; theta)$

This justifies the name 'logistic regression'. Data is fit into linear regression model, which then be acted upon by a logistic function predicting the target categorical dependent variable.

Types of Logistic Regression

1. Binary Logistic Regression

The categorical response has only two 2 possible outcomes. Example: Spam or Not

2. Multinomial Logistic Regression

Three or more categories without ordering. Example: Predicting which food is preferred more (Veg, Non-Veg, Vegan)

3. Ordinal Logistic Regression

Three or more categories with ordering. Example: Movie rating from 1 to 5

Decision Boundary

To predict which class a data belongs, a threshold can be set. Based upon this threshold, the obtained estimated probability is classified into classes.

Say, if predicted_value ≥ 0.5, then classify email as spam else as not spam.

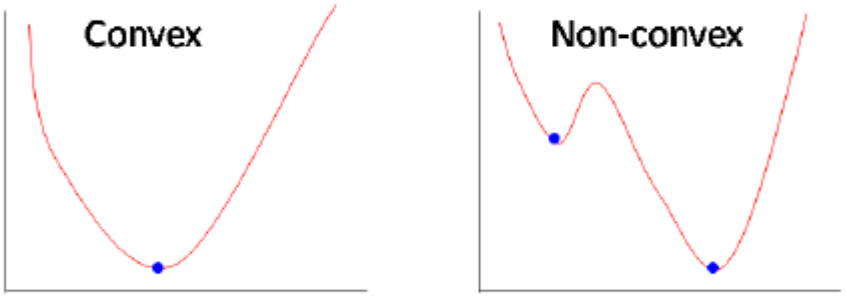
Decision boundary can be linear or non-linear. Polynomial order can be increased to get complex decision boundary.

Cost Function

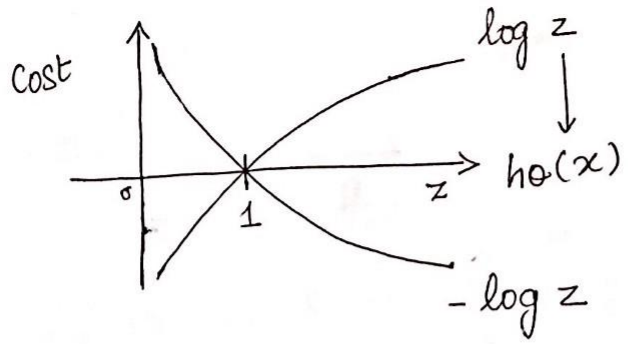
$$\text{Cost}(h_\Theta(x), Y(\text{actual})) = -\log(h_\Theta(x)) \text{ if } y=1$$
$$-\log(1-h_\Theta(x)) \text{ if } y=0$$

Why cost function which has been used for linear can not be used for logistic?

Linear regression uses mean squared error as its cost function. If this is used for logistic regression, then it will be a non-convex function of parameters (theta). Gradient descent will converge into global minimum only if the function is convex.



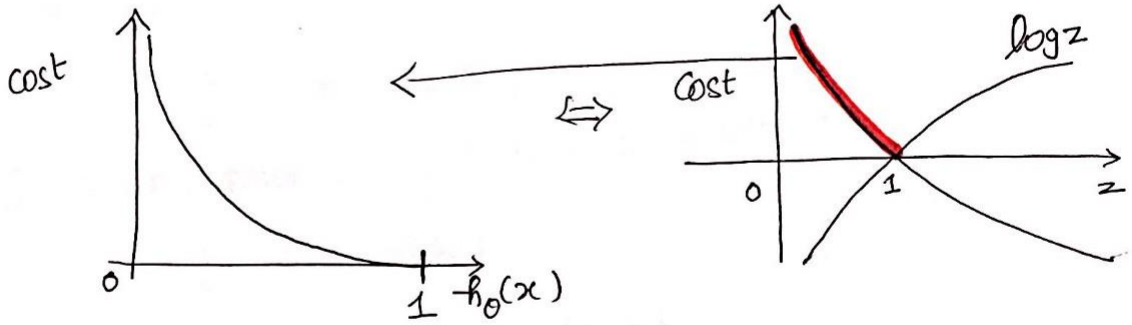
Cost function explanation



$$\text{Cost}(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) & \text{if } y = 1 \\ -\log(1 - h_\theta(x)) & \text{if } y = 0 \end{cases}$$

If $y = 1$,

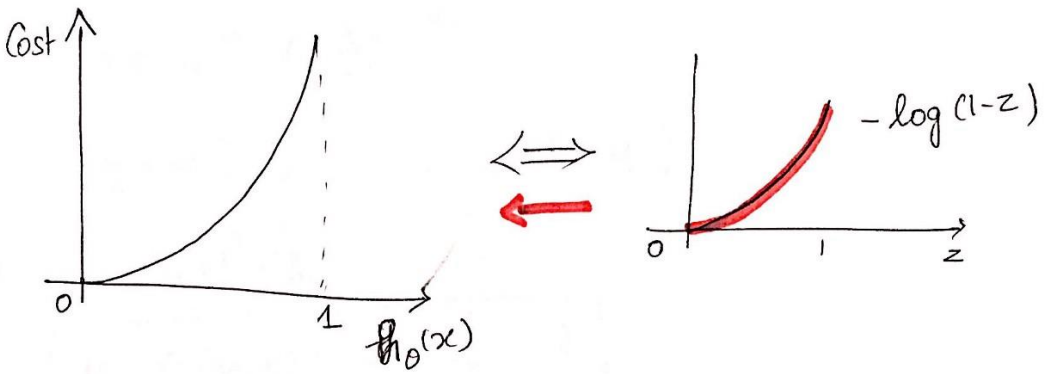
$$\text{Cost}(h_\theta(x), y) = -\log(h_\theta(x))$$



If, Cost = 0 $\Rightarrow$ $y = 1$ $\Rightarrow h_\theta(x) = 1$

Cost = unfinity for $h_\theta(x) = 0$

If $h_\theta(x) = 0$, it is similar to predicting $P(y = 1 | x; \theta) = 0$



If cost = 0 $\Rightarrow h_\theta(x) = 0 \Rightarrow y = 0$

cost = ∞ $\Rightarrow h_\theta(x) = 1$

If $h_\theta(x) = 1$, it is similar to predicting

$P(y=0|x;\theta) = 0$

Why this cost function?

Let us consider,

* $\hat{y} = P(y=1|x)$

  $\hat{y}$ is the probability that $Y=1$, given x

* $1-\hat{y} = P(y=0|x)$

* $P(y|x) = \hat{y}^{y} \cdot (1-\hat{y})^{(1-y)}$

If $y=1 \Rightarrow P(y|x) = \hat{y}$

$$\Rightarrow \quad \log\left(\hat{y}^{\,y} \cdot (1-\hat{y})^{(1-y)}\right)$$

$$\Rightarrow \quad y \log \hat{y} + (1-y) \log (1-\hat{y})$$

$$\Rightarrow \quad -L(\hat{y}, y)$$

$$\boxed{\log P(y|x) = -L(\hat{y}, y)}$$

This negative function is because when we train, we need to maximize the probability by minimizing loss function. Decreasing the cost will increase the maximum likelihood assuming that samples are drawn from an identically independent distribution.

Gradient

$$z = w_1 x_1 + w_2 x_2 + b \longrightarrow \hat{y} = a = \sigma(z) \longrightarrow L(\hat{y}, y)$$

$$\Leftrightarrow a = \hat{y}$$

$$w_1 \Rightarrow \frac{\partial(L)}{\partial w_1} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial(z)}{\partial w_1}$$

$$\frac{\partial L}{\partial a} = \frac{\partial}{\partial a}(-y \log a - (1-y)\log(1-a))$$

$$= -y\left(\frac{1}{a}\right) - (-1)\frac{(1-y)}{(1-a)}$$

$$\boxed{\frac{\partial L}{\partial a} = \left(\frac{-y}{a}\right) + \left(\frac{1-y}{1-a}\right)}$$

$$\boxed{\frac{\partial a}{\partial z} = a(1-a)}$$

$$\boxed{\frac{\partial z}{\partial w_1} = x_1}$$

$$\frac{\partial L}{\partial w_1} = \left(\left(\frac{-y}{a} + \frac{(1-y)}{1-a} \right) \cdot (a)(1-a) \right) \cdot x_1$$

$$= (a-y) \cdot x_1$$

Update for w_1,

$$\frac{\partial L}{\partial w_1} = (a-y) \cdot x_1$$

Here, $(a-y) = \frac{\partial L}{\partial z}$

$$\boxed{w_1 = w_1 - \alpha \frac{\partial L}{\partial w_1}}$$

Similarly, for all parameters

$$\boxed{w_i = w_i - \alpha \frac{\partial L}{\partial w_i}} \qquad i = 1, 2, \ldots, m$$

m = no of parameters

$$\boxed{b = b - \alpha \frac{\partial L}{\partial b}}$$

Where, $\frac{\partial L}{\partial b} = (a-y)$

Cost vs Number_of_Iterations

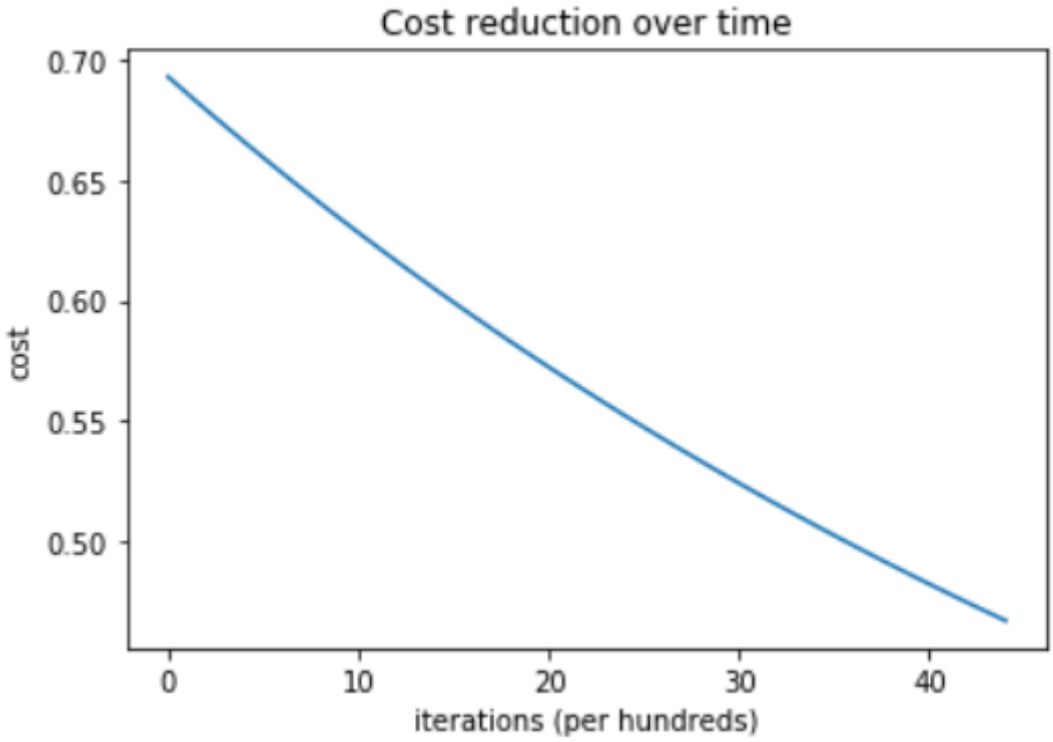


Figure 13: Cost Reduction

Train and test accuracy of the system is 100 %

This implementation is for binary logistic regression. For data with more than 2 classes, softmax regression has to be used.

*ESTIMATING PARAMETERS

Parameter estimation is defined as **the experimental determination of values of parameters that govern the system behavior, assuming that the structure of the process is known**

Parameter estimates (also called coefficients) are the change in the response associated with a one-unit change of the predictor, all other predictors being held constant.

The unknown model parameters are estimated using least-squares estimation.

A coefficient describes the size of the contribution of that predictor; a near-zero coefficient indicates that variable has little influence on the response. The sign of the coefficient indicates the direction of the relationship, although the sign can change if more terms are added to the model, so the interpretation is not particularly useful. A confidence interval expresses the uncertainty in

the estimate, under the assumption of normally distributed errors. Due to the central limit theorem, violation of the normality assumption is not a problem if the sample size is moderate.

- For quantitative terms, the coefficient represents the rate of change in the response per 1 unit change of the predictor, assuming all other predictors are held constant. The units of measurement for the coefficient are the units of response per unit of the predictor.

  For example, a coefficient for Height of 0.75, in a simple model for the response Weight (kg) with predictor Height (cm), could be expressed as 0.75 *kg per cm* which indicates a 0.75 kg weight increase per 1 cm in height.

  When a predictor is a logarithm transformation of the original variable, the coefficient is the rate of change in the response per 1 unit change in the log of the predictor. Commonly *base 2* log and *base 10* log are used as transforms. For *base 2* log the coefficient can be interpreted as the rate of change in the response when for a doubling of the predictor value. For *base 10* log the coefficient can be interpreted as the rate of change in the response when the predictor is multiplied by 10, or as the % change in the response per % change in the predictor.
- For categorical terms, there is a coefficient for each level:

- For nominal predictors the coefficients represent the difference between the level mean and the grand mean.

  Analyse-it uses *effect coding* for nominal terms (also known as the *mean deviation coding*). The sum of the parameter estimates for a categorical term using effect coding is equal to 0.
- For ordinal predictors, the coefficients represent the difference between the level mean and the baseline mean.

  Analyse-it uses *reference coding* for ordinal terms. The first level is used as the baseline or reference level.
- For the constant term, the coefficient is the response when all predictors are 0, and the units of measurement are the same as the response variable.

  A standardized parameter estimate (commonly known as standardized beta coefficient) removes the unit of measurement of predictor and response variables. They represent the change in standard deviations of the response for 1 standard deviation change of the predictor. You can use them to compare the relative effects of predictors measured on different scales.

  VIF, the variance inflation factor, represents the increase in the variance of the parameter estimate due to correlation (collinearity) between predictors. Collinearity between the predictors can lead to unstable parameter estimates. As a rule of thumb, VIF should be close to the minimum value of 1, indicating no collinearity. When VIF is greater than 5, there is high collinearity between predictors.

  A t-test formally tests the null hypothesis that the parameter is equal to 0, against the alternative hypothesis that it is not equal to 0. When the p-value is small, you can reject the null hypothesis and conclude that the parameter is not equal to 0 and it does contribute to the model.

  When a parameter is not deemed to contribute statistically to the model, you can consider removing it. However, you should be cautious of removing terms that are known to contribute by some underlying mechanism, regardless of the statistical significance of a hypothesis test, and recognize that removing a term can alter the effect of other terms

*ACCURACY

Data accuracy is one of the components of data quality. **It refers to whether the data values stored for an object are the correct values**. To be correct, a data values must be the right value and must be represented in a consistent and unambiguous form

*TIME SERIES ANALYSIS

For as long as we have been recording data, time has been a crucial factor. In time series analysis, time is a significant variable of the data. Times series analysis helps us study our world and learn how we progress within it.

Time series analysis is a specific way of analyzing a sequence of data points collected over an interval of time. In time series analysis, analysts record data points at consistent intervals over a set period of time rather than just recording the data points intermittently or randomly. However, this type of analysis is not merely the act of collecting data over time

What sets time series data apart from other data is that the analysis can show how variables change over time. In other words, time is a crucial variable because it shows how the data adjusts over the course of the data points as well as the final results. It provides an additional source of information and a set order of dependencies between the data.

Time series analysis typically requires a large number of data points to ensure consistency and reliability. An extensive data set ensures you have a representative sample size and that analysis can cut through noisy data. It also ensures that any trends or patterns discovered are not outliers and can account for seasonal variance. Additionally, time series data can be used for forecasting—predicting future data based on historical data

Time series analysis helps organizations understand the underlying causes of trends or systemic patterns over time. Using data visualizations, business users can see seasonal trends and dig deeper into why these trends occur. With modern analytics platforms, these visualizations can go far beyond line graphs.

When organizations analyze data over consistent intervals, they can also use time series forecasting to predict the likelihood of future events. Time series forecasting is part

of predictive analytics. It can show likely changes in the data, like seasonality or cyclic behavior, which provides a better understanding of data variables and helps forecast better.

For example, Des Moines Public Schools analyzed five years of student achievement data to identify at-risk students and track progress over time. Today's technology allows us to collect massive amounts of data every day and it's easier than ever to gather enough consistent data for comprehensive analysis

*TIME SERIES ANALYSIS EXAMPLE

Time series analysis is used for non-stationary data—things that are constantly fluctuating over time or are affected by time. Industries like finance, retail, and economics frequently use time series analysis because currency and sales are always changing. Stock market analysis is an excellent example of time series analysis in action, especially with automated trading algorithms. Likewise, time series analysis is ideal for forecasting weather changes, helping meteorologists predict everything from tomorrow's weather report to future years of climate change. Examples of time series analysis in action include:

- Weather data
- Rainfall measurements
- Temperature readings
- Heart rate monitoring (EKG)
- Brain monitoring (EEG)
- Quarterly sales
- Stock prices
- Automated stock trading
- Industry forecasts
- Interest rates

TIME SERIES ANALYSIS TYPE

Because time series analysis includes many categories or variations of data, analysts sometimes must make complex models. However, analysts can't account for all variances, and they can't generalize a specific model to every sample. Models that are too complex or that try to do too many things can lead to a lack of fit. Lack of fit or overfitting models lead to those models not distinguishing between random error and true relationships, leaving analysis skewed and forecasts incorrect.

Models of time series analysis include:

- **Classification:** Identifies and assigns categories to the data.

- **Curve fitting:** Plots the data along a curve to study the relationships of variables within the data.
- **Descriptive analysis:** Identifies patterns in time series data, like trends, cycles, or seasonal variation.
- **Explanative analysis:** Attempts to understand the data and the relationships within it, as well as cause and effect.
- **Exploratory analysis:** Highlights the main characteristics of the time series data, usually in a visual format.
- **Forecasting:** Predicts future data. This type is based on historical trends. It uses the historical data as a model for future data, predicting scenarios that could happen along future plot points.
- **Intervention analysis:** Studies how an event can change the data.
- **Segmentation:** Splits the data into segments to show the underlying properties of the source information

Data classification

Further, time series data can be classified into two main categories:

- **Stock time series data** means measuring attributes at a certain point in time, like a static snapshot of the information as it was.
- **Flow time series data** means measuring the activity of the attributes over a certain period, which is generally part of the total whole and makes up a portion of the results.

Data variations

In time series data, variations can occur sporadically throughout the data:

- **Functional analysis** can pick out the patterns and relationships within the data to identify notable events.
- **Trend analysis** means determining consistent movement in a certain direction. There are two types of trends: deterministic, where we can find the underlying cause, and stochastic, which is random and unexplainable.
- **Seasonal variation** describes events that occur at specific and regular intervals during the course of a year. Serial dependence occurs when data points close together in time tend to be related.

Time series analysis and forecasting models must define the types of data relevant to answering the business question. Once analysts have chosen the relevant data they want to analyze, they choose what types of analysis and techniques are the best fit.

Important Considerations for Time Series Analysis

While time series data is data collected over time, there are different types of data that describe how and when that time data was recorded. For example:

- **Time series data** is data that is recorded over consistent intervals of time.
- **Cross-sectional data** consists of several variables recorded at the same time.
- **Pooled data** is a combination of both time series data and cross-sectional data.

Time Series Analysis Models and Techniques

Just as there are many types and models, there are also a variety of methods to study data. Here are the three most common.

- **Box-Jenkins ARIMA models:** These univariate models are used to better understand a single time-dependent variable, such as temperature over time, and to predict future data points of variables. These models work on the assumption that the data is stationary. Analysts have to account for and remove as many differences and seasonalities in past data points as they can. Thankfully, the ARIMA model includes terms to account for moving averages, seasonal difference operators, and autoregressive terms within the model.
- **Box-Jenkins Multivariate Models:** Multivariate models are used to analyze more than one time-dependent variable, such as temperature and humidity, over time.
- **Holt-Winters Method:** The Holt-Winters method is an exponential smoothing technique. It is designed to predict outcomes, provided that the data points include seasonality

***MOVING AVERAGES**

In statistics, a moving average is a calculation used to analyze data points by creating a series of averages of different subsets of the full data set. In finance, a moving average (MA) is a stock indicator that is commonly used in technical analysis

Moving average is a simple, technical analysis tool. Moving averages are usually calculated to identify the trend direction of a stock or to determine its support and resistance levels. It is a trend-following—or lagging—indicator because it is based on past prices.

The longer the time period for the moving average, the greater the lag. So, a 200-day moving average will have a much greater degree of lag than a 20-day MA because it contains prices for the past 200 days. The 50-day and 200-day moving average figures for stocks are widely followed by investors and traders and are considered to be important trading signals.

Moving averages are a totally customizable indicator, which means that an investor can freely choose whatever time frame they want when calculating an average. The most common time periods used in moving averages are 15, 20, 30, 50, 100, and 200 days. The shorter the time span used to create the average, the more sensitive it will be to price changes. The longer the time span, the less sensitive the average will be.

Investors may choose different time periods of varying lengths to calculate moving averages based on their trading objectives. Shorter moving averages are typically used for short-term trading, while longer-term moving averages are more suited for long-term investors.

There is no correct time frame to use when setting up your moving averages. The best way to figure out which one works best for you is to experiment with a number of different time periods until you find one that fits your strategy.

Predicting trends in the stock market is no simple process. While it is impossible to predict the future movement of a specific stock, using technical analysis and research can help you make better predictions.

A rising moving average indicates that the security is in an uptrend, while a declining moving average indicates that it is in a downtrend. Similarly, upward momentum is confirmed with a bullish crossover, which occurs when a short-term moving average crosses above a longer-term moving average. Conversely, downward momentum is confirmed with a bearish crossover, which occurs when a short-term moving average crosses below a longer-term moving average.[1]

While calculating moving averages are useful in their own right, the calculation can also form the basis for other technical analysis indicators, such as the moving average convergence divergence (MACD).

The moving average convergence divergence (MACD) is used by traders to monitor the relationship between two moving averages. It is generally calculated by subtracting a 26-day exponential moving average from a 12-day exponential moving average.

When the MACD is positive, the short-term average is located above the long-term average. This is an indication of upward momentum. When the short-term average is below the long-term average, this is a sign that the momentum is downward. Many traders will also watch for a move above or below the zero line. A move above zero is a signal to buy, while a cross below zero is a signal to sell.

Types of Moving Averages

Simple Moving Average

The simplest form of a moving average, known as a simple moving average (SMA), is calculated by taking the arithmetic mean of a given set of values over a specified period of time. In other words, a set of numbers–or prices in the case of financial instruments–are added together and then divided by the number of prices in the set. The formula for calculating the simple moving average of a security is as follow

$$SMA = \frac{A_1 + A_2 + \ldots + A_n}{n}$$

where:*A*=Average in period *nn*=Number of time periods

Exponential Moving Average (EMA)

The exponential moving average is a type of moving average that gives more weight to recent prices in an attempt to make it more responsive to new information. To calculate an EMA, you must first compute the simple moving average (SMA) over a particular time period. Next, you must calculate the multiplier for weighting the EMA (referred to as the "smoothing factor"), which typically follows the formula: [2/(selected time period + 1)]. So, for a 20-day moving average, the multiplier would be [2/(20+1)]= 0.0952. Then you use the smoothing factor combined with the previous EMA to arrive at the current value. The EMA thus gives a higher weighting to recent prices, while the SMA assigns an equal weighting to all values

$$EMA_t = \left[V_t \times \left(\frac{s}{1+d}\right)\right] + EMA_y \times \left[1 - \left(\frac{s}{1+d}\right)\right]$$

where:

EMA_t = EMA today

V_t = Value today

EMA_y = EMA yesterday

s = Smoothing

d = Number of days

Simple Moving Average (SMA) vs. Exponential Moving Average (EMA)

The calculation for EMA puts more emphasis on the recent data points. Because of this, EMA is considered a weighted average calculation.

In the figure below, the number of time periods used in each average is identical–15–but the EMA responds more quickly to the changing prices than the SMA. You can also observe in the figure that the EMA has a higher value when the price is rising than the SMA (and it falls faster than the SMA when the price is declining). This responsiveness to price changes is the main reason why some traders prefer to use the EMA over the SMA.

Example of a Moving Average

The moving average is calculated differently depending on the type: SMA or EMA. Below, we look at a simple moving average (SMA) of a security with the following closing prices over 15 days:

- Week 1 (5 days): 20, 22, 24, 25, 23
- Week 2 (5 days): 26, 28, 26, 29, 27
- Week 3 (5 days): 28, 30, 27, 29, 28

A 10-day moving average would average out the closing prices for the first 10 days as the first data point. The next data point would drop the earliest price, add the price on day 11 and take the average.

Example of a Moving Average Indicator

A Bollinger Band® technical indicator has bands generally placed two standard deviations away from a simple moving average. In general, a move toward the upper band suggests the asset is becoming overbought, while a move close to the lower band suggests the asset is becoming oversold. Since standard deviation is used as a statistical measure of volatility, this indicator adjusts itself to market conditions.

What Does a Moving Average Indicate?

A moving average is a statistic that captures the average change in a data series over time. In finance, moving averages are often used by technical analysts to keep track of prices trends for specific securities. An upward trend in a moving average might signify an upswing in the price or momentum of a security, while a downward trend would be seen as a sign of decline. Today, there is a wide variety of moving averages to choose from, ranging from simple measures to complex formulas that require a computer program to efficiently calculate.

What Are Moving Averages Used for?

Moving averages are widely used in technical analysis, a branch of investing that seeks to understand and profit from the price movement patterns of securities and indices. Generally, technical analysts will use moving averages to detect whether a change in momentum is occurring for a security, such as if there is a sudden downward move in a security's price. Other times, they will use moving averages to confirm their suspicions that a change might be underway. For example, if a company's share price rises above its 200-day moving average, that might be taken as a bullish signal.

What Are Some Examples of Moving Averages?

Many different types of moving averages have been developed for use in investing. For example, the exponential moving average (EMA) is a type of moving average that gives more weight to more recent trading days. This type of moving average might be more useful for short-term traders for whom longer-term historical data might be less relevant. A simple moving average, on the other hand, is calculated by averaging a series of prices while giving equal weight to each of the prices involved

*MISSING VALUES

Missing value occurs when there is no data value for a variable in an observation. The phenomenon of missing value is universal in clinical researches involving big data

Missing value occurs when there is no data value for a variable in an observation. The phenomenon of missing value is universal in clinical researches involving big data. Nurses may forget to record urine output at a certain time point. Patients may have only one measurement of blood lactate, while the researcher is interested in exploring the impact of lactate trend on mortality outcome. Other reasons of missing values include but not limited to coding errors, faulty equipment and nonresponses (1). In statistical packages, some commands (e.g., logistic regression) may automatically deleted observations with missing values. There is no problem if there are a few incomplete observations. However, when there are a large number of observations with missing values, the default listwise deletion may result in significant loss of information. In such situation, analysts should take a close look at the missing patterns and find appropriate means to cope with it. The present article will introduce how missing values are handled in R, and provide some basic skills in dealing with missing values.

HOW MISSING VALUE IS HANDLED IN R

Missing value is represented by the symbol NA (not available) in R. When you read an Excel spreadsheet containing empty cells into R console, these empty cells will be replaced by NAs. This is different from STATA where empty cells are replaced with ".". The same missing value symbol is used in R for both numeric and character variables. R provides several functions for handling missing value

Table 1

R functions to handle missing values

Functions	Description
is.na()	To indicate which element is NA
na.rm ()	It is used within other functions, as an optional argument
na.fail()	It can be used to detect missing values in a dataset
na.omit()	It returns the object with incomplete cases removed
complete.cases()	It returns a logical vector indicating which cases have no missing values
na.tree.replace()	Adds a new level called "NA" to any discrete predictor in a data frame that contains NAs

Functions	Description
na.gam.replace()	A discrete variable with missing value is replaced by a new level labelled “NA”; A missing numeric vector has its missing values replaced by the mean of the non-missing values

The function is.na () is the most commonly used method to indicate which element is NA. It returns logical values (FALSE or TRUE) and have the same length as its argument. Suppose we have six patients. Five lactate values are recoded and one is missing.

The returning vector of is.na () have the length of six. In the fourth place the value is TRUE, indicating lactate value is missing in the forth patient.

Someone may think of using logical test (e.g., lactate==NA) to examine the missing pattern. This can never be TRUE because missing values are considered non-comparable and you have to use missing value functions. The “==” operator returns NA when either argument is NA. By using which() function, you can locate which element of a vector contains NA. In the example, which() function returns 4, indicating the forth patient has missing lactate.

Next, you may want to describe lactate levels of the six patients. In statistical description, mean, variance and standard deviation are among the most commonly used statistics

. All of these functions return NA value because the vector lactate contains missing values. Fortunately, there is a function na.rm () to remove NAs when applying statistical functions.

The results are exactly what you want. Both mean and standard deviation are calculated based on the five patients with lactate values available.

DATA FRAME FROM MISSING VALUE

In real world setting, what you encounter is missing values in a data frame. Thus, this section focuses on how to handle data frames with missing values. First we create a data frame of three variables and five observations.

```
>ptid<-c(1,2,3,4,5)
>sex<-c("m","f",NA,"f","m")
>lactate<-c(0.2,3.3,4.5,NA,6.1)
> data<-data.frame(ptid,sex,lactate)
> data
```

	ptid	sex	lactate
1	1	m	0.2
2	2	f	3.3
3	3	<NA>	4.5
4	4	f	NA
5	5	m	6.1

Note that the third patient have missing value on sex, and the forth patient have missing value on lactate.

```
> na.fail(data)
Error in na.fail.default(data) : missing values in c
```

The function na.fail() can be used to detect missing values in a dataset. If there is no missing value it returns the assigned object (see below). If there is missing value it returns an error message indicating there is one or more missing values in the dataset.

Although some good default settings in regression model can effectively ignore observations with missing values, it is useful to create a new data frame by omitting observations with missing values.

```
> na.omit(data)
```

	ptid	sex	lactate
1	1	m	0.2
2	2	f	3.3
5	5	m	6.1

The function na.omit() returns the object with incomplete cases removed. As you can see, the third and forth patients with missing value in one variable are removed. Alternatively, the same purpose can be achieved by using the following code.

```
> complete.data<- data[complete.cases(dat
> complete.data
```

	ptid	sex	lactate
1	1	m	0.2
2	2	f	3.3
5	5	m	6.1

By applying na.fail() to the complete dataset obtained by na.omit(), the error message is replaced by a new complete dataset. The test passed when the new dataset contains no NAs.

```
> na.fail(complete.data)
  ptid sex lactate
1    1   m     0.2
2    2   f     3.3
5    5   m     6.1
```

Sometimes you may want to locate which variable of a patient contains missing value. Try functions is.na() and which() as described previously.

```
> is.na(data)
      ptid  sex  lactate
[1,] FALSE FALSE FALSE
[2,] FALSE FALSE FALSE
[3,] FALSE TRUE  FALSE
[4,] FALSE FALSE TRUE
[5,] FALSE FALSE FALSE
> which(is.na(data))
[1] 8 14
```

It returns 8 and 14 indicating the location of missing values. R counts the place in a column-wise fashion. There is no problem when the dataset is not large. However, if the dataset is large (always the case in big data exploration) you are overwhelmed by a long list of location numbers. The following code helps you to select observations with at least one missing value

As expected, the returned results indicate the third and forth patients who have at least one missing variable. The function lapply () returns a list of the locations of missing values across variables ptid, sex and lactate. Unlist () simplifies list structure to produce a vector. When there is more than one missing value in an observation, unique() simplified it to list observations with missing values only once.

In occasions, you may be interested in counting missing values for a variable (number of missing values in a column). Then you can restrict your analysis to variables with less than one fifth missing variables

Missing values in regression model

Regression model building is probably the most commonly used in statistical analysis. However, details on missing values are always omitted in regression model fitting. With

small number of missing values, it is safe to fit a model by default argument. Problem may arise with substantial number of missing values and analysts have to understand how missing values are handled in model building. Here the difference between na.omit() and na.exclude() will be shown.

For the illustration purpose, I will regress lactate on patient id. Of course, this has no practical meaning.

As you can see, the na.exclude() function pads the residuals and fitted values with NAs where there are missing values. However, the na.omit() function simply exclude observations with missing values. In this regard, na.exclude() is a placeholder for missing values

Some advanced functions for missing values

There are situations when you don't want to simply delete observations with missing values. Or missing values may have special clinical relevance. In our example, missing lactate values may indicate that the patient have been fully recovered from shock (e.g., lactate is a biomarker of tissue hypoperfusion and hypoxia and clinicians typically do not order lactate for patients who are hemodynamically stable). Therefore, NAs in lactate indicates stable patients. Because NAs contain important information, it is wise to add a new category called "NA" to replace the missing values. You can try na.tree.replace() in the tree package for this purpose (2). However, this function is limited by the fact that it is only applicable for discrete variables with missing value. I create a new data frame for the illustration.

The data frame data.discrete contains three variables and the last two discrete variables have NAs. The error message returned by na.fail() indicates NAs in the data frame data.discrete. Then, a new data frame called newdata.discrete is created by using na.tree.replace(). The na.fail() function returns the new data frame without error. As you can see missing values are replaced by string value "NA". The limitation of na.tree.replace() is that it stops if any continuous variable contains NAs.

```
> na.tree.replace(data)
Error in na.tree.replace(data):
    continuous variable lactate contained N
```

Note that the data frame newdata pass the na.fail () test and returns the new dataset. For variable sex the missing value is replaced by string value "NA", and the missing numeric value of lactate is replaced by the mean of the available lactate values.

***SERIAL COLLERATION**

Serial correlation occurs in a time series when a variable and a lagged version of itself (for instance a variable at times T and at T-1) are observed to be correlated with one another over periods of time. Repeating patterns often show serial correlation when the level of a variable

affects its future level. In finance, this correlation is used by technical analysts to determine how well the past price of a security predicts the future price.

Serial correlation is similar to the statistical concepts of autocorrelation or lagged correlation.

Serial Correlation Explained

Serial correlation is used in statistics to describe the relationship between observations of the same variable over specific periods. If a variable's serial correlation is measured as zero, there is no correlation, and each of the observations is independent of one another. Conversely, if a variable's serial correlation skews toward one, the observations are serially correlated, and future observations are affected by past values. Essentially, a variable that is serially correlated has a pattern and is not random.

Error terms occur when a model is not completely accurate and results in differing results during real-world applications. When error terms from different (usually adjacent) periods (or cross-section observations) are correlated, the error term is serially correlated. Serial correlation occurs in time-series studies when the errors associated with a given period carry over into future periods. For example, when predicting the growth of stock dividends, an overestimate in one year will lead to overestimates in succeeding years.

Serial correlation can make simulated trading models more accurate, which helps the investor develop a less risky investment strategy.

Technical analysis uses measures of serial correlation when analyzing a security's pattern. The analysis is based entirely on a stock's price movement and associated volume rather than a company's fundamentals. Practitioners of technical analysis, if they use serial correlation correctly, identify and validate the profitable patterns or a security or group of securities and spot investment opportunities.

The Concept of Serial Correlation

Serial correlation was originally used in engineering to determine how a signal, such as a computer signal or radio wave, varies compared to itself over time. The concept grew in popularity in economic circles as economists and practitioners of econometrics used the measure to analyze economic data over time.

Almost all large financial institutions now have quantitative analysts, known as quants, on staff. These financial trading analysts use technical analysis and other statistical inferences to analyze and predict the stock market. These modelers attempt to identify the structure of the correlations to improve forecasts and the potential profitability of a strategy. In addition, identifying the correlation structure improves the realism of any simulated time series based on the model. Accurate simulations reduce the risk of investment strategies.

Quants are integral to the success of many of these financial institutions since they provide market models that the institution then uses as the basis for its investment strategy.

Serial correlation among these quants is determined using the Durbin-Watson (DW) test. The correlation can be either positive or negative. A stock price displaying positive serial correlation has a positive pattern. A security that has a negative serial correlation has a negative influence on itself over time

***AUTOCORRELATION**

Autocorrelation **represents the degree of similarity between a given time series and a lagged version of itself over successive time intervals**. Autocorrelation measures the relationship between a variable's current value and its past values

Types of Autocorrelation

The most common form of autocorrelation is **first-order serial correlation**, which can either be positive or negative.

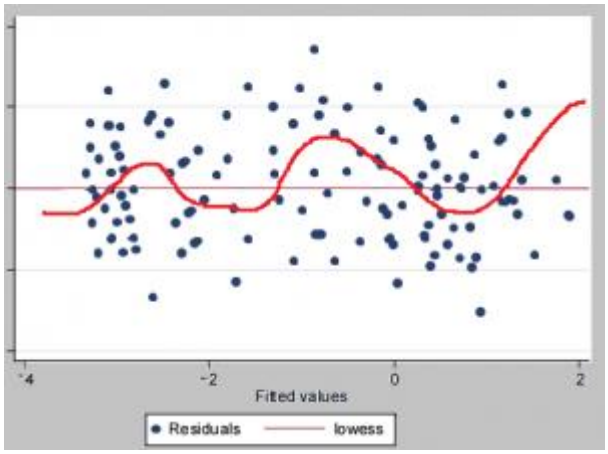
- Positive serial correlation is where a positive error in one period carries over into a positive error for the following period.
- Negative serial correlation is where a negative error in one period carries over into a negative error for the following period.

Second-order serial correlation is where an error affects data two time periods later. This can happen when your data has seasonality. Orders higher than second-order do happen, but they are rare.

Testing for Autocorrelation

You can test for autocorrelation with:

- A plot of residuals. Plot e_t against t and look for clusters of successive residuals on one side of the zero line. You can also try adding a Lowess line, as in the image below.
- A Durbin-Watson test.
- A Lagrange Multiplier Test.
- Ljung Box Test.
- A correlogram. A pattern in the results is an indication for autocorrelation. Any values above zero should be looked at with suspicion.
- The Moran's I statistic, which is similar to a correlation coefficient.



Measuring Serial Correlations

Serial correlations, when they exist, can be either **positive** or **negative**.

- Positive serial correlations indicate that value changes between the current price of a security and future prices are likely to be similar to the value changes between recent past prices and the current price.
- A negative serial correlation indicates that value changes between the current price and future prices are likely to move in the opposite direction as the value changes between past prices and the current price.

When the variable of a security's current price and its price in a prior time period exhibit positive serial correlation, they display what is known as mean aversion.

Aversion from the mean indicates that price changes in the security are prone to following trends and that, over periods of time, they will show higher standard deviations than would be the case with no correlation.

There is a wide variety of complex statistical formulas that can be used to measure serial correlation; however, most formulas calculate serial correlation with values ranging from -1 to +1.

A serial correlation value of zero indicates that no correlation exists. In other words, there is no observable relationship or pattern that exists between the current value of a variable and its value during previous time periods. Values nearer to +1 indicate a positive serial correlation, while values between zero and -1 indicate a negative serial correlation.

Use of Serial Correlation in Financial Modeling

Detecting and implementing the use of serial correlations in building financial models has become increasingly popular since the initial widespread use of computer technology in the 1980s.

Investment banks and other financial institutions now regularly incorporate the study of serial correlations to help improve forecast models for investment returns by detecting patterns that may occur in price changes over time.

By improving the accuracy of financial models, the use of serial correlation measures can serve to help maximize returns on investment, reduce investment risk, or both.

The study of serial correlations did not actually originate in the financial services industry – it originated in the world of engineering. The first studies of serial correlations were studies of how signals, such as radio broadcast signals, varied over successive time periods.

After such studies proved fruitful, economists and financial analysts gradually began to consider serial correlations between the values of security prices and various economic metrics, such as interest rates or gross domestic product (GDP).

Correlations can be measured using the =CORREL formula in Excel.

Example – Spotting Momentum Stocks

An example of how serial correlation can be used in predicting future price movements of a security can be found in momentum stocks.

Momentum stocks are stocks which, historically, have exhibited price movements that reveal sustained trends. That is, once a stock price begins moving in one direction, it tends to gain momentum and continue moving in the same direction over successive time periods.

Momentum stocks can be identified because they will exhibit positive serial correlation. The current price of the stock can be shown to have a positive correlation with the stock's price in previous time periods.

An investor can use this knowledge to profit from buying into identified momentum stocks once they begin exhibiting a price trend.

The investor purchases the stock based on the assumption that future price changes will tend to resemble recent past price changes – in other words, the stock will continue trending for at least some time period into the future.

Additional Resources

Thank you for reading CFI's guide to Serial Correlation. To keep learning and developing your knowledge of financial analysis, we highly recommend the additional resources below:

- Decoupling
- Negative Correlation
- Multiple Linear Regression
- Financial Forecasting

***INTRODUCTION TO SURVIVAL ANALYSIS**

Introduction to Survival Analysis

Understand the basic concepts of survival analysis and what tasks it can be used for!

In our extremely competitive times, all businesses face the problem of customer churn/retention. To quickly give some context, churn happens when the customer stops using the services of a company (stops purchasing, cancels the subscription, etc.). Retention refers to keeping the clients of a business active (the definition of active highly depends on the business model).

Intuitively, companies want to increase retention by preventing churn. This way, their relationship with the customers is longer and thus potentially more profitable. What is more, in most cases the company's cost of retaining a customer is much lower than that of acquiring a new customer, for example, via performance marketing. For businesses, the concept of retention is closely connected to customer lifetime value (CLV), which the businesses want to maximize. But that is a topic for another article.

With this article, I want to start a short series focusing on survival analysis, which is often an underestimated, yet very interesting branch of statistical learning. In this article, I provide a general introduction to survival analysis and its building blocks. First I explain the required concepts and then describe different approaches to analyzing time-to-event data. Let's start!

Introduction to Survival Analysis

Survival analysis is a field of statistics that focuses on analyzing the expected time until a certain event happens. Originally, this branch of statistics developed around measuring the effects of medical treatment on patients' survival in clinical trials. For example, imagine a group of cancer patients who are administered a certain new form of treatment. Survival analysis can be used for analyzing the results of that treatment in terms of the patients' life expectancy.

However, survival analysis is not restricted to investigating deaths and can be just as well used for determining the time until a machine fails or — what may at first sound a bit counterintuitively— a user of a certain platform converts to a premium service. That is possible because survival analysis focuses on the time

- the event of interest is clearly defined and well-specified, so there is no ambiguity about whether it happened or not,
- the event can occur only once for each subject — this is clear in case of death, but if we applied the analysis to churn, this might be a more complicated case, as a churned user might be reactivated and churn again.

We have already established that survival analysis is used for modeling the **time-to-event series**, in other words, lifetimes (hence also the name of the Python library which is the go-to tool for this kind of analyses). Generally speaking, we can use survival analysis to try to answer questions like:

- what percentage of the population will survive past a certain time?

- of the survivors, what will be their death/failure rate?
- how do particular characteristics (for example, such features as age, gender, geographical location, etc.) affect the probability of survival?

Having briefly described the general idea of survival analysis, it is time to introduce a few concepts that are crucial for a thorough understanding of the subject.

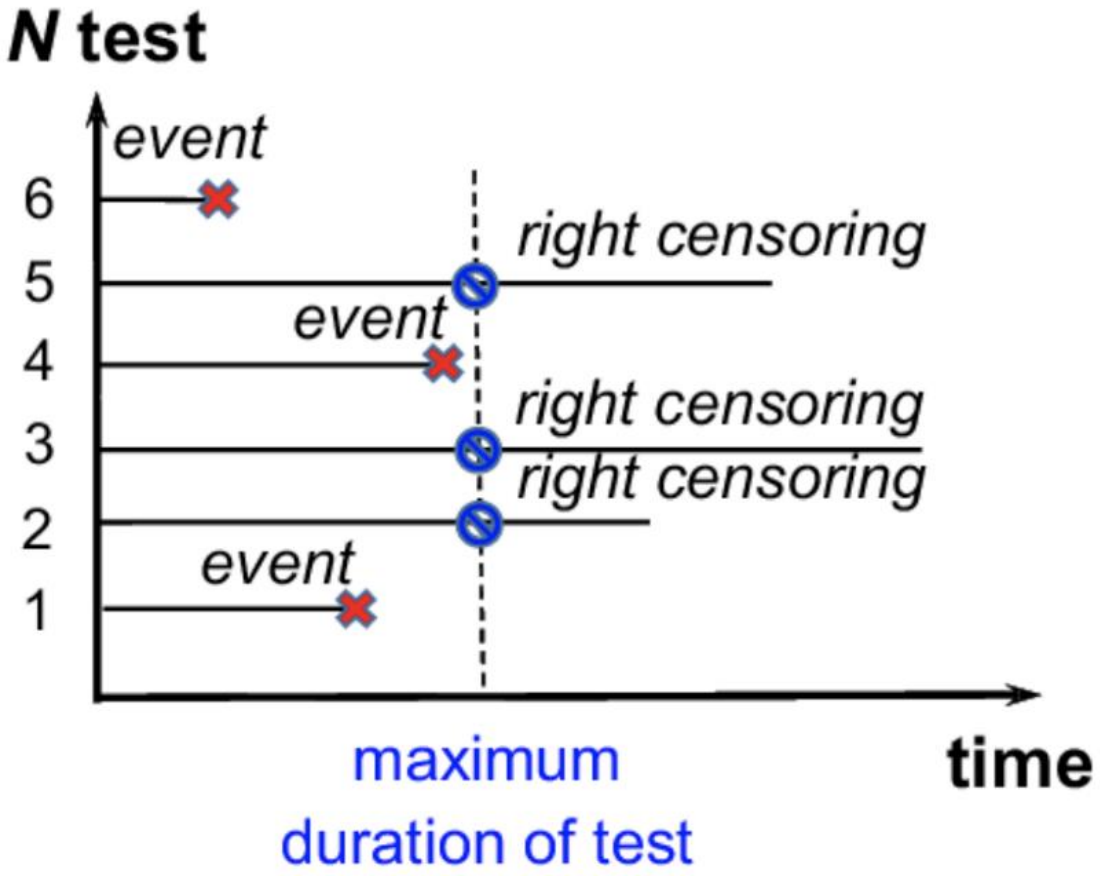
Censoring

Censoring can be described as the missing data problem in the domain of survival analysis. Observations are *censored* when the information about their survival time is incomplete. There are different kinds of censoring, such as:

- right-censoring,
- interval-censoring,
- left-censoring.

To keep this section short, we just discuss the one that is encountered most frequently — **right-censoring**. Let's come back to the example with cancer treatment. Imagine, that the study of the effects of the new medicine lasts 5 years (this is an arbitrary number, not actually based on anything). It can happen that after 5 years, some of the patients survived and thus have not experienced the death event. At the same time, the authors of the study lost contact with some patients — they might have relocated to another country, they might have actually died, but no confirmation was ever received. Those cases are affected by right-censoring, that is, their true survival time

is equal to or greater than the observed survival time (in this case, the 5 years of the study). The following image illustrates right-censoring



The existence of censoring is also the reason why we cannot use simple OLS for problems in the survival analysis. That is because OLS effectively draws a regression line that minimizes the sum of squared errors. But for censored data, the error terms are unknown and therefore we cannot minimize the MSE. Applying some simple solutions such as using the censorship date as the date of the death event or dropping the censored observations can severely bias the results.

For information regarding different kinds of censoring, please go here.

The Survival Function

The **survival function** is a function of time (*t*) and can be represented as

$$S(t) = \Pr(T > t)$$

where *Pr()* stands for the probability and *T* for the time of the event of interest for a random observation from the sample. We can interpret the survival function as the probability of the event of interest (for example, the death event) not occurring by the time *t.*

The survival function takes values in the range between 0 and 1 (inclusive) and is a non-increasing function of *t.*

The Hazard Function

We can think of the **hazard function** (or hazard rate) as the probability of the subject experiencing the event of interest within a small (or to be more precise, infinitesimal) interval of time, assuming that the subject has survived up until the beginning of the said interval. The hazard function can be represented as:

$$\lambda(t) = \lim_{dt \to 0} \frac{\Pr\{t \leq T < t + dt | T \geq t\}}{dt}.$$

where the expression in the numerator is the conditional probability of the event of interest occurring in the given time interval, provided it has not happened before. *dt* in the denominator is the width of the considered interval of time. When we divide the former by the latter, we effectively obtain the rate of the event's occurrence per unit of time. Lastly, by taking

the limit as the width of the interval goes to zero, we end up with the instantaneous rate of occurrence, so the risk of an event happening at a particular point in time.

You might wonder why the hazard rate is defined using this small interval of time. The reason for that lies in the fact that the probability of a continuous random variable being equal to a particular value is zero. That is why we need to consider the probability of the event happening in a very small interval of time.

Technical note: to be theoretically correct, it is important to mention that the hazard function is not actually a probability and the name *hazard rate* is the more fitting one. That is because even though the expression in the numerator is the probability, the *dt* in the denominator can actually result in a value of the hazard rate greater than 1 (it is still limited to 0 at the lower interval).

Lastly, the survival and hazard functions are related to each other as specified by the following formula:

$$S(t) = \exp\{-\int_0^t \lambda(x)dx\}$$

Different approaches to Survival Analysis

As survival analysis is an entire domain of different statistical methods for working with time-to-event series, there are naturally many different approaches we could follow. On a high level, we could split them into three main groups:

- **Non-parametric** — with these approaches, we make no assumptions about the underlying distribution of data. Perhaps the most popular example from this group is the **Kaplan-Meier curve**, which — in short — is a method of estimating and plotting the survival probability as a function of time.

- **Semi-parametric** — as you could have guessed, this group is in between the two extremes and makes very few assumptions. Most importantly, there are no assumptions about the shape of the hazard function/rate. The most popular method from this group is the **Cox regression**, which we can use to identify the relationship between the hazard function and a set of explanatory variables (predictors).

- **Parametric** — you might have encountered this approach while doing your studies. The idea is to use some statistical distributions (some of the popular ones include exponential, log, Weibull, or Lomax) to estimate how long a subject will survive. Often, we use maximum likelihood estimation (MLE) to fit the distribution (or actually the distribution's parameters) to the data for the best performance.

XXXX