Experimental Design: Principles and Practices and Statistics Review

1

1.1 The Strategy of Experimentation

Observing a system or process while it is in operation is an important part of the learning process and is an integral part of understanding how systems and processes work. The great New York Yankees catcher Yogi Berra said that "… you can observe a lot just by watching." However, to understand what happens to a process when you change certain input factors, you have to do more than just watch—you actually have to change the factors. This means that to really understand cause-and-effect relationships in a system, you must deliberately change the input variables to the system and observe the changes in the system output that these changes to the inputs produce. In other words, you need to conduct **experiments** on the system. Observations on a system or process can lead to theories or hypotheses about what makes the system work, but experiments of the type described above are required to demonstrate that these theories are correct.

Experiments are conducted in virtually all fields of science, engineering, industry, and business, usually to discover something about a particular process, product, or system. Each experimental **run** is really just a **test**. Obviously, testing plays a vital role in how products are designed and improved, how processes are developed and optimized, and how systems perform. We define an **experiment** as a series of tests in which deliberate changes are made to the controllable input variables of a process or system with the objective of identifying the reasons for the observed changes in the output or response variables of the system. Specifically, we usually want to determine which input variables are responsible for the observed changes in response to develop a model relating the response to the important input variables and to use this model for making decisions about the system. The person conducting the experiment is called the **experimenter**.

This book is about planning and conducting experiments as well as analyzing the resulting data so that valid and objective conclusions are obtained. Our focus is on experiments in engineering, science, and business. Experimentation plays an important role in **technology commercialization** and **product realization** activities, which consist of new product design and formulation, manufacturing process development, and process improvement. The objective in many cases may be to develop a **robust** process, that is, a process affected minimally by external sources of variability. There are also many applications of designed experiments in a nonmanufacturing or non-product-development setting, such as marketing, service operations, e-commerce, and general business operations.

As an example of an experiment, suppose that a biomedical engineer is designing a new pump for the intravenous delivery of a drug. The pump should deliver a constant quantity or dose of the drug over a specified period of time. The primary design parameter of the pump that the engineer is considering at this stage is the diameter of the pump cylinder. Two diameters are feasible: 3/16th in. and 1/4th in. Here, the objective of the experimenter is to determine which diameter produces the appropriate flow rate for the particular type of pump. In this type of experiment, a common approach would be to build several prototypes and subject each prototype to a standard test to determine the effect of the cylinder diameter on the drug delivery rate. The average delivery rate

that is observed for each cylinder diameter would be used to determine the best pump design to use in the product.

This is a relatively simple experiment. It only has one **factor**, the cylinder diameter, and the factor only has two **levels**. However, with a little thought, a number of important questions come to mind:

1. Are these two diameters the only ones that should be considered?
2. Are there any other factors that might potentially affect the drug delivery rate? Should those be investigated in this experiment (such as the length of the cylinder, or the length of the plunger, or the pressure at which the pump operates)?
3. How many prototypes of each cylinder should be tested?
4. How should the experiment actually be conducted; that is, in what order should the prototypes be tested?
5. What method of data analysis should be used?
6. What difference in average drug delivery rate between the two cylinder diameters will be considered important?

These are all important questions. All of them, and perhaps many others, need to be answered satisfactorily before conducting the experiment.

As a second example, consider the design of the home page for an e-commerce business. A great deal of business activity is generated through this website, so its design is a very important aspect of business operations. Suppose that the website has the following components: (1) a photo flash image, (2) a main headline, (3) a sub headline, (4) a main text copy, and (5) a main image on the right. The response variable is the click-through rate, that is, the number of visitors who click through into the site divided by the total number of visitors to the site. The click-through rate can be tracked automatically for a particular website. Suppose that there are five choices for the photo flash image, four choices for the main headline, three choices for the sub headline, two choices for the main text copy, and two choices for the main image.

This is a more complex experiment involving five factors. The most conservative classical approach to a multifactor experiment is to test all possible combinations of factor levels. This is running a **factorial design**. If we use a factorial design for this experiment, Web pages for all possible combinations of these factor levels must be constructed and tested (240 Web pages). Even if a website design can be tested in a few hours once it is in place, the logistical aspects of designing 240 Web pages are formidable. Fortunately, in multifactor experiments, we often do not need to run a complete factorial; a **fractional factorial** experiment that uses a small number of the possible Web page designs would likely be successful. This experiment would require a fractional factorial where the factors have different numbers of levels.

Finally, consider designing an aircraft engine. The engine is a commercial turbofan, intended to operate in the cruise configuration at 40,000 ft and 0.8 Mach. The design parameters that need to be studied include inlet flow, fan pressure ratio, overall pressure, stator outlet temperature, and many other factors. The output response variables in this system are specific fuel consumption and engine thrust. In designing this system, it would be prohibitively expensive to build prototypes or actual test articles early in the design process, so the engineers use a **computer model** of the system that allows them to focus on the key design parameters of the engine and to vary them in an effort to optimize the performance of the engine. Designed experiments can be employed with the computer model of the engine to determine the most important design parameters and their optimal settings. Engineers frequently employ various types of computer models (finite element models, computational fluid dynamic models, electrical circuit simulators, traffic flow models, and so on) for design activities, and experimental design techniques are an integral part of how to employ those computer models to make the engineering design and development activities more effective and efficient.

Experimentation is a vital part of the **scientific** (or **engineering**) **method**. Of course, there are situations where the scientific phenomena are so well understood that useful results including mathematical models can be developed directly by applying these well-understood principles. The models of such phenomena that follow directly from the physical mechanism are usually called **mechanistic models**. A simple example is the familiar equation for current flow in an electrical circuit, Ohm's law, $E = IR$. However, most problems in science, engineering, and business require **observation** of the system at work and **experimentation** to elucidate information about why and how it works. Well-designed experiments can often lead to a model of system performance; such experimentally determined models are called **empirical models**. The focus of this book is on designing an experiment so that the experimenter can obtain a valid and useful empirical model of the system under study. These empirical models can be manipulated and analyzed just as a mechanistic model can.

A well-designed experiment is important because the results and conclusions that can be drawn from the experiment depend, to a large extent, on the method of data collection. To illustrate this point, suppose that the bioengineer in the drug delivery pump experiment used all of the smaller diameter cylinders in morning tests and all of the larger diameter cylinders in afternoon tests. She plans to compare the sample average drug delivery rates from the two sets of tests to draw conclusions. It is possible that there is a temperature difference between the morning and afternoon test period and that temperature could have some impact on the performance of the pump. Now, when the average drug delivery rates are compared, she is unable to say how much of the observed difference is the result of the two different time periods and how much is the result of inherent differences between the cylinder diameters. Thus, the method of data collection has directly impacted the conclusions that can be drawn from the experiment.

In general, experiments are used to study the performance of processes and systems.

The process or system can be represented by the model shown in Figure 1.1. We can usually visualize the process as a combination of resources (which could include operations, machines, methods, people, computer code and software, but are not limited to those explicit items) that transforms some set of input variables or design factors (which could be component parts, raw materials, personnel or operator actions, or information, for example) into an output that has one or more observable response variables. Some of the design factors $x_1, x_2, \ldots, x_p$ are controllable, whereas other variables $z_1, z_2, \ldots, z_q$ are uncontrollable (although they may be controllable for purposes of a test). The objectives of the experiment may include the following:

1. Determining which factors are most influential on the response y
2. Determining where to set the influential xs so that the response y is almost always at or near its most desired value
3. Determining where to set the influential xs so that variability in y is small
4. Determining where to set the influential xs so that the effects of the uncontrollable variables $z_1, z_2, \ldots z_q$ are minimized.

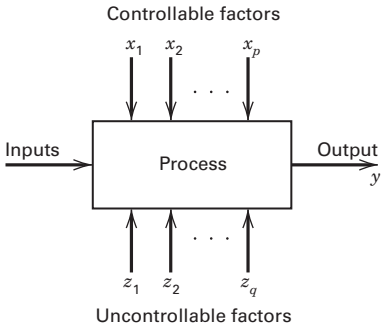


FIGURE 1.1 General Model of a System or Process

As you can see from the examples that we presented, experiments often involve several factors. The general approach to planning and conducting the experiment is called the **strategy of experimentation**. We will illustrate three strategies with an extension of the drug delivery pump experiment described earlier. The performance of this pump depends on more than the cylinder diameter. Suppose that the other factors include the length of the cylinder, the fit or clearance between the cylinder wall and the plunger, the plunger length, and the diameter of the tubing that connects the pump to the patient's blood vessel. This is a total of five design factors. There are other factors that could be considered, such as the choice of material for the tubing and the diameter of the needle that is inserted into the patient's vein, but the bioengineer decides that these factors can be ignored; that is, these factors are not important because their effects are likely to be so small that they have no practical value. Engineers, scientists, and business analysts often make these types of decisions about some of the factors they are considering.

Now let us consider how the five design factors that we have identified as potentially important could be experimentally tested to determine their effect on the drug delivery rate. Each trial of the experiment requires assembling and testing a prototype with a specific configuration of the five factors. Suppose that a maximum of eight prototypes can be tested over the course of the experiment. One approach would be to select an essentially arbitrary combination of these factors, test them, and see the result. Then, depending on the observed results, another combination of factors is selected, a second prototype assembled, and another test performed. This approach could be continued, switching the values of one or more design factors from test to test until all eight prototypes have been built and tested. This strategy of experimentation, called the **best-guess approach**, is frequently used in practice. It can work reasonably well, too, because the experimenters often have a great deal of technical (theoretical) knowledge or insight about the system they are studying, as well as considerable practical experience.

The best-guess approach has at least two disadvantages. First, suppose the initial "best guess" does not produce the desired results. Now the experimenter has to take another guess at the correct combination of factor levels. This could continue for a long time, without any guarantee of success. It is certainly possible that the entire budget for the experiment (eight prototypes) could be exhausted without finding a satisfactory solution. In fact, the optimal solution could be some set of factor values that lie in between the specific values of the factors that are being tested. Second, suppose the initial best guess (or one of the early ones) produces an acceptable result. Now the experimenter is tempted to stop testing, although there is no guarantee that the *best* solution has been found. A competitor that uses a better experimental approach can possibly develop a product that will outperform yours.

Another strategy of experimentation that is used extensively in practice is the **one-factor-at-a-time (OFAT)** approach. This method consists of selecting a starting point, or **baseline** set of levels, for each factor and then successively varying each factor over its range with the other factors held constant at the baseline level. After all tests are performed, a series of graphs are usually constructed showing how the response variable is affected by varying each factor with all other factors held constant.

To illustrate the OFAT approach, suppose that each one of the five factors in the drug delivery pump has two values (or levels) that we want to test. We will call these levels "low" and "high" and denote that by -1 and $+1$ respectively. Table 1.1 shows an OFAT experiment for this problem. The first thing we notice about the OFAT is that it only requires six runs. This is an apparent efficiency, as we were prepared to perform eight tests. But further examination of the experiment shows that the six runs are used very poorly. For example, only run 1 and run 2 tell you anything about changing the cylinder diameter. Furthermore, only run 1 and run 3 tell you anything about changing the cylinder length. In fact, the effect of changing any one variable is determined by using the results from only two of the six runs. This is very inefficient use of the data—every conclusion is based on only two tests even though six tests were performed. Furthermore, if there is some variability in the drug delivery rate observations (as there always will be), then this variability could have an effect on the conclusions.

Table 1.1 A One-factor-at-a-time Experiment for the Drug Delivery Pump

	Factor				
Run	**Cylinder diameter**	**Cylinder length**	**Fit**	**Plunger length**	**Tubing diameter**
1	−1	−1	−1	−1	−1
2	+1	−1	−1	−1	−1
3	−1	+1	−1	−1	−1
4	−1	−1	+1	−1	−1
5	−1	−1	−1	+1	−1
6	−1	−1	−1	−1	+1

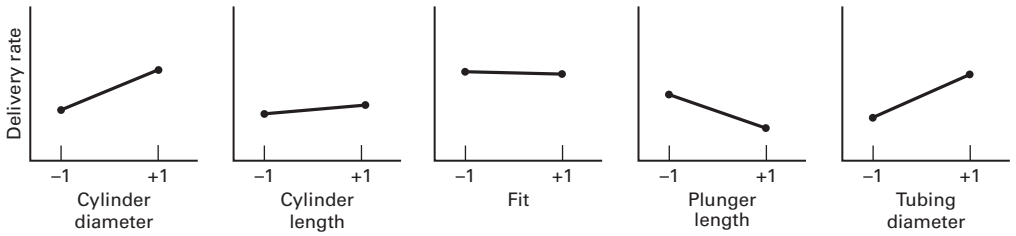


FIGURE 1.2 Results from the OFAT in Table 1.1

Figure 1.2 shows a set of hypothetical graphs that represent the results from the OFAT in Table 1.1. A straight line connects the observed response at the high and low levels of each factor. The slope of the line determines the effect of the factor; a positive slope means that increasing the level of the factor from low to high increases the drug delivery rate, while a slope that is zero (or nearly so) indicates the factor has no effect. Each one of the data values that determine the slope in each graph is the result of a single run.

Another major disadvantage of the OFAT strategy is that it fails to consider any possible **interaction** effect of the factors (an interaction is the failure of one factor to produce the same effect on the response at different levels of another factor). Figure 1.3 shows an interaction between the factors cylinder diameter and cylinder length for the drug delivery pump experiment. Notice that if the engineer uses the smaller diameter cylinder, the length of the cylinder has virtually no effect on the drug delivery rate, but if she uses the larger diameter cylinder, much different results are obtained by changing the cylinder length. Interactions between factors are very common, and if they occur, the OFAT strategy will usually produce poor results. Many engineers do not recognize this, and, consequently, OFAT experiments are run frequently in practice. (Some individuals actually think that this strategy is related to the scientific method or that it is a "sound" engineering principle.) OFAT experiments are always less efficient than other methods based on a statistical approach to design.

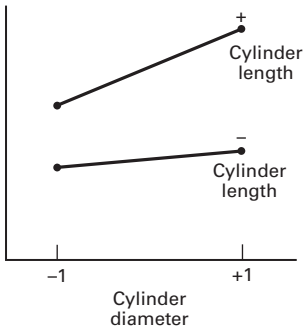


FIGURE 1.3 An Interaction Between Cylinder Diameter and Cylinder Length

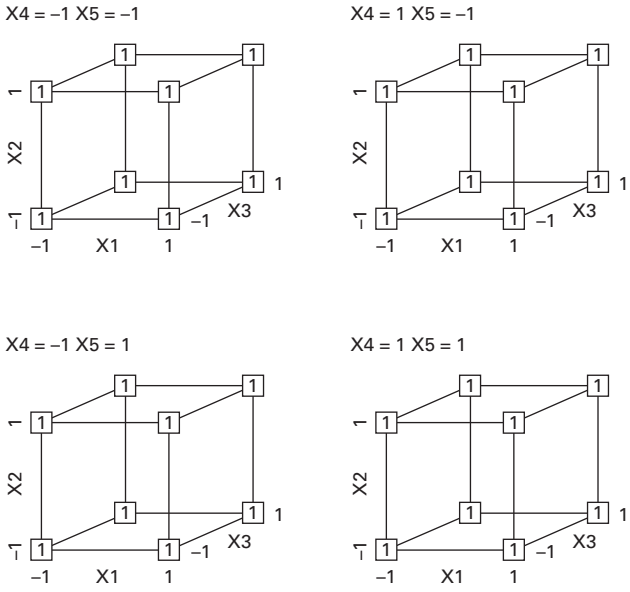


FIGURE 1.4 The 2^5 Full-factorial Design (the Symbol 1 at Each Cube Corner Indicates That a Run Should Be Made at That Location)

The correct approach to dealing with several factors is to conduct a **factorial** experiment. This is an experimental strategy in which factors are varied together, instead of one at a time. The factorial experimental design concept is extremely important, and we will illustrate how it can be applied to the drug delivery pump experiment. We will continue to assume that each factor has two levels, low and high, represented symbolically by -1 and $+1$, respectively. A complete or full-factorial experiment would require that we run every possible combination of the five factors across two levels each. This design has $2^5 = 32$ runs (see Figure 1.4). Because the experimenter only has a budget that will allow building eight prototypes, the full-factorial strategy will not work here. Fortunately, if there are four to five or more factors, a full factorial is usually not necessary and we can use a smaller fraction of the original runs. Figure 1.5 shows a one-half fraction of

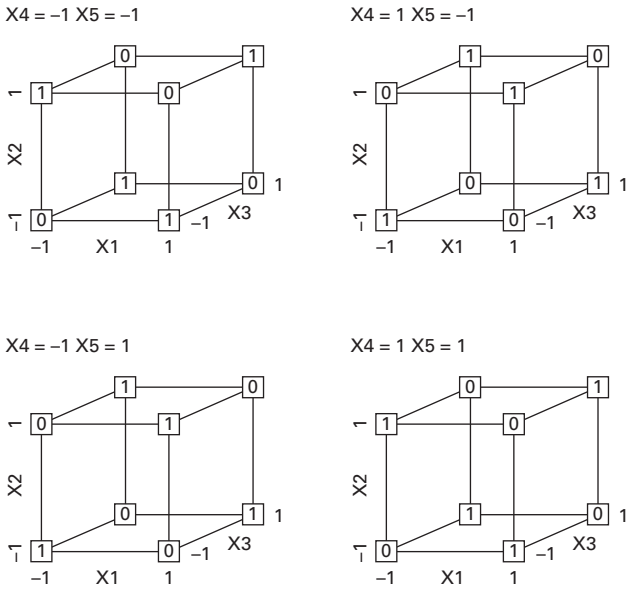


FIGURE 1.5 The One-half Fraction of the 2^5 Design (the Cells with 1 Indicate Where Runs Should Be Made)

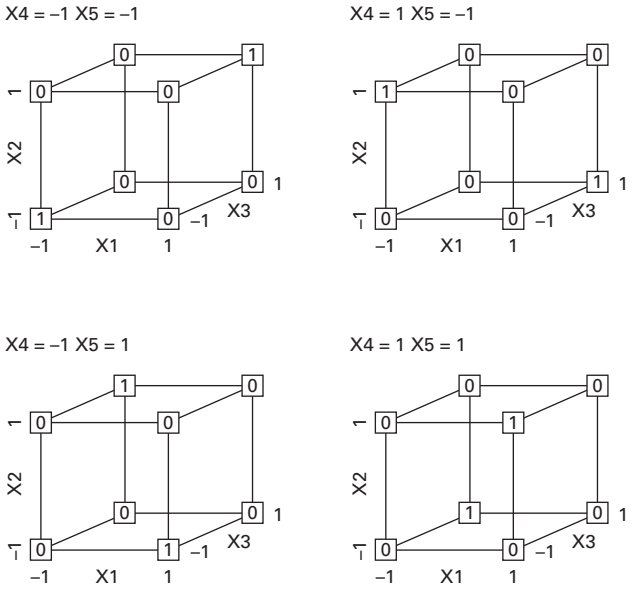


FIGURE 1.6 The One-quarter Fraction of the 2^5 Design (the Cells with 1 Indicate Where Runs Should Be Made)

the 2^5. This design has 16 runs, so it still is not feasible. This is an excellent design, however, because it allows the experimenter to estimate the effects of all five factors and all ten of the two-factor interactions involving these five factors. There is also an eight-run fraction of the 2^5, called a one-quarter fraction. This design is shown in Figure 1.6 and Table 1.2.

One very important feature of the factorial experiment is evident from this simple example; namely, factorials and fractional factorials make the most efficient use of the experimental data. Notice from Table 1.2 that this experiment has eight observations, and for every factor there are four observations on the drug delivery rate with the factor at the high level (+1) and four observations with the factor at the low level (−1). So if we take the difference in the averages of these groups of four observations as a measure of the effect of a factor on the drug delivery rate, every factor effect is estimated using all eight runs. No other strategy of experimentation makes such efficient use of the data. This is an important and useful feature of factorials.

Table 1.2 A One-quarter Fraction of the 2^5 Design for the Drug Delivery Pump Experiment

	Factor				
Run	Cylinder diameter	Cylinder length	Fit	Plunger length	Tubing diameter
1	−1	−1	−1	+1	+1
2	+1	−1	−1	−1	−1
3	−1	+1	−1	−1	+1
4	+1	+1	−1	+1	−1
5	−1	−1	+1	+1	−1
6	+1	−1	+1	−1	+1
7	−1	+1	+1	−1	−1
8	+1	+1	+1	+1	+1

1.2 Basic Principles

If an experiment is going to be performed most efficiently, a scientific approach to planning the experiment must be employed. **Statistical design of experiments** refers to the process of planning the experiment so that appropriate data will be collected and analyzed by statistical methods, resulting in valid and objective conclusions. A statistical approach to experimental design is necessary if we wish to draw meaningful conclusions from the data. When the problem involves data that are subject to experimental errors, statistical methods are the only **objective** approach to analysis. Thus, there are two aspects to any experimental problem: the design of the experiment and the statistical analysis of the data. These two aspects are closely related because the possibilities for analysis depend directly on the design employed, both of which are addressed in this book.

The four basic principles of experimental design are **the factorial principle**, **randomization**, **replication**, and **blocking**. We discussed the factorial principle in Section 1.1. The core idea here is that the most efficient way of acquiring information is to systematically vary multiple factors in a carefully constructed series of tests. Randomization is the cornerstone underlying the use of statistical methods in experimental design. By randomization, we mean that both the allocation of the experimental material and the order in which the individual runs or trials of the experiment are to be performed are randomly determined. Statistical methods require that the observations (or errors) be independently distributed random variables. Randomization usually makes this assumption valid. By properly randomizing the experiment, we also assist in "averaging out" the effects of extraneous factors that may be present. For example, suppose that the pressure in the drug delivery pump experiment increases throughout the day and the drug delivery rate may be affected by pump pressure. If all the smaller cylinder diameter prototypes are tested in the morning and all of the larger diameter cylinder prototypes are tested in the afternoon, we may be introducing systematic bias into the experimental results. This bias handicaps one of the cylinder diameters and consequently invalidates our results. Testing the prototypes in random order throughout the day is the solution to this problem.

Computer software programs are widely used to assist experimenters in selecting and constructing experimental designs. These programs often present the runs in the experimental design in random order. This random order is created by using a random number generator. Even with such a computer program, it is still often necessary to assign units of experimental material, operators, gauges or measurement devices, and so forth for use in the experiment. Sometimes, experimenters encounter situations where randomization of some aspect of the experiment is difficult. For example, in a chemical process, temperature may be a very hard-to-change variable as we may want to change it less often than we change the levels of other factors. In an experiment of this type, **complete randomization** would be difficult because it would add time and cost. There are statistical design methods for dealing with **restricted randomization**. Some of these approaches are discussed in subsequent chapters.

By **replication** we mean an independent repeat of each factor combination in the experiment. In the drug delivery pump experiment discussed in Section 1.1, replication would consist of testing a prototype with the smaller cylinder diameter and a prototype with the larger cylinder diameter. Thus, if four prototypes of each type are tested, we say the experiment has been replicated four times, or that four **replicates** have been obtained. Each of the eight total observations should be run in random order. Replication has two important properties. First, it allows the experimenter to obtain an estimate of the experimental error. This estimate of error becomes a basic unit of measurement for determining whether observed differences in the data are really **statistically significantly** different. Second, if the sample mean ($\bar{y}$) is used to estimate the true mean response for one of the factor levels in the experiment, replication permits the experimenter to obtain a more precise estimate of this parameter. For example, if σ^2 is the variance of an individual observation and there are n replicates, the variance of the sample mean is

$$\sigma^2_{\bar{y}} = \frac{\sigma^2}{n}$$

The practical implication of this is that replication improves the precision of estimation of the true mean of the response variable. For example, if we had only $n = 1$ replicate and observed $y_1 = 10.1$ ml/h for the smaller diameter cylinder and $y_2 = 10.3$ ml/h for the larger diameter cylinder, we would be unable to make a statistically sound decision about the effect of the cylinder diameter on the drug delivery rate—that is, the observed difference from this experiment could be the result of experimental error. The point is that without replication we have no way of knowing why the two observations are different. On the other hand, if n was reasonably large and the experimental error was sufficiently small and if we observed $\bar{y}_1 < \bar{y}_2$, we would be reasonably safe in concluding that the larger cylinder diameter results in a higher drug delivery rate than the smaller cylinder diameter.

Often when the runs in an experiment are randomized, two (or more) consecutive runs will have exactly the same levels for some of the factors. For example, suppose we have three factors in an experiment: pressure, temperature, and time. When the experimental runs are randomized, we find the following:

Run number	Pressure (psi)	Temperature (°C)	Time (min)
i	30	100	30
$i + 1$	30	125	45
$i + 2$	40	125	45

Notice that between runs i and $i + 1$, the levels of pressure are identical and between runs $i + 1$ and $i + 2$, the levels of both temperature and time are identical. To obtain a true replicate, the experimenter needs to "twist the pressure knob" to an intermediate setting between runs i and $i + 1$, and reset pressure to 30 psi for run $i + 1$. Similarly, temperature and time should be reset to intermediate levels between runs $i + 1$ and $i + 2$ before being set to their design levels for run $i + 2$. Part of the experimental error is the variability associated with hitting and holding factor levels.

There is an important distinction between **replication** and **repeated measurements**. For example, suppose that a silicon wafer is etched in a single-wafer plasma etching process, and a critical dimension on this wafer is measured three times. These measurements are not replicates; they are a form of repeated measurements, and in this case, the observed variability in the three repeated measurements is a direct reflection of the inherent variability in the measurement system or gauge. As another illustration, suppose that as part of an experiment in semiconductor manufacturing four wafers are processed simultaneously in an oxidation furnace at a particular gas flow rate and time and then a measurement is taken on the oxide thickness of each wafer. Once again, the measurements on the four wafers are not replicates but repeated measurements. In this case, they reflect differences among the wafers and other sources of variability within that particular furnace run. Replication reflects sources of variability both **between** runs and (potentially) **within** runs.

Blocking is a design technique used to improve the precision with which comparisons among the factors of interest are made. Often blocking is used to reduce or eliminate the variability transmitted from **nuisance factors**—that is, factors that may influence the experimental response but in which we are not directly interested. For example, an experiment in a pharmaceutical manufacturing may require two batches of the active drug to make all the required runs. However, there could be differences between the batches due to variability in the process that produces the active drug, and if we are not specifically interested in this effect, we would think of the batches of active drug as a nuisance factor. Generally, a block is a set of relatively homogeneous experimental conditions. In the pharmaceutical manufacturing example, each batch of the active drug would form a block, because the variability within a batch would be expected to be smaller than

the variability between batches. Typically, as in this example, each level of the nuisance factor becomes a block. Then the experimenter divides the observations from the statistical design into groups that are run in each block. We study blocking in detail in several places in the text.

The four basic principles of experimental design, the factorial principle, randomization, replication, and blocking, deserve consideration in every experiment. We illustrate and emphasize them repeatedly throughout this book.

1.3 Practical Guidelines for Designing an Experiment

To use the statistical approach in designing and analyzing an experiment, it is necessary for everyone involved in the experiment to have a clear idea in advance of exactly what is to be studied, how the data are to be collected, and at least a qualitative understanding of how these data are to be analyzed. An outline of the recommended procedure is given in Table 1.3. We now give a brief discussion of this outline and elaborate on some of the key points. For more information about practical aspects of design, see Coleman and Montgomery (1993) and the references therein.

1.3.1 Recognition of and Statement of the Problem

This may seem to be a rather obvious point, but in practice it is often neither simple to realize that a problem requiring experimentation exists nor is it easy to develop a clear and generally accepted statement of this problem. It is necessary to develop all ideas about the objectives of the experiment. Usually, it is important to solicit input from all concerned parties: engineering, quality assurance, manufacturing, marketing, management, operating personnel (who usually have much insight and who are too often ignored), and when possible the customer. For this reason, a **team approach** to designing experiments is recommended.

It is usually helpful to prepare a list of specific problems or questions that need to be addressed by the experiment. A clear statement of the problem often contributes substantially to better understanding of the phenomenon being studied and the final solution of the problem. It is also important to keep the overall objective in mind; for example, is this a new process or system—in which case the initial objective is likely to be **characterization** or **factor screening**—or is it a mature or reasonably well-understood system that has been previously characterized—in which case the objective might be **optimization**? There are many possible objectives of an experiment, including **confirmation** (is the system performing the same way now that it did in the past?), **discovery** (what happens if we explore new materials, variables, operating conditions, etc.?), and **stability** or **robustness** (under what conditions do the response variables of interest seriously

Table 1.3 Guidelines for Designing an Experiment

1. Recognition of and statement of the problem	Preexperimental planning
2. Selection of the response variable[a]	
3. Choice of factors, levels, and ranges[a]	
4. Choice of experimental design	
5. Performing the experiment	
6. Statistical analysis of the data	
7. Conclusions and recommendations	

[a] In practice, steps are often done simultaneously or in reverse order.

degrade? or, how can we reduce the variability in the response variable that arises from sources that we cannot directly control?). Obviously, the specific questions to be addressed in the experiment relate directly to the overall objectives. An important aspect of problem formulation is the recognition that one large comprehensive experiment may not answer the key questions satisfactorily. A single comprehensive experiment requires the experimenters know the answers to a lot of questions, and if they are wrong, the results may be disappointing. This leads to wasting time, materials, and other resources and may result in never answering the original research questions satisfactorily. A **sequential** approach employing a series of smaller experiments, each with a specific objective, such as factor screening, is a better strategy.

1.3.2 Selection of the Response Variable

In selecting the response variable, the experimenter should be certain that this variable really provides useful information about the process under study. Most often, the average or standard deviation (or both) of the measured characteristic will be the response variable. Multiple responses are not unusual. Gauge capability (or measurement error) is also an important factor. If gauge capability is inadequate, only relatively large factor effects will be detected by the experiment or perhaps additional replication will be required. In some situations where gauge capability is poor, the experimenter may decide to measure each experimental unit several times and use the average of the repeated measurements as the observed response. It is usually critically important to identify issues related to defining the responses of interest and how they are to be measured *before* conducting the experiment. Sometimes, designed experiments are employed to study and improve the performance of measurement systems. (As noted in Table 1.3, steps 2 and 3 are often done simultaneously or in the reverse order.)

1.3.3 Choice of Factors, Levels, and Ranges

When considering the factors that may influence the performance of a process or system, the experimenter usually discovers that these factors can be classified as either **potential design factors** or nuisance factors. The potential design factors are those factors that the experimenter may wish to vary in the experiment. Often we find that there are a lot of potential design factors, and some further classification of them is helpful. Some useful classifications are **design factors**, **held-constant factors**, and **allowed-to-vary** factors. The design factors are the factors actually selected for study in the experiment. Held-constant factors are variables that may exert some effect on the response, but for purposes of the present experiment these factors are not of interest, so they will be held at a specific level. For example, in an etching experiment in the semiconductor industry, there may be an effect that is unique to the specific plasma etch tool used in the experiment. However, this factor would be very difficult to vary in an experiment, so the experimenter may decide to perform all of the experimental runs on one particular (ideally "typical") etcher. Thus, this factor has been held constant. As an example of allowed-to-vary factors, the experimental units or the "materials" to which the design factors are applied are usually nonhomogeneous, yet we often ignore this unit-to-unit variability and rely on randomization to balance out any material or experimental unit effect. We often assume that the effects of held-constant factors and allowed-to-vary factors are relatively small.

Nuisance factors, on the other hand, may have large effects that must be accounted for, yet we may not be interested in them in the context of the present experiment. Nuisance factors are often classified as **controllable**, **uncontrollable**, or **noise factors**. A controllable nuisance factor is one whose levels may be set by the experimenter. For example, the experimenter can select different batches of raw material or different days of the week when conducting the experiment. The blocking principle, discussed in the previous section, is often useful in dealing with

controllable nuisance factors. If a nuisance factor is uncontrollable in the experiment, but it can be measured, an analysis procedure called the **analysis of covariance** can often be used to compensate for its effect. For example, the relative humidity in the process environment may affect process performance, and if the humidity cannot be controlled, it can be measured and treated as a covariate. When a factor that varies naturally and uncontrollably in the process can be controlled for purposes of an experiment, we often call it a noise factor. In these kinds of experimental situations, our objective is usually to find the settings of the controllable design factors that minimize the variability transmitted from the noise factors. This is sometimes called a process robustness study or a robust design problem. Blocking, analysis of covariance, and process robustness studies are discussed later in the text.

Once the experimenter has selected the design factors, he or she must choose the ranges over which these factors will be varied and the specific levels at which runs will be made. Thought must also be given to how these factors are to be controlled at the desired values and how they are to be measured. For instance, in the drug delivery pump experiment, the engineer has defined five factors that may affect the drug delivery rate. The experimenter will also have to decide on a region of interest for each factor (that is, the range over which each factor will be varied) and on how many levels of each variable to use. **Process knowledge** is required to do this. This process knowledge is usually a combination of practical experience and theoretical understanding. It is important to investigate all factors that may be of importance and to be not overly influenced by past experience, particularly when we are in the early stages of experimentation or when the process is not very mature.

When the objective of the experiment is **factor screening** or **process characterization**, it is usually best to keep the number of factor levels low. Often, two levels are adequate in factor screening studies. Choosing the region of interest is also important. In factor screening, the region of interest should be relatively large—that is, the range over which the factors are varied should be broad. As we learn more about which variables are important and which levels produce the best results, the region of interest will usually become narrower.

The **cause-and-effect diagram** can be a useful technique for organizing some of the information generated in preexperimental planning. Figure 1.7 is the cause-and-effect diagram constructed while planning an experiment to resolve problems with wafer charging (charge accumulation on the wafers) encountered in an etching tool used in semiconductor manufacturing. The cause-and-effect diagram is also known as a **fishbone diagram** because the "effect" of interest or the response variable is drawn along the spine of the diagram and the potential causes or design factors organized in a series of ribs. The cause-and-effect diagram uses the traditional

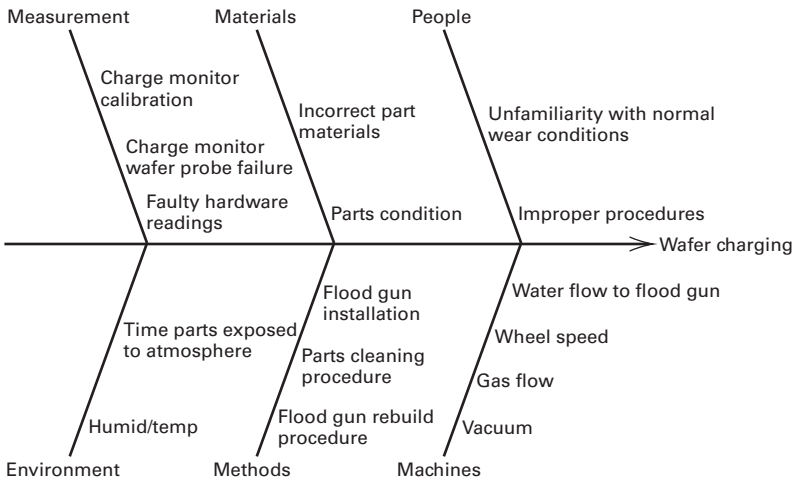


FIGURE 1.7 A Cause-and-effect Diagram for the Wafer Charging Experiment

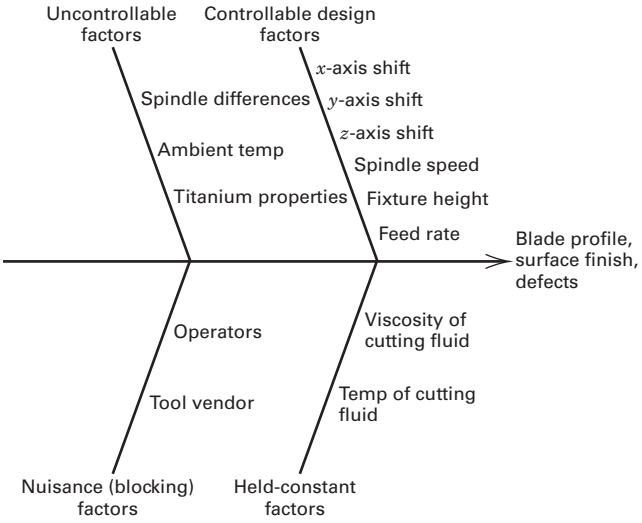


FIGURE 1.8 A Cause-and-effect Diagram for the CNC Experiment

causes of measurement, materials, people, environment, methods, and machines to organize the information and potential design factors. Notice that some of the individual causes will probably lead directly to a design factor that will be included in the experiment (such as wheel speed, gas flow, and vacuum), while others represent potential areas that will need further study to turn them into design factors (such as operators following improper procedures), and still others may lead to either factors that will be held constant during the experiment or blocked (such as temperature and relative humidity). Figure 1.8 is a cause-and-effect diagram for an experiment to study the effect of several factors on the turbine blades produced on a computer-numerical-controlled (CNC) machine. This experiment has three response variables: blade profile, blade surface finish, and surface finish defects in the finished blade. The causes are organized into groups of controllable factors from which the design factors for the experiment may be selected, uncontrollable factors whose effects will probably be balanced out by randomization, nuisance factors that may be blocked, and factors that may be held constant when the experiment is conducted. It is not unusual for experimenters to construct several different cause-and-effect diagrams to assist and guide them during preexperimental planning.

We reiterate how crucial it is to bring out all points of view and process information in steps 1 through 3. We refer to this as **preexperimental planning**. Coleman and Montgomery (1993) provided worksheets that can be useful in preexperimental planning. It is unlikely that one person has all the knowledge required to do this adequately in many situations. Therefore, we advocate for a team effort in planning the experiment. Most of your success will hinge on how well the preexperimental planning is done.

1.3.4 Experimental Design Generation

If the above preexperimental planning activities are done correctly, this step is relatively easy. Generation of the design considers several decision variables, including the sample size (number of replicates), selection of a suitable run order for the experimental trials, and determination of whether or not blocking or other randomization restrictions are involved. It also considers selecting a tentative **empirical model** to describe the results. The model is a quantitative relationship (equation) between the response and the important design factors. In many cases, a low-order polynomial model will be adequate. A **first-order** model in two variables is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where y is the response, the xs are the design factors, the βs are unknown parameters that will be estimated from the data in the experiment, and ε is a random error term that accounts for the experimental error in the system that is being studied. First-order models are used extensively in screening or characterization experiments. A common extension of the first-order model is to add an **interaction** term, say

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon$$

where the cross-product term $x_1 x_2$ represents the two-factor interaction involving the design factors. Because such interactions are relatively common, the first-order model with interaction is widely used. Higher-order interactions can also be included in experiments with more than two factors if necessary. Another widely used model is the **second-order** model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \varepsilon$$

Second-order models are often used in optimization experiments.

There are several interactive statistical software packages that support this phase of experimental design. The experimenter can enter information about the number of factors, levels, and ranges, and the model that is anticipated to be appropriate, and these programs will then allow the experimenter to generate a design that is, in some sense, the "best" design for the problem. The software will also provide diagnostic information about how the design will perform for the anticipated model. This allows the experimenter to generate more than one design for his or her problem and compare the design choices available. These programs will usually also provide a worksheet (with the order of the runs randomized) for use in conducting the experiment.

In selecting the design, it is important to keep the experimental objectives in mind. In many engineering experiments, we already know at the outset that some of the factor levels will result in different values for the response. Consequently, we are interested in identifying *which* factors cause this difference and in estimating the **magnitude** of the response change. In other situations, we may be more interested in verifying **uniformity**. For example, two production conditions A and B may be compared, A being the standard and B being a more cost-effective alternative. The experimenter will then be interested in demonstrating that using B does not lower the yield relative to the use of A.

1.3.5 Performing the Experiment

When running the experiment, it is vital to monitor the process carefully to ensure that everything is being done according to planning. Errors in experimental procedure at this stage will usually destroy experimental validity. Up-front planning is crucial to success. It is easy to underestimate the logistical and planning aspects of running a designed experiment in a complex manufacturing or research and development environment. Coleman and Montgomery (1993) suggested that prior to conducting the experiment a few trial runs or pilot runs are often helpful. These runs provide information about the consistency of experimental material, a check on the measurement system, a rough idea of experimental error, and a chance to practice the overall experimental technique. This also provides an opportunity to revisit the decisions made in steps 1–4, if necessary.

1.3.6 Statistical Analysis of the Data

Statistical methods should be used to analyze the data so that results and conclusions are **objective** rather than judgmental in nature. If the experiment has been designed correctly and performed according to the design, the statistical methods required are not elaborate. Software packages can assist in data analysis, and most of the programs used in step 4 to select the design provide a

seamless, direct interface to the statistical analysis. Often we find that simple **graphical methods** play an important role in data analysis and interpretation, and computer software plays an important role in visualizing the results of the experiment. Because many of the questions that the experimenter wants to answer can be cast into a hypothesis-testing framework, statistical hypothesis testing and confidence interval estimation procedures are very useful in analyzing the data from a designed experiment. Fitting the **empirical model** for the experiment is also important. Residual analysis and model adequacy checking are also well-advised analysis techniques. We discuss these issues in detail later. Remember that statistical methods cannot prove that a factor (or factors) has a particular effect. They only provide guidelines as to the reliability and validity of results. When properly applied, statistical methods do not allow anything to be proved experimentally, but they do allow us to measure the likely error or risk in a conclusion. The primary advantage of statistical methods is that they add **objectivity** to the decision-making process. Statistical techniques coupled with good engineering or process knowledge and common sense generally lead to sound conclusions.

1.3.7 Conclusions and Recommendations

Once the data have been analyzed, the experimenter must draw *practical* conclusions about the results and recommend a course of action. Graphical methods are often useful in this stage, particularly in presenting the results to others. **Follow-up runs** and **confirmation testing** should also be performed to validate the conclusions from the experiment. Throughout this entire process, it is important to keep in mind that experimentation is an important part of the learning process, where we tentatively formulate hypotheses about a system, perform experiments to investigate these hypotheses, and on the basis of the results formulate new hypotheses, and so on. This suggests that experimentation is **iterative**. It is often a major mistake to design a single, large, comprehensive experiment at the start of a study. A successful experiment requires knowledge of the important factors, the ranges over which these factors should be varied, the appropriate number of levels to use, the proper units of measurement for these variables, and the correct form of the model for the results.

Generally, we do not perfectly know the answers to these questions, but we learn about them as we go along. As an experimental program progresses, we often drop some input variables, add others, change the region of exploration for some factors, or add new response variables. Consequently, it is desirable to experiment **sequentially**. This will ensure that sufficient resources are available to perform confirmation runs and ultimately accomplish the final objective of the experiment.

1.4 A Brief History of Designed Experiments

There have been four eras in the modern development of statistical experimental design. The agricultural era was led by the pioneering work of Sir Ronald A. Fisher in the 1920s and early 1930s. During that time, Fisher was responsible for statistics and data analysis at the Rothamsted Agricultural Experimental Station near London, England. Fisher recognized that a poor method for generating data often hampered the data analysis. By interacting with scientists and researchers in many fields, he developed insights that led to the four basic principles of experimental design that we discussed in Section 1.2: the factorial principle, randomization, replication, and blocking. Fisher systematically introduced statistical thinking and principles into designing experimental investigations, including the likelihood concept and the analysis of variance. His two books (the most recent editions are Fisher (1958, 1966)) had profound influence on the use of statistics, particularly in agricultural and related life sciences. It is not an overstatement to say that Fisher's work transformed agricultural science. For an excellent biography of Fisher, see Box (1978).

Although applications of statistical design in industrial settings certainly began in the 1930s, the second, or industrial, era was catalyzed by the development of response surface methodology (RSM) by Box and Wilson (1951). They recognized and exploited the fact that many industrial experiments are fundamentally different from their agricultural counterparts in two ways: (1) the response variable can usually be observed (nearly) immediately and (2) the experimenter can quickly learn crucial information from a small group of runs that can be used to plan the next experiment. Box (1999) calls these two features of industrial experiments **immediacy** and **sequentiality**. Over the next 30 years, RSM and other design techniques spread throughout the chemical and the process industries, mostly in research and development work. George Box was the intellectual leader of this movement. However, the application of statistical design at the plant or manufacturing process level was still not extremely widespread. Some of the reasons for this include an inadequate training in basic statistical concepts and methods for engineers, scientists, and business personnel, the lack of computing resources and user-friendly software to support its implementation.

It was during the second industrial era that work on **optimal** design of experiments began. Kiefer (1959, 1961) and Kiefer and Wolfowitz (1959) proposed a formal approach to selecting a design based on specific objective optimality criteria. Their initial approach was to select a design that would result in the model parameters being estimated with the best possible precision. This approach did not find much application because of the lack of computer tools for its implementation. However, there have been great advances in both algorithms for generating optimal designs and computing capability over the last 20 years.

The increasing interest of Western industry in quality improvement that began in the late 1970s ushered in the third era of statistical design. The work of Genichi Taguchi (Taguchi and Wu (1980), Kackar (1985), and Taguchi (1987, 1991)) had a significant impact on expanding the interest in and the use of designed experiments. Taguchi advocated using designed experiments for what he termed **robust parameter design**, or

1. Making processes insensitive to environmental factors or other factors that are difficult to control
2. Making products insensitive to variation transmitted from components
3. Finding levels of the process variables that force the mean to a desired value while reducing variability around this value.

Taguchi suggested highly fractionated factorial designs and other orthogonal arrays along with some novel statistical methods to solve these problems. The resulting methodology generated much discussion and controversy. Part of the controversy arose because Taguchi's methodology was advocated in the West initially (and primarily) by entrepreneurs, and the underlying statistical science had not been adequately peer reviewed. By the late 1980s, the results of peer review indicated that although Taguchi's engineering concepts and objectives were well founded, there were substantial problems with his experimental strategy and methods of data analysis. For specific details of these issues, see Box (1988), Box, Bisgaard, and Fung (1988), Hunter (1985, 1989), Pignatiello and Ramberg (1992), and Myers, Montgomery, and Anderson-Cook (2016). Many of these concerns are also summarized in the extensive panel discussion in the May 1992 issue of *Technometrics* (see Nair et al. (1992)).

There were several positive outcomes of the Taguchi controversy. First, designed experiments became more widely used in the discrete parts of industries, including automotive and aerospace manufacturing, electronics and semiconductors, and many other industries that had previously made little use of the technique. It was also one of the drivers of the fourth era of statistical design, because not only did applications of designed experiments increase but also these applications spawned new research problems and a lot of highly useful new methodology appeared. An important aspect of this new methodology was renewed interest in optimal design. Because

computers were more powerful and widely available, algorithms to construct optimal designs were developed and began to be implemented in commercial software.

This fourth era of statistical design has included an expanded general interest in optimal design by researchers and software developers, as well as practitioners. Today the optimal design approach is a viable and attractive way to approach experimental design problems in the industrial and business world. We use the optimal design approach extensively throughout this book. This era has also produced other new methodology, including alternatives to Taguchi's technical methods that allow his engineering concepts to be carried into practice efficiently and effectively. Formal education in statistical experimental design is becoming part of many engineering programs in universities, at both undergraduate and graduate levels. More students in business are being exposed to the concepts as well. The successful integration of good experimental design practice into engineering, science, and business is a key factor in future competitiveness. Applications of designed experiments have grown far beyond the agricultural origins. There is not a single area of science, engineering, or business that has not successfully employed statistically designed experiments. There has also been considerable utilization of designed experiments in many other areas, including the service sector of business, financial services, government operations, e-commerce, and many nonprofit business sectors. An article appeared in *Forbes* magazine on March 11, 1996, entitled The New Mantra: MVT," where MVT stands for "multivariable testing," a term that the authors use to describe factorial designs. The article describes the many successes that a diverse group of companies have had through their use of statistically designed experiments.

1.5 A Review: Using Statistical Techniques in Experimentation

Much of the research in engineering, science, business, and industry is empirical and makes extensive use of experimentation. Statistical methods can greatly increase the efficiency of these experiments and often strengthen the conclusions so obtained. The proper use of statistical techniques in experimentation requires that the experimenter keep the following points in mind:

1. **Use your nonstatistical knowledge of the problem.** Experimenters are usually highly knowledgeable in their fields. For example, a civil engineer working on a problem in hydrology typically has considerable practical experience and formal academic training in this area, and a marketing research analyst has a lot of knowledge about the specific business problem under investigation. In some fields, there is a large body of physical theory from which relationships between factors and responses are drawn. This type of nonstatistical knowledge is invaluable in choosing factors, determining factor levels, deciding how many replicates to run, interpreting the results of the analysis, and so forth. Using a designed experiment is no substitute for thinking about the problem.

2. **Keep the design and analysis as simple as possible.** Don't be overzealous in the use of complex, sophisticated statistical techniques. Relatively simple design and analysis methods are adequate and easier to explain. This is a good place to reemphasize steps 1–3 of the procedure recommended in Section 1.3. If you do the preexperimental planning carefully and select a reasonable design, the analysis will almost always be relatively straightforward. In fact, a well-designed experiment will sometimes almost analyze itself! However, if you botch the preexperimental planning and execute the experimental design badly, it is unlikely that even the most complex and elegant statistics can save the day.

3. **Recognize the difference between practical and statistical significance.** Just because two experimental conditions produce mean responses that are statistically significantly different, there is no assurance that this difference is large enough to have any practical value.

For example, an engineer may determine that a modification to an automobile fuel injection system may produce a true mean improvement in gasoline mileage of 0.1 mi/gal and be able to determine that this is a statistically significant result. However, if the cost of the modification is $1000, the 0.1 mi/gal difference will be too small to be of any practical value.

4. **Experiments are usually iterative.** Remember that in most situations it is unwise to design too comprehensive an experiment at the start of a study. Successful design requires the knowledge of important factors, the ranges over which these factors are varied, the appropriate number of levels for each factor, and the proper methods and units of measurement for each factor and response. Generally, we are not well equipped to answer these questions at the beginning of the experiment, but we learn the answers as we go along. This argues in favor of the **iterative**, or **sequential**, approach discussed previously. Of course, there are situations where comprehensive experiments are entirely appropriate, but as a general rule most experiments should be iterative. Often these first efforts are primarily learning experiences, and some resources must be available to accomplish the final objectives of the experiment.

1.6 Review of Some Basic Statistical Concepts and Methods

1.6.1 Data Description

Statistics is the **science of data**. Collecting, describing, displaying, analyzing, and drawing conclusions from data are some of the many aspects of statistics. As an example, the United States Golf Association tests golf balls to ensure that they conform to the rules of golf. Balls are tested for weight, diameter, roundness, and conformance to an overall distance standard. The overall distance test is conducted by hitting balls with a driver swung by a mechanical device nicknamed Iron Byron, after the legendary great player Byron Nelson, whose swing the machine is said to emulate. The distances in yards achieved for 10 balls of a particular brand are as follows: 292, 270, 271, 278, 279, 268, 284, 271, 262, and 283. These distances are the total of the carry in the air and the distance the ball bounces or rolls after contact with the ground. We think of these observations as a **sample** of the possible distances that golf balls of this particular type could possibly travel. Much of statistics is involved with drawing conclusions about a population from sample data that is collected from that population.

Notice that the individual observations in the sample differ, so there is fluctuation, or **noise**, in the observed distances. This noise is usually called **experimental error** or simply **error**. It is a statistical error, meaning that it arises from variation that is uncontrolled and generally unavoidable. In this case, the error arises from ball-to-ball differences, differences in Iron Byron's swing from shot to shot, the differences in performance of the golf club from shot to shot (such as the flex in the shaft or the point of contact of the ball on the club face), the point at which the ball contacts the ground, which impacts the amount of roll and other factors. The presence of error or noise implies that the response variable, overall distance, is a **random variable**. A random variable may be either discrete or continuous. If the set of all possible values of the random variable is either finite or countably infinite, then the random variable is discrete, whereas if the set of all possible values of the random variable is an interval, then the random variable is continuous.

1.6.1.1 Graphical and Numerical Description of Variability

We often use simple graphical methods to assist in analyzing the data from an experiment. The dot diagram, illustrated in Figure 1.9 for the golf ball distance data, is a very useful device for displaying a small body of data (say up to about 20 observations). The dot diagram enables the

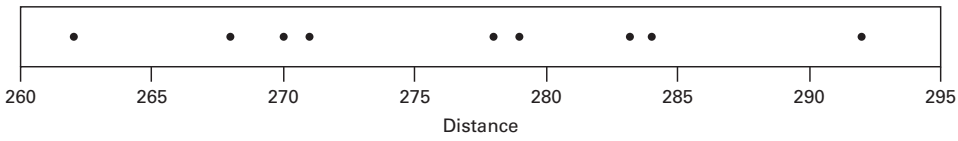


FIGURE 1.9
A Dot Diagram of the Golf Ball Distance Data

experimenter to see quickly the general location or central tendency of the observations and their spread. For example, in the golf ball overall distance data, the dot diagram reveals that the center of the data is about 275 yards, but that there is a lot of variability, with some balls traveling as little as 262 yards and some balls traveling as far as 292 yards, a range of 30 yards.

We can also describe data numerically. The **sample average** or **sample mean** is a very good way to measure the central tendency in sample data. The sample mean is defined as

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} \tag{1.1}$$

where y_i are the observations in the sample and n is the sample size. The sample average for the golf ball distance data is

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{292 + 270 + 271 + \cdots + 283}{10} = \frac{2758}{10} = 275.8 \text{ yards}$$

The sample average is a measure of the middle of the data in the sense that if you view the dot diagram as a system of unit weights the diagram balances exactly at the sample average. It is analogous to the first moment in mechanics.

Variability is usually measured by the **sample variance** or **sample standard deviation**. The sample variance is

$$S^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} \tag{1.2}$$

A large sample variance implies large variability in the sample data. The sample standard deviation S is the positive square root of the sample variance. Many individuals prefer to work with the standard deviation because it is expressed in the same units as the observations y (yards in the golf ball distance data), while the sample variance is in the square of the original units. The sample variance and standard deviation for the golf ball distance data are

$$S^2 = \frac{\sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}}{n-1} = \frac{761,384.0 - \frac{(2758)^2}{10}}{9} = 80.84$$

$$S = 8.99 \text{ yards}$$

If the data are fairly numerous, the dots in a dot diagram become difficult to distinguish and a histogram may be preferable. Figure 1.10 presents a histogram for 200 observations on the metal recovery, or yield, from a smelting process. The histogram shows the central tendency, spread, and

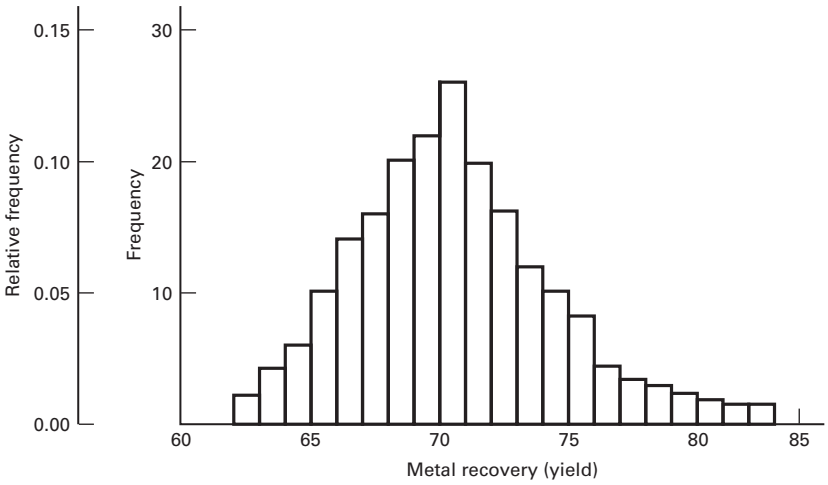


FIGURE 1.10 Histogram of 200 Observations on Metal Recovery (Yield) from a Smelting Process

general shape of the distribution of the data. Recall that a histogram is constructed by dividing the horizontal axis into bins (usually of equal length) and drawing a rectangle over the jth bin with the area of the rectangle proportional to n_j, the number of observations that fall in that bin. If the bins are of equal length, then the height of the rectangle will equal the number of sample observations in that bin. Histograms are large-sample tools and should generally not be used unless there are at least 30 or more observations.

The **box plot** (or **box-and-whisker plot**) is a useful way to display data. A box plot displays the minimum, the maximum, the lower and upper quartiles (the 25th percentile and the 75th percentile, respectively), and the median (the 50th percentile) on a rectangular box aligned either horizontally or vertically. The box extends from the lower quartile to the upper quartile, and a line is drawn through the box at the median. Lines (or whiskers) extend from the ends of the box to (typically) the minimum and maximum values. (There are several variations of box plots that have different rules for denoting the extreme sample points; see Montgomery and Runger (2018) for more details.) Figure 1.11 presents the box plot for the golf ball distance data. The median distance is 274.5 yards, which is very similar to the sample average (275.8 yards), and

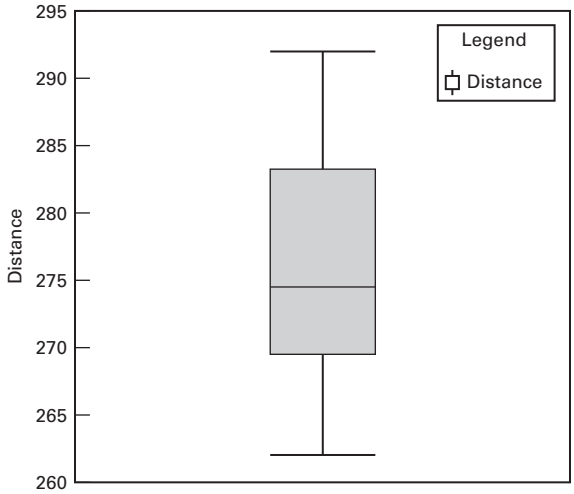


FIGURE 1.11 Box Plot of the Golf Ball Distance Data

both whiskers are similar in length. These are indications that the sample of data we have collected has a reasonably symmetric shape.

Dot diagrams, histograms, and box plots are useful graphical tools for summarizing the information in a sample of data, and the sample mean and sample variance or standard deviation are widely used numerical descriptive summaries. To describe the observations that might occur in a sample more completely, we can also use the concept of a probability distribution.

1.6.1.2 Probability Distributions

A random variable, say y, is characterized by its probability distribution. If y is discrete, we often call the probability distribution of y, say $p(y)$, the probability mass function of y. If y is continuous, the probability distribution of y, say $f(y)$, is often called the probability density function for y.

Figure 1.12 illustrates hypothetical discrete and continuous probability distributions.

Notice that in the discrete probability distribution (Figure 1.12a), it is the height of the function $p(y_j)$ that represents probability, whereas in the continuous case (Figure 1.12b), it is the area under the curve $f(y)$ associated with a given interval that represents probability. The properties of probability distributions may be summarized quantitatively as follows:

y discrete:

$$0 \le p(y_j) \le 1 \qquad \text{all values of } y_j$$

$$P(y = y_j) = p(y_j) \qquad \text{all values of } y_j$$

$$\sum_{\substack{\text{all values} \\ \text{of } y_j}} p(y_j) = 1$$

y continuous:

$$0 \le f(y)$$

$$P(a \le y \le b) = \int_a^b f(y)\, dy$$

$$\int_{-\infty}^{\infty} f(y)\, dy = 1$$

1.6.1.3 Mean, Variance, and Expected Values

The **mean**, μ, of a probability distribution is a measure of its central tendency or location. Mathematically, we define the mean as

$$\mu = \begin{cases} \int_{-\infty}^{\infty} yf(y)\, dy & y \text{ continuous} \\ \sum_{\text{all } y} yp(y_j) & y \text{ discrete} \end{cases} \tag{1.3}$$

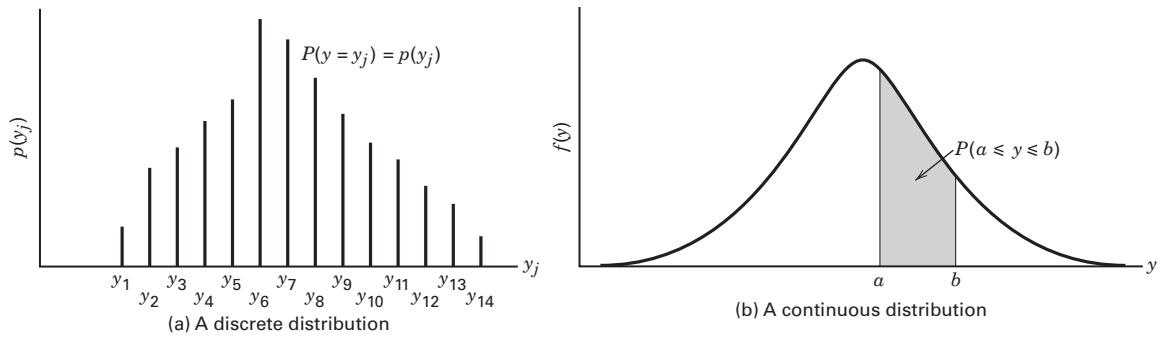


FIGURE 1.12 (a) Discrete and (b) Continuous Probability Distributions

We may also express the mean in terms of the **expected value** or the long-run average value of the random variable y as

$$\mu = E(y) = \begin{cases} \int_{-\infty}^{\infty} yf(y)\,dy & y \text{ continuous} \\ \sum_{\text{all } y} yp(y_j) & y \text{ discrete} \end{cases} \tag{1.4}$$

where E denotes the **expected value operator**.

The variability or dispersion of a probability distribution can be measured by the **variance**, defined as

$$\sigma^2 = \begin{cases} \int_{-\infty}^{\infty} (y-\mu)^2 f(y)\,dy & y \text{ continuous} \\ \sum_{\text{all } y} (y-\mu)^2 p(y_j) & y \text{ discrete} \end{cases} \tag{1.5}$$

Note that the variance can be expressed entirely in terms of expectation because

$$\sigma^2 = E[(y-\mu)^2] \tag{1.6}$$

Finally, the variance is used so extensively that it is convenient to define a **variance operator** V such that

$$V(y) = E[(y-\mu)^2] = \sigma^2 \tag{1.7}$$

The concepts of expected value and variance are used extensively throughout this book, and it may be helpful to review several elementary results concerning these operators. If y is a random variable with mean μ and variance σ^2 and c is a constant, then

1. $E(c) = c$
2. $E(y) = \mu$
3. $E(cy) = cE(y) = c\mu$
4. $V(c) = 0$
5. $V(y) = \sigma^2$
6. $V(cy) = c^2V(y) = c^2\sigma^2$

If there are two random variables, say, y_1 with $E(y_1) = \mu_1$ and $V(y_1) = \sigma_1^2$ and y_2 with $E(y_2) = \mu_2$ and $V(y_2) = \sigma_2^2$, we have

7. $E(y_1 + y_2) = E(y_1) + E(y_2) = \mu_1 + \mu_2$

It is possible to show that

8. $V(y_1 + y_2) = V(y_1) + V(y_2) + 2\text{Cov}(y_1, y_2)$

where

$$\text{Cov}(y_1, y_2) = E[(y_1 - \mu_1)(y_2 - \mu_2)] \tag{1.8}$$

is the **covariance** of the random variables y_1 and y_2. The covariance is a measure of the linear association between y_1 and y_2. More specifically, we may show that if y_1 and y_2 are independent, then $\text{Cov}(y_1, y_2) = 0$. We may also show that

9. $V(y_1 - y_2) = V(y_1) + V(y_2) - 2\text{Cov}(y_1, y_2)$

If y_1 and y_2 are **independent**, we have

10. $V(y_1 \pm y_2) = V(y_1) + V(y_2) = \sigma_1^2 + \sigma_2^2$

and

11. $E(y_1 \cdot y_2) = E(y_1) \cdot E(y_2) = \mu_1 \cdot \mu_2$

However, note that, in general

12. $E\left(\frac{y_1}{y_2}\right) \neq \frac{E(y_1)}{E(y_2)}$

regardless of whether or not y_1 and y_2 are independent.

1.6.2 Random Samples, Statistics and Sampling Distributions

The objective of statistical inference is to draw conclusions about a population using a sample from that population. Most of the methods that we will study assume that **random samples** are used. That is, if the population contains N elements and a sample of n of them is to be selected, and if each of the $N!/[(N-n)!n!]$ possible samples has an equal probability of being chosen, then the procedure employed is called **random sampling**. A key feature of a random sample is that the observations in the random sample are independent random variables. In practice, it is sometimes difficult to obtain random samples, and random numbers generated by a computer program may be helpful.

Statistical inference makes considerable use of quantities computed from the observations in the sample. We define a **statistic** as any function of the observations in a sample that does not contain unknown parameters. For example, the sample mean $\bar{y}$ and the sample variance S^2 are both statistics. The sample mean is a **point estimator** of the population mean μ, and the sample variance S^2 is a **point estimator** of the population variance σ^2. In general, an estimator of an unknown parameter is a statistic that corresponds to that parameter. Note that a point estimator is a random variable. A particular numerical value of an estimator, computed from sample data, is called an **estimate**. For example, suppose we wish to estimate the mean and variance of the breaking strength of a particular type of textile fiber. A random sample of $n = 25$ fiber specimens is tested, and the breaking strength is recorded for each. The sample mean and variance are computed according to Eqs. (1.1) and (1.2), respectively, and are $\bar{y} = 21.7$ and $S^2 = 1.20$. Therefore, the estimate of μ is 21.7, and the estimate of σ^2 is 1.20.

We know that the variance of the sample mean is $\sigma_{\bar{y}}^2 = \sigma^2/n$. The square root of this quantity, $\sigma_{\bar{y}} = \sigma/\sqrt{n}$, is called the **standard error** of the mean. A small value of the standard error indicates that the mean has been estimated with high precision. When the population variance is unknown, we can replace σ by the sample standard deviation S, and $\hat{\sigma}_{\bar{y}} = S/\sqrt{n}$ is called the **estimated standard error**.

Several properties are required of good point estimators. Two of the most important are the following:

1. An **unbiased** point estimator is often desirable. That is, the long-run average or expected value of the point estimator should be equal to the parameter that is being estimated.

   Although unbiasedness is desirable, this property alone does not always make an estimator a good one.

2. An unbiased estimator should have minimum variance. This property states that the minimum variance point estimator has a variance that is smaller than the variance of any other estimator of that parameter.

We may easily show that $\bar{y}$ and S^2 are unbiased estimators of μ and σ^2, respectively. First consider $\bar{y}$. Using the properties of expectation, we have

$$\begin{aligned} E(\bar{y}) &= E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(y_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned}$$

because the expected value of each observation y_i is μ. Thus, $\bar{y}$ is an unbiased estimator of μ.

Now consider the sample variance S^2. We have

$$\begin{aligned} E(S^2) &= E\left[\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}\right] \\ &= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right] \\ &= \frac{1}{n-1} E(SS) \end{aligned}$$

where $SS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ is the **corrected sum of squares** of the observations y_i. Now

$$\begin{aligned} E(SS) &= E\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right] \\ &= E\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right] \end{aligned} \tag{1.9}$$

$$\begin{aligned} &= \sum_{i=1}^{n} (\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) \\ &= (n-1)\sigma^2 \end{aligned} \tag{1.10}$$

Therefore,

$$E(S^2) = \frac{1}{n-1} E(SS) = \sigma^2$$

Therefore, S^2 is an unbiased estimator of σ^2.

1.6.2.1 Degrees of Freedom

The quantity $n-1$ in Eq. (1.10) is called the number of degrees of freedom of the sum of squares SS. This is a very general result; that is, if y is a random variable with variance σ^2 and $SS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ has ν degrees of freedom, then

$$E\left(\frac{SS}{\nu}\right) = \sigma^2 \tag{1.11}$$

The number of degrees of freedom of a sum of squares is equal to the number of independent elements in that sum of squares. For example, SS in Eq. (1.9) consists of the sum of squares of the n elements $y_1, y_2, \ldots, y_n$. These elements are not all independent because $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$; in fact, only $n-1$ of them are independent, implying that SS has $n-1$ degrees of freedom.

1.6.2.2 The Normal and Other Sampling Distributions

Often we are able to determine the probability distribution of a particular statistic if we know the probability distribution of the population from which the sample was drawn. The probability distribution of a statistic is called a sampling distribution. We now briefly discuss several useful sampling distributions.

One of the most commonly used sampling distributions is the normal distribution. If y is a normal random variable, the probability distribution of y is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)[(y-\mu)/\sigma]^2} \quad -\infty < y < \infty \tag{1.12}$$

where $-\infty < \mu < \infty$ is the mean of the distribution and $\sigma^2 > 0$ is the variance. The normal distribution is shown in Figure 1.13.

Because sample runs that differ as a result of experimental error often are well described by the normal distribution, it plays a central role in the analysis of data from designed experiments. Many important sampling distributions may also be defined in terms of normal random variables. We often use the notation $y \sim N(\mu, \sigma^2)$ to denote that y is distributed normally with mean μ and variance σ^2.

An important special case of the normal distribution is the **standard normal distribution**; that is, $\mu = 0$ and $\sigma^2 = 1$. We see that if $y \sim N(\mu, \sigma^2)$, the random variable

$$z = \frac{y - \mu}{\sigma} \tag{1.13}$$

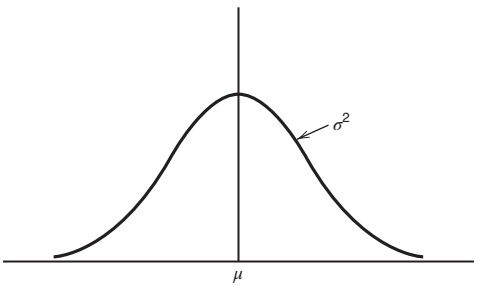


FIGURE 1.13 The Normal Distribution

follows the standard normal distribution, denoted $z \sim N(0, 1)$. The operation demonstrated in Eq. (1.13) is often called **standardizing** the normal random variable. It is often useful to be able to calculate probabilities associated with the standard normal random variable.

Many statistical techniques assume that the random variable is normally distributed. The central limit theorem is often a justification of approximate normality.

The Central Limit Theorem

If $y_1, y_2, \ldots, y_n$ is a sequence of n independent and identically distributed random variables with $E(y_i) = \mu$ and $V(y_i) = \sigma^2$ (both finite) and $x = y_1 + y_2 + \cdots + y_n$, then the limiting form of the distribution of

$$z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}}$$

as $n \to \infty$, is the standard normal distribution.

This result states essentially that the sum of n independent and identically distributed random variables is approximately normally distributed. In many cases, this approximation is good for very small n, say $n < 10$, whereas in other cases large n is required, say $n > 100$. Frequently, we think of the error in an experiment as arising in an additive manner from several independent sources; consequently, the normal distribution becomes a plausible model for the combined experimental error.

An important sampling distribution that can be defined in terms of normal random variables is the **chi-square** or χ^2 **distribution**. If $z_1, z_2, \ldots, z_k$ are normally and independently distributed random variables with mean 0 and variance 1, abbreviated NID(0, 1), then the random variable

$$x = z_1^2 + z_2^2 + \cdots + z_k^2$$

follows the **chi-square distribution with k degrees of freedom**. The density function of chi-square is

$$f(x) = \frac{1}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} x^{(k/2)-1} e^{-x/2} \quad x > 0 \tag{1.14}$$

Several chi-square distributions are shown in Figure 1.14. The distribution is asymmetric, or **skewed**, with the mean and variance

$$\mu = k$$
$$\sigma^2 = 2k$$

respectively. The appendix provides JMP scripting language (JSL) commands for calculating the percentage points of the chi-square distribution.

As an example of a random variable that follows the chi-square distribution, suppose that $y_1, y_2, \ldots, y_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution. Then

$$\frac{SS}{\sigma^2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sigma^2} \sim \chi^2_{n-1} \tag{1.15}$$

That is, SS/σ^2 is distributed as chi-square with $n - 1$ degrees of freedom.

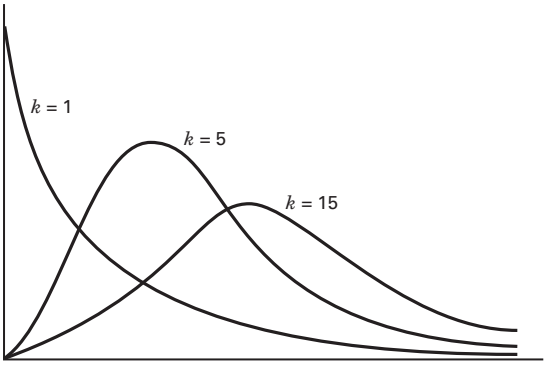


FIGURE 1.14 The Chi-square Distribution for Several Degrees of Freedom

Many of the techniques used in this book involve the computation and manipulation of sums of squares. The result given in Eq. (1.15) is extremely important and occurs repeatedly; a sum of squares in normal random variables when divided by σ^2 follows the chi-square distribution.

The sample variance can be written as

$$S^2 = \frac{SS}{n-1} \tag{1.16}$$

If the observations in the sample are NID(μ, σ^2), then the distribution of S^2 is $[\sigma^2/(n-1)]\chi^2_{n-1}$. Thus, the sampling distribution of the sample variance is a constant times the chi-square distribution if the population is normally distributed.

If z and χ^2_k are independent standard normal and chi-square random variables, respectively, the random variable

$$t_k = \frac{z}{\sqrt{\chi^2_k/k}} \tag{1.17}$$

follows the ***t*-distribution with *k* degrees of freedom**, denoted t_k. The density function of t is

$$f(t) = \frac{\Gamma[(k+1)/2]}{\sqrt{k\pi}\,\Gamma(k/2)} \frac{1}{[(t^2/k)+1]^{(k+1)/2}} \qquad -\infty < t < \infty \tag{1.18}$$

and the mean and variance of t are $\mu = 0$ and $\sigma^2 = k/(k-2)$ for $k > 2$, respectively. Several t-distributions are shown in Figure 1.15. Note that if $k = \infty$, the t-distribution becomes the standard normal distribution. The appendix provides JSL commands for calculating the percentage points of the t-distribution.

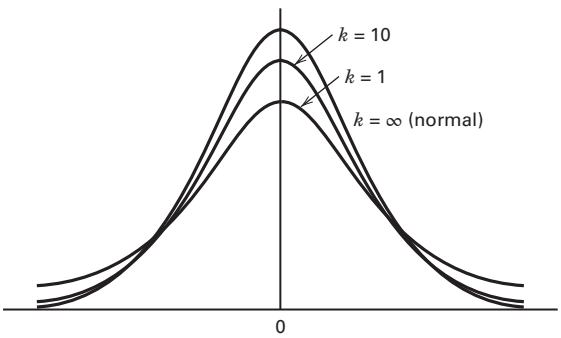


FIGURE 1.15 The *t*-distribution for Several Different Degrees of Freedom

If $y_1, y_2, \ldots, y_n$ is a random sample from the $N(\mu, \sigma^2)$ distribution, then the quantity

$$t = \frac{\bar{y} - \mu}{S/\sqrt{n}} \tag{1.19}$$

is distributed as t with $n - 1$ degrees of freedom.

The final sampling distribution that we will consider is the ***F*-distribution**. If χ_u^2 and χ_v^2 are two independent chi-square random variables with u and v degrees of freedom, respectively, then the ratio

$$F_{u,v} = \frac{\chi_u^2/u}{\chi_v^2/v} \tag{1.20}$$

follows the ***F*-distribution with *u numerator* degrees of freedom** and ***v denominator* degrees of freedom**. If x is an F random variable with u numerator and v denominator degrees of freedom, then the probability distribution of x is

$$h(x) = \frac{\Gamma\left(\frac{u+v}{2}\right)\left(\frac{u}{v}\right)^{u/2} x^{(u/2)-1}}{\Gamma\left(\frac{u}{x}\right)\Gamma\left(\frac{v}{2}\right)\left[\left(\frac{u}{v}\right)x + 1\right]^{(u+v)/2}} \quad 0 < x < \infty \tag{1.21}$$

Several F-distributions are shown in Figure 1.16. This distribution is very important in the statistical analysis of designed experiments. The appendix provides JSL commands for calculating the percentage points of the F-distribution.

As an example of a statistic that is distributed as F, suppose we have two independent normal populations with common variance σ^2. If $y_{11}, y_{12}, \ldots, y_{1n_1}$ is a random sample of n_1 observations from the first population, and if $y_{21}, y_{22}, \ldots, y_{2n_2}$ is a random sample of n_2 observations from the second, then

$$\frac{S_1^2}{S_2^2} \sim F_{n_1-1, n_2-1} \tag{1.22}$$

where S_1^2 and S_2^2 are the two sample variances. This result follows directly from Eqs. (1.15) and (1.20).

1.6.3 Statistical Intervals and Tests of Hypotheses

The field of statistical inference consists of those methods used to make decisions or to draw conclusions about a population. These methods use the information contained in a random sample

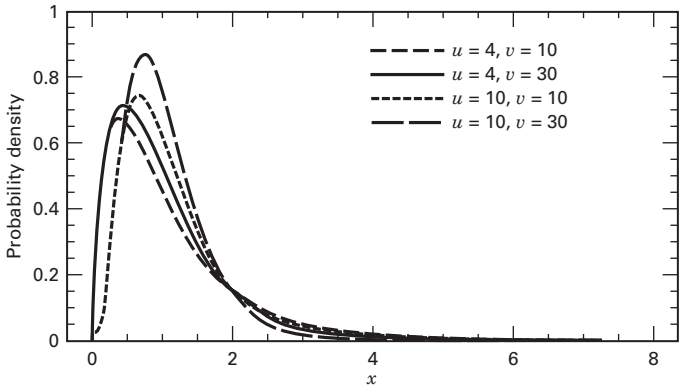


FIGURE 1.16 The F-distribution for Several Different Choices of the Two Parameters

from the population in drawing conclusions. Statistical inference may be divided into two major areas: parameter estimation and hypothesis testing. As an example of a parameter estimation problem, suppose that an engineer is analyzing the tensile strength of a component used in an automobile chassis. Because variability in tensile strength is naturally present between the individual components because of differences in raw material batches, manufacturing processes, and measurement procedures (for example), the engineer is interested in estimating the mean tensile strength of the components. Knowledge of the statistical sampling properties of the estimator used would enable the engineer to establish the precision of the estimate.

Now consider a different situation involving the strength of the automobile chassis components. The engineer conjectures that the mean strength of these components is greater than 1000 psi. Statistical hypothesis testing is a framework for solving problems of this type. In this case, the hypothesis would be that the mean strength exceeds 1000 psi. There is no emphasis on estimating strength; instead, the focus is on drawing conclusions about the stated hypothesis.

1.6.3.1 Comparing a Single Mean to a Specified Value, Variance Known

Some experiments involve comparing only one population mean μ to a specified value, say μ_0. The hypotheses are

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu \neq \mu_0 \end{aligned} \tag{1.23}$$

The statement H_0 is called the null hypothesis and H_1 is called the alternative hypothesis. The alternative hypothesis here is two-sided (or two-tailed) because we are interested in rejecting the null hypothesis for any value of the mean μ that differs from the value μ_0.

If the population is normal with known variance, or if the population is nonnormal but the sample size is large enough so that the central limit theorem applies, then the hypothesis may be tested using a direct application of the normal distribution. The **one-sample Z-test** statistic is

$$Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} \tag{1.24}$$

If the null hypothesis is true, the distribution of the test statistic is standard normal, that is, $N(0,1)$. Notice that the test statistic is just the ratio of the difference between the sample mean and the hypothesized value of the population mean divided by the standard error. It is a dimensionless quantity that shows how far $\bar{y}$ is from μ_0 in standard deviation units.

The standard normal distribution is called the **reference distribution** for the test. This implies that when we take a random sample from the population of interest and compute the test statistic we expect Z_0 to be close to zero. That is, a small positive or small negative value of the test statistic is evidence in support of the null hypothesis. However, if Z_0 is not very close to zero, this is evidence against H_0. Suppose that we are willing to incorrectly reject the null hypothesis 5% of the time; that is, reject H_0 and conclude that the mean μ is not equal to μ_0 when it is actually equal to μ_0. Let $\Phi(Z)$ be the cumulative standard normal probability associated with the value Z; that is, $\Phi(Z)$ is the area (probability) to the left of Z on the standard normal distribution. Suppose that we set this probability equal to 0.975. That is, $\Phi(1.96) = 0.975$. We refer to this value of Z as the **upper 2½ percent point of the standard normal distribution** and we denote it by $Z_{0.025}$. Now, the standard normal distribution is symmetric around zero, so $-Z_{0.025} = -1.96$ would be the **lower 2½ percent point of the standard normal distribution** because the probability below this value is 0.025. These two Z-values, $Z_{0.025} = 1.96$ and $-Z_{0.025} = -1.96$, are the **critical values** of the statistical test and values of Z above 1.96 and below -1.96 are called the **critical region** for the test statistic. If the value of the test statistic Z_0 in Eq. (1.24) is greater than 1.96 or less than -1.96, we would reject the null hypothesis. This choice of critical values ensures that if the null hypothesis is true we will only wrongly reject H_0 5% of the time. Rejecting the null hypothesis

when it is true is called a **type I error**, and in our example the probability of a type I error was selected to be 0.05. In general, we use the symbol α to represent the type I error probability. The probability of making a type I error is also called the **significance level** of the statistical test.

There is nothing "magic" about the choice of 0.05 as the probability of a type I error. The experimenter can select any value that he or she thinks is appropriate. However, values in the range 0.01–0.10 are widely used in practice, with 0.05 being a "default" choice in much engineering and scientific research. You should consider the consequences (economic or otherwise) when choosing the type I error rate.

To generalize, suppose that the type I error probability is α. Let $Z_{\alpha/2}$ be the upper $\alpha/2$ percentage point of the standard normal distribution. Then to test the hypotheses in Eq. (1.23), we would take a random sample of size n from the population of interest and compute the value of the test statistic Z_0 in Eq. (1.24). We would reject the null hypothesis H_0 if either

$$Z_0 > Z_{\alpha/2}$$

or

$$Z_0 < -Z_{\alpha/2}$$

and we should fail to reject H_0 if

$$-Z_{\alpha/2} \le Z_0 \le Z_{\alpha/2}$$

Figure 1.17a shows the location of the critical regions and the critical values for this test.

Sometimes we are interested in rejecting the null hypothesis only if the mean is greater than μ_0. The hypotheses are

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu > \mu_0 \end{aligned} \tag{1.25}$$

This is called a one-sided or upper-tailed alternative hypothesis. In this situation, only values of the sample mean $\bar{y}$ that are greater than μ_0 are consistent with the Alternative hypothesis, so only positive values of the test statistic should lead to rejection of H_0. As a result, we should reject only if

$$Z_0 > Z_\alpha.$$

Refer to Figure 1.17b. Finally, consider the other one-sided test

$$\begin{aligned} H_0 &: \mu = \mu_0 \\ H_1 &: \mu < \mu_0 \end{aligned} \tag{1.26}$$

This is a one-sided lower-tailed alternative hypothesis. In this situation, only values of the sample mean $\bar{y}$ that are less than μ_0 are consistent with the Alternative hypothesis, so only negative values of the test statistic should lead to rejection of H_0. As a result, we should reject only if

$$Z_0 < Z_\alpha.$$

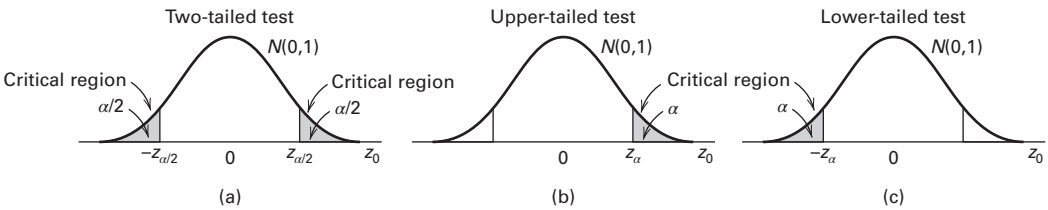


FIGURE 1.17 The distribution of Z_0 when $H_0 : \mu = \mu_0$ is true, with critical region for (a) The two-sided alternative $H_1 : \mu \neq \mu_0$. (b) The one-sided alternative $H_1 : \mu > \mu_0$. (c) The one-sided alternative $H_1 : \mu < \mu_0$

EXAMPLE 1.1 Golf Balls

Recall the golf ball overall distance data introduced earlier. Ten balls were selected at random from the balls produced by a golf equipment manufacturer and tested using a mechanical robot to hit the balls. The distances in yards traveled by the balls (carry plus roll) are as follows: 292, 270, 271, 278, 279, 268, 284, 271, 262, and 283. Assume that the standard deviation of overall distance is known to be $\sigma = 10$ yards. Can the manufacturer support the claim that the mean overall distance for this type of golf ball is at least 270 yards? Use $\alpha = 0.05$.

The appropriate hypotheses for this problem are

$$H_0 : \mu = 270$$
$$H_1 : \mu > 270$$

The sample average is $\bar{y} = 275.8$, so the value of the test statistic is

$$Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}} = \frac{275.8 - 270}{10/\sqrt{10}} = 1.83$$

Because we have chosen $\alpha = 0.05$, we will reject the null hypothesis if the test statistic value Z_0 is greater than $Z_{\alpha/2} = 1.645$. Since $Z_0 = 1.83 > Z_{\alpha/2} = 1.645$, we reject H_0 and conclude that the mean overall distance for this type of golf ball exceeds 270 yards.

1.6.3.2 The Use of *P*-values in Hypothesis Testing

The method that we have described to report the results of a hypothesis test essentially states that the null hypothesis was or was not rejected at a specified α-value or level of significance. This is called **fixed significance level** testing. The fixed significance level approach to hypothesis testing is useful in that it leads directly to methods for determining the appropriate sample sizes to use in hypothesis testing. However, the fixed significance level approach does have some disadvantages.

For example, in the golf ball problem above, we can say that H_0: $\mu = 270$ was rejected at the 0.05 level of significance. This statement of conclusions may be often inadequate because it gives the decision-maker no idea about whether the computed value of the test statistic was just barely in the rejection region or whether it was very far into this region. Furthermore, stating the results this way imposes the predefined level of significance on other users of the information. This approach may be unsatisfactory because some decision-makers might be uncomfortable with the risks implied by $\alpha = 0.05$.

To avoid these difficulties, the *P*-value approach has been adopted widely in practice. Roughly speaking, the *P*-value is the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis H_0 is true. Thus, a *P*-value directly measures the weight of evidence against H_0; the smaller the *P*-value, the stronger the evidence against H_0. Therefore, a decision-maker can draw a conclusion at any specified level of significance. We now give a formal definition:

> The *P*-value is the smallest level of significance that would lead to rejection of the null hypothesis H_0 with the given data.

Another way to think of this is that the *P*-value is the probability that you are wrong if you reject the null hypothesis.

Let us consider the case

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

To test the null hypothesis using the *P*-value approach, we would find the probability of observing a value of the sample mean that is at least as extreme as the observed value $\bar{y}$, given that the null hypothesis is true. The standard normal *Z*-value that corresponds to $\bar{y}$ is found from the test statistic:

$$Z_0 = \frac{\bar{y} - \mu_0}{\sigma/\sqrt{n}}$$

In terms of the standard normal cumulative distribution, the probability we are seeking is $1 - \Phi(|Z_0|)$. The reason that the argument of the standard normal cumulative distribution is $|Z_0|$ is that the value of Z_0 could be either positive or negative, depending on the observed sample mean.

Because this is a two-tailed test, this is only one-half of the *P*-value. Therefore, for the two-sided alternative hypothesis we are considering, the *P*-value is

$$P = 2[1 - \Phi(|Z_0|)] \tag{1.27}$$

Figure 1.18a illustrates the computation of the *P*-value.

Now let us consider the one-sided alternatives. Suppose that we are testing

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

In this situation, only values of the sample mean $\bar{y}$ that are greater than μ_0 are consistent with the alternative hypothesis, so only positive values of the test statistic should lead to rejection of H_0. Therefore, the *P*-value would be the probability that the standard normal random variable is greater than the observed value of the test statistic Z_0. This *P*-value is computed as

$$P = 1 - \Phi(Z_0) \tag{1.28}$$

Refer to Figure 1.18b. Finally, consider the other one-sided test:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

In this situation, only values of the sample mean $\bar{y}$ that are less than μ_0 are consistent with the alternative hypothesis, so only negative values of the test statistic should lead to rejection of H_0.

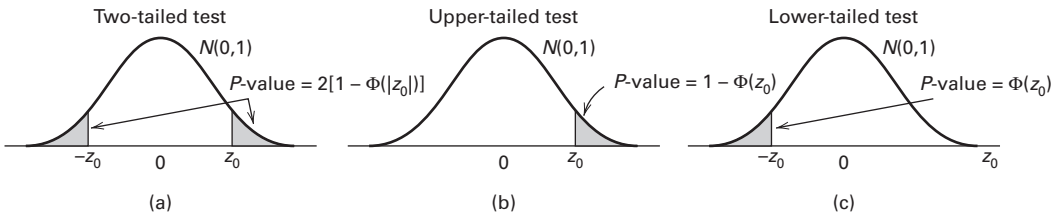


FIGURE 1.18 The *P*-value for a *z*-test. (a) The two-sided alternative $H_1 : \mu \neq \mu_0$. (b) The one-sided alternative $H_1 : \mu > \mu_0$. (c) The one-sided alternative $H_1 : \mu < \mu_0$

Therefore, the *P*-value would be the probability that the standard normal random variable is less than the observed value of the test statistic Z_0. This *P*-value is computed as

$$P = \Phi(Z_0) \tag{1.29}$$

Refer to Figure 1.18c.

EXAMPLE 1.2 Golf Balls

Reconsider the golf ball test data from Example 1.1. Recall that the hypotheses for this problem are

$$H_0 : \mu = 270$$
$$H_1 : \mu > 270$$

The sample average is $\bar{y} = 275.8$, so the value of the test statistic is $Z_0 = 1.83$. Because this is an upper-tailed one-sided test, the *P*-value is computed from Eq. (1.28):

$$P = 1 - \Phi(Z_0) = 1 - \Phi(1.83) = 0.0336$$

This is a small *P*-value, smaller than the "default" level of significance specified in the original problem statement, so the decision-maker should reject the null hypothesis. More formally, we can say that H_0 is rejected at the 0.0336 level of significance.

1.6.3.3 Confidence Interval on the Mean

In many situations, a point estimate does not provide enough information about a parameter. For example, in the golf ball problem, we have rejected the null hypothesis H_0: $\mu = 270$ and our estimate of the mean overall distance is $\bar{y} = 275.8$ yards. This is called a **point estimate**. It is never the case that the true mean overall distance is exactly equal to the point estimate 275.8 yards. In many cases, it would be preferable to have an **interval** in which we would expect to find the true mean distance. One way to accomplish this is with an interval estimate called a **confidence interval** (CI).

An interval estimate of the unknown parameter, such as μ, is an interval of the form

$$L \le \mu \le U$$

where the endpoints *L* and *U* depend on the numerical value of the sample mean for a particular sample. Because different samples will produce different values of the sample mean and, consequently, different values of the endpoints *L* and *U*, these endpoints are values of random variables. From the sampling distribution of the sample mean, we will be able to determine the values of *L* and *U* such that the following probability statement is true:

$$P(L \le \mu \le U) = 1 - \alpha \tag{1.30}$$

where $0 < \alpha < 1$. Thus, we have a probability of $1 - \alpha$ of selecting a sample that will produce an interval containing the true value of μ. The resulting interval

$$L \le \mu \le U \tag{1.31}$$

is the $100(1 - \alpha)\%$ CI for the mean μ. The quantities *L* and *U* are called the lower- and upper-confidence limits, respectively, and $1 - \alpha$ is called the **confidence coefficient**.

A strict interpretation of a CI is that, if an infinite number of random samples are collected and a $100(1-\alpha)\%$ CI for μ is computed from each sample, $100(1-\alpha)\%$ of these intervals will contain the true value of μ.

In practice, we obtain only one random sample and calculate one CI. Because this interval either will or will not contain the true value of μ, it is not reasonable to attach a probability level to this specific event. The appropriate statement is that the observed interval $[L, U]$ brackets the true value of μ with confidence $100(1-\alpha)$. This statement has a frequency interpretation; that is, we do not know whether the statement is true for this specific sample, but the method used to obtain the interval $[L, U]$ yields correct statements $100(1-\alpha)\%$ of the time.

The length $U - L$ of the observed two-sided CI is an important measure of the quality of the information obtained from the sample. The half-interval length $\mu - L$ or $U - \mu$ is called the **precision** of the estimator. The longer the CI, the more confident we are that the interval actually contains the true value of μ. On the other hand, the longer the interval is, the less information we have about the true value of μ. In an ideal situation, we obtain a relatively short interval with high confidence.

It is very easy to find the quantities L and U that define the two-sided CI for μ. We know that the sampling distribution of $\overline{X}$ is normal with mean μ and variance σ^2/n. Therefore, the distribution of the statistic

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is a standard normal distribution.

Therefore

$$P\{-Z_{\alpha/2} \le Z \le Z_{\alpha/2}\} = 1 - \alpha$$

so that

$$P\left\{-Z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}\right\} = 1 - \alpha$$

This can be rearranged as

$$P\left\{\overline{X} - \frac{Z_{\alpha/2}\sigma}{\sqrt{n}} \le \mu \le \overline{X} + \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}\right\} = 1 - \alpha$$

From the consideration of Eq. (1.30), the lower and upper limits of the inequalities in the last equation are the lower- and upper-confidence limits L and U, respectively. This leads to the following definition:

The $100(1-\alpha)\%$ CI on the mean μ with known standard deviation is

$$\overline{y} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{y} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{1.32}$$

where $Z_{\alpha/2}$ is the upper $100\alpha/2$ percentage point and $-Z_{\alpha/2}$ is the corresponding lower $100\alpha/2$ percentage point of the standard normal distribution.

To illustrate the construction of a CI, reconsider the golf ball distance data. Suppose that we want a 95% CI. Then $Z_{\alpha/2} = Z_{0.025} = 1.96$ and the 95% CI is

$$\overline{y} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{y} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

$$275.8 - 1.96\frac{10}{\sqrt{10}} \le \mu \le 275.8 + 1.96\frac{10}{\sqrt{10}}$$
$$275.8 - 1.96(3.16) \le \mu \le 275.8 + 1.96(3.16)$$
$$269.6 \le \mu \le 282.0$$

Remember how to interpret the CI; this specific interval either contains the true mean or it does not (and we do not know which), but because of the procedure that we used to construct the CI, in repeated sampling 95% of the intervals that we would compute will contain the true value of μ. It is reasonable to think of this interval as a range of highly likely values for the mean.

Both hypothesis testing and CI construction can be done using software. The output from JMP for the golf ball data is shown in Figure 1.19. JMP presents a normal quantile plot, which is useful in determining if the distribution of overall distance is normal, and since the observations are reasonably well approximated by a straight line, it is reasonable to assume that overall distance is at least approximately normally distributed. The quantiles are shown next, followed by the statistical test of the hypothesis that the mean overall distance equals 270 yards. We supplied the standard deviation as input to the software, so JMP performed the Z-test described above. The test statistic agrees closely with the value we computed manually. JMP reports a two-tailed P-value (0.0666) and both one-sided P-values. Because the alternative hypothesis was an upper-tailed one-sided test, the appropriate P-value is 0.0333. This agrees with our previous results.

1.6.3.4 The Experimental Design Problem

The experimental design problem in the case of statistical inference (hypothesis tests, CIs) on the mean of a population with known variance is fairly simple; it consists mainly of determining the appropriate sample size to use. Determination of sample size requires discussion of a type II error of a statistical test. Recall that a type I error is made if the null hypothesis is rejected when it is actually true. You control the probability of making a type I error when you select the significance level (α) for the test. The type II error is defined as failing to reject the null hypothesis when it is false. The probability of the type II error is denoted by β. Sometimes, it is more convenient to work in terms of the power of the test, where power $= 1 - \beta$. Therefore, power is the probability of rejecting the null hypothesis when it is in fact false. High power (or small β) in a statistical test is desirable.

The type II error probability depends on four parameters: the type I error probability, the sample size, the population standard deviation, and the true value of the mean. Fortunately, the type I error probability is controlled and the standard deviation is known. So to determine the appropriate sample size, all we need to do is to define a value of the true population mean, μ, different from the hypothesized value μ_0, such that if the true mean really is this value μ we want to reject the null hypothesis with high probability. Let the true value of the mean be

$$\mu = \mu_0 + \delta$$

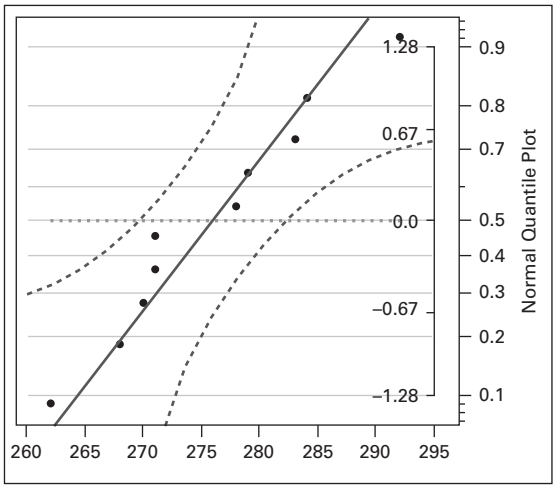
Then the relationship between β, δ, α, and the sample size n for the two-sided alternative hypothesis is given by

$$\beta = \Phi\left(Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) - \Phi\left(-Z_{\alpha/2} - \frac{\delta\sqrt{n}}{\sigma}\right) \tag{1.33}$$

See Montgomery and Runger (2018) for the development of this equation. This equation can be solved for the sample size n, yielding

$$n \cong \frac{(Z_{\alpha/2} + Z_{\beta})^2\sigma^2}{\delta^2} \tag{1.34}$$

Distance



Quantiles

100.0%	maximum	292
99.5%		292
97.5%		292
90.0%		291.2
75.0%	quartile	283.25
50.0%	median	274.5
25.0%	quartile	269.5
10.0%		262.6
2.5%		262
0.5%		262
0.0%	minimum	262

Test Mean

Hypothesized Value	270
Actual Estimate	275.8
DF	9
Std Dev	8.99135
Sigma given	10

	z Test
Test Statistic	1.8341
Prob > |z|	0.0666
Prob > z	0.0333*
Prob < z	0.9667

FIGURE 1.19 JMP Output for the Golf Ball Distance Data

For the one-sided alternatives, the analogous equation is

$$n = \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{\delta^2} \tag{1.35}$$

When using Eqs. (1.34) and (1.35), round up if the sample size is not an integer.

EXAMPLE 1.3 Golf Balls

Recall that in the golf ball overall distance problem we are testing

$$H_0 : \mu = 270$$
$$H_1 : \mu > 270$$

with $\alpha = 0.05$. We want to reject the null hypothesis with a high probability, say power of 0.90, if the mean is as large as 275 yards. What sample size is required?

Because $\alpha = 0.05$ and $\beta = 1 - \text{power} = 1 - 0.90 = 0.10$, we find $Z_\alpha = Z_{0.05} = 1.645$ and $Z_\beta = Z_{0.10} = 1.28$. We also know that $\sigma = 10$ and since the true value of the mean that we want to "detect" is $\mu = 275, \delta = \mu - \mu_0 = 275 - 270 = 5$. Using Eq. (1.35) the required sample size is

$$n = \frac{(Z_\alpha + Z_\beta)^2\sigma^2}{\delta^2} = \frac{(1.645 + 1.28)^2(10)^2}{5^2} = 34.22 \cong 35$$

1.6.3.5 Comparing a Single Mean to a Specified Value, Variance Unknown

We now consider the case of hypothesis testing and CI construction on a single mean where the population variance is unknown. If the sample size is large, say at least 30 or 40, the procedures presented previously for the known variance case can be used. In other words, they are **large-sample methods**. However, there are many situations where the sample size is small, so the methods we will describe are extremely useful. When the variance of the population is unknown, we must make the additional assumption that the population is normally distributed, although moderate departures from normality will not seriously affect the results.

Suppose that $y_1, y_2, \ldots, y_n$ is a random sample of size n from a population that has a normal distribution with mean μ and variance σ^2, and let $\bar{y}$ and S^2 be the sample average and sample variance of the observations, respectively. Then as noted in Eq. (1.19), the ratio

$$t = \frac{\bar{y} - \mu}{S/\sqrt{n}}$$

has a t-distribution with $n - 1$ degrees of freedom. Let $t_{\alpha/2,n-1}$ be the upper $\alpha/2$ percentage point of the t-distribution with $n - 1$ degrees of freedom; that is, the probability beyond $t_{\alpha/2,n-1}$ is $\alpha/2$. Because the t-distribution is symmetric about zero, the lower-tail percentage point that has probability $\alpha/2$ below it is $-t_{\alpha/2,n-1}$.

Consider testing the two-sided hypotheses

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0 \tag{1.36}$$

The **test statistic** is

$$t_0 = \frac{\bar{y} - \mu_0}{S/\sqrt{n}} \tag{1.37}$$

Notice that the test statistic is just the difference between the sample and hypothesized values of the mean divided by the estimated standard error. The decision procedure is very similar to the Z-test, except that the t-distribution is used to find the critical values instead of the standard normal. If the null hypothesis is true, t_0 has a t-distribution with $n - 1$ degrees of freedom, and we would reject H_0 if $t_0 > t_{\alpha/2,n-1}$ or if $t_0 < -t_{\alpha/2,n-1}$. If the alternative hypothesis is one-sided, say

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu > \mu_0$$

we would reject H_0 if $t_0 > t_{\alpha,n-1}$, and if the hypotheses are

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu < \mu_0$$

we would reject H_0 if $t_0 < -t_{\alpha,n-1}$.

EXAMPLE 1.4 Cloud Seeding

Cloud seeding has been studied for many decades as a weather modification procedure (for an interesting study of this subject, see the article in *Technometrics*, "A Bayesian Analysis of a Multiplicative Treatment Effect in Weather Modification," Vol. 17, 1975, pp. 161–166). The rainfall in acre-feet from 20 clouds that were selected at random and seeded with silver nitrate is as follows: 18.0, 30.7, 19.8, 27.1, 22.3, 18.8, 31.8, 23.4, 21.2, 27.9, 31.9, 27.1, 25.0, 24.7, 26.9, 21.8, 29.2, 34.8, 26.7, and 31.6. Can you support the claim that mean rainfall from seeded clouds exceeds 25 acre-feet? Use $\alpha = 0.05$.

From the sample data, we find that $\bar{y} = 26.035$ and $S = 4.785$. The hypotheses we are going to test are

$$H_0 : \mu = 25$$
$$H_1 : \mu > 25$$

The value of the test statistic is

$$t_0 = \frac{\bar{y} - \mu_0}{S/\sqrt{n}} = \frac{26.035 - 25}{4.785/\sqrt{20}} = 0.9674$$

Since this is an upper-tailed one-sided test, we would reject the null hypothesis if the value of the test statistic $t_0 > t_{\alpha,n-1} = t_{0.05,19}$. Using JSL commands from the appendix, we find that $t_{0.05,19} = 1.729$, so we cannot reject the null hypothesis. There is little evidence in the data to support the claim that the mean rainfall from seeded clouds exceeds 25 acre-feet.

1.6.3.6 *P*-values in the *t*-test

We can use P-values in the t-test in a manner similar to their usage in the Z-test described earlier. Statistics software packages calculate and display P-values. However, in working problems by hand, it is useful to be able to find the P-value for a t-test. Fortunately, it is easy to find lower and upper bounds on the P-value by using JMP scripting functions described in the appendix. It is also helpful to refer to Figure 1.20.

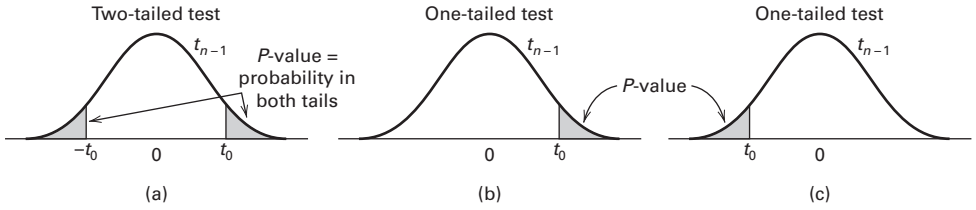


FIGURE 1.20
Calculating the P-value for a t-test: (a) $H_1 : \mu \neq \mu_0$; (b) $H_1 : \mu > \mu_0$; (c) $H_1 : \mu < \mu_0$

To illustrate, suppose that we are conducting an upper-tailed t-test (so H_1: $\mu > \mu_0$) with 14 degrees of freedom. The relevant critical values from the appendix are as follows:

Critical value:	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
Tail area:	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0025	0.001	0.0005

After calculating the test statistic, we find that $t_0 = 2.8$. Now, $t_0 = 2.8$ is between two tabulated values 2.624 and 2.977. Therefore, the P-value must be between 0.01 and 0.005. These are effectively the upper and lower bounds on the P-value.

This illustrates the procedure for an upper-tailed test. If the test is lower tailed, just change the sign on the lower and upper bounds for t_0 and proceed as above. Remember that for a two-tailed test, the level of significance associated with a particular critical value is twice the corresponding tail area in the column heading. This consideration must be taken into account when we compute the bound on the P-value. For example, suppose that $t_0 = 2.8$ for a two-tailed alternative based on 14 degrees of freedom. The value of the test statistic $t_0 > 2.624$ (corresponding to $\alpha = 2 \times 0.01 = 0.02$) and $t_0 < 2.977$ (corresponding to $\alpha = 2 \times 0.005 = 0.01$), so the lower and upper bounds on the P-value would be $0.01 < P < 0.02$ for this case.

We can also find CIs on the mean μ of a normal distribution where the variance is unknown. The CI equation is as follows:

The $100(1 - \alpha)\%$ CI on the mean μ of a normal distribution with unknown standard deviation is

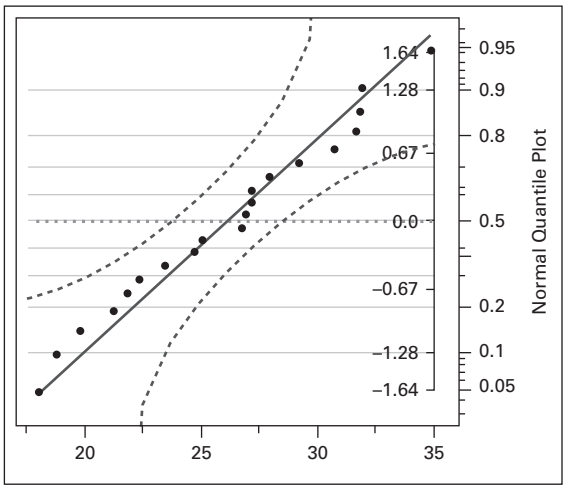
$$\bar{y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \tag{1.38}$$

where $t_{\alpha/2,n-1}$ is the upper $100\alpha/2$ percentage point and $-t_{\alpha/2,n-1}$ is the corresponding lower $100\alpha/2$ percentage point of the t-distribution with $n - 1$ degrees of freedom in the appendix.

The JMP output for the cloud seeding problem is shown in Figure 1.21. Notice that the P-value reported is 0.1728. You should be able to verify that $0.10 < P < 0.25$, so we cannot reject the null hypothesis. There is little evidence to indicate that the mean rainfall from seeded clouds exceeds 25 acre-feet. JMP also provides the 95% CI from Eq. (1.38).

We can also determine sample sizes for the t-test. Unfortunately, there are no closed-form equations for this as there were in the Z-test because the probability calculations involve the noncentral t-distribution, which has a very messy probability density function. When the standard deviation is unknown, we must work in terms of the difference between the true mean and the hypothesized mean that we want to detect as a ratio, say δ/σ. If we are interested in detecting small differences between the means, we should select a small ratio, say $0 < \delta/\sigma \leq 1$. For detecting moderate differences, we could select $1 < \delta/\sigma \leq 2$, and for detecting large differences we could

Rainfall



Quantiles

100.0%	maximum	34.8
99.5%		34.8
97.5%		34.8
90.0%		31.89
75.0%	quartile	30.325
50.0%	median	26.8
25.0%	quartile	21.925
10.0%		18.9
2.5%		18
0.5%		18
0.0%	minimum	18

Moments

Mean	26.035
Std Dev	4.7847647
Std Err Mean	1.0699059
Upper 95% Mean	28.274339
Lower 95% Mean	23.795661
N	20

Test Mean

Hypothesized Value	25
Actual Estimate	26.035
DF	19
Std Dev	4.78476

t Test

Test Statistic	0.9674
Prob > |t|	0.3455
Prob > t	0.1728
Prob < t	0.8272

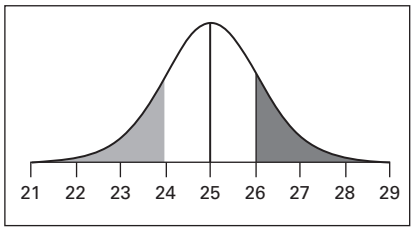


FIGURE 1.21 JMP Output for the Rainfall Data

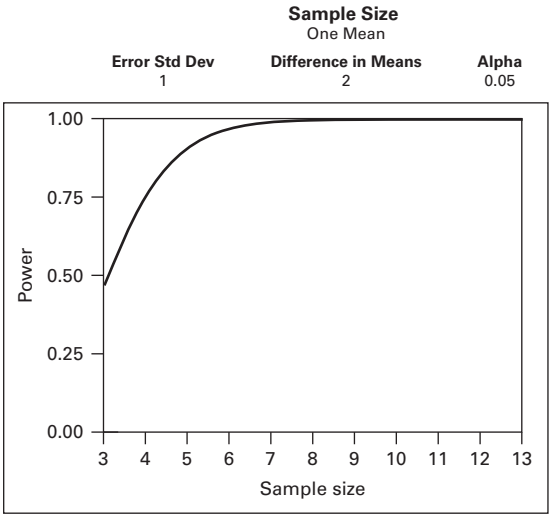


FIGURE 1.22 Power Versus Sample Size Curve for the Cloud Seeding Data

select $\delta/\sigma > 2$. Software can be used to determine the appropriate sample size for many situations. Suppose that in the cloud seeding problem we wanted to detect a moderately large difference between the true mean and the hypothesized value of 25. Figure 1.22 is an output from the JMP power and sample size calculator for $\delta/\sigma = 2$. If the sample size is at least $n = 8$, the statistical test will have very high power, exceeding 0.99.