2

Data Collection

Data collection is arguably the most important activity of engineering statistics. Often, properly collected data will essentially speak for themselves, making formal inferences rather like frosting on the cake. On the other hand, no amount of cleverness in post-facto data processing will salvage a badly done study. So it makes sense to consider carefully how to go about gathering data.

This chapter begins with a discussion of some general considerations in the collection of engineering data. It turns next to concepts and methods applicable specifically in enumerative contexts, followed by a discussion of both general principles and some specific plans for engineering experimentation. The chapter concludes with advice for the step-by-step planning of a statistical engineering study.

2.1 General Principles in the Collection of Engineering Data

Regardless of the particulars of a statistical engineering study, a number of common general considerations are relevant. Some of these are discussed in this section, organized around the topics of measurement, sampling, and recording.

2.1.1 Measurement

Good measurement is indispensable in any statistical engineering study. An engineer planning a study ought to ensure that data on relevant variables will be collected by well-trained people using measurement equipment of known and adequate quality.

When choosing variables to observe in a statistical study, the concepts of measurement validity and precision, discussed in Section 1.3, must be remembered. One practical point in this regard concerns how directly a measure represents a system property. When a **direct measure** exists, it is preferable to an indirect measure, because it will usually give much better precision.

Example 1

Exhaust Temperature Versus Weight Loss

An engineer working on a drying process for a bulk material was having difficulty determining when a target dryness had been reached. The method being used was monitoring the temperature of hot air being exhausted from the dryer. Exhaust temperature was a valid but very imprecise indicator of moisture content.

Someone suggested measuring the weight loss of the material instead of exhaust temperature. The engineer developed an ingenious method of doing this, at only slightly greater expense. This much more direct measurement greatly improved the quality of the engineer's information.

It is often easier to identify appropriate measures than to carefully and unequivocally define them so that they can be used. For example, suppose a metal cylinder is to be turned on a lathe, and it is agreed that cylinder diameter is of engineering importance. What is meant by the word *diameter*? Should it be measured on one end of the cylinder (and if so, which?) or in the center, or where? In practice, these locations will differ somewhat. Further, when a cylinder is gauged at some chosen location, should it be rolled in the gauge to get a maximum (or minimum) reading, or should it simply be measured as first put into the gauge? The cross sections of real-world cylinders are not exactly circular or uniform, and how the measurement is done will affect how the resulting data look.

It is especially necessary—and difficult—to make careful **operational definitions** where qualitative and count variables are involved. Consider the case of a process engineer responsible for an injection-molding machine producing plastic auto grills. If the number of abrasions appearing on these is of concern and data are to be gathered, how is *abrasion* defined? There are certainly locations on a grill where a flaw is of no consequence. Should those areas be inspected? How big should an abrasion be in order to be included in a count? How (if at all) should an inspector distinguish between abrasions and other imperfections that might appear on a grill? All of these questions must be addressed in an operational definition of "abrasion" before consistent data collection can take place.

Once developed, operational definitions and standard measurement procedures must be communicated to those who will use them. Training of technicians has to be taken seriously. Workers need to understand the importance of adhering to the standard definitions and methods in order to provide consistency. For example, if instructions call for zeroing an instrument before each measurement, it must always be done.

The performance of any measuring equipment used in a study must be known to be adequate—both before beginning and throughout the study. Most large industrial concerns have regular programs for both recalibrating and monitoring the precision of their measuring devices. The second of these activities sometimes goes under the name of **gauge R and R studies**—the two R's being **repeatability** and **reproducibility**. Repeatability is variation observed when a single operator uses the gauge to measure and remeasure one item. Reproducibility is variation in measurement attributable to differences among operators. (A detailed discussion of such studies can be found in Section 2.2.2 of *Statistical Quality Assurance Methods for Engineers* by Vardeman and Jobe.)

Calibration and precision studies should assure the engineer that instrumentation is adequate at the beginning of a statistical study. If the time span involved in the study is appreciable, the *stability* of the instrumentation must be maintained over the study period through checks on calibration and precision.

2.1.2 Sampling

Once it is established how measurement/observation will proceed, the engineer can consider how much to do, who is to do it, where and under what conditions it is to be done, etc. Sections 2.2, 2.3, and 2.4 consider the question of choosing what observations to make, first in enumerative and then in experimental studies. But first, a few general comments about the issues of "How much?", "Who?", and "Where?".

How much data?

The most common question engineers ask about data collection is "How many observations do I need?" Unfortunately, the proper answer to the question is typically "it depends." As you proceed through this book, you should begin to develop some intuition and some rough guides for choosing sample sizes. For the time being, we point out that the only factor on which the answer to the sample size question really depends is the variation in response that one expects (coming both from unit-to-unit variation and from measurement variation).

This makes sense. If objects to be observed were all alike and perfect measurement were possible, then a single observation would suffice for any purpose. But if there is increase either in the measurement noise or in the variation in the system or population under study, the sample size necessary to get a clear picture of reality becomes larger.

However, one feature of the matter of sample size sometimes catches people a bit off guard—the fact that in enumerative studies (provided the population size is large), sample size requirements do not depend on the population size. That is, sample size requirements are not relative to population size, but, rather, are absolute. If a sample size of 5 is adequate to characterize compressive strengths of a lot of 1,000 red clay bricks, then a sample of size 5 would be adequate to characterize compressive strengths for a lot of 100,000 bricks with similar brick-to-brick variability.

Who should collect data?

The "Who?" question of data collection cannot be effectively answered without reference to human nature and behavior. This is true even in a time when automatic data collection devices are proliferating. Humans will continue to supervise these and process the information they generate. Those who collect engineering data must not only be well trained; they must also be convinced that the data they collect will be used and in a way that is in their best interests. Good data must be seen as a help in doing a good job, benefiting an organization, and remaining employed, rather than as pointless or even threatening. If those charged with collecting or releasing data believe that the data will be used against them, it is unrealistic to expect them to produce useful data.

Example 2

Data—An Aid or a Threat?

One of the authors once toured a facility with a company industrial statistician as guide. That person proudly pointed out evidence that data were being collected and effectively used. Upon entering a certain department, the tone of the conversation changed dramatically. Apparently, the workers in that department had been asked to collect data on job errors. The data had pointed unmistakably to poor performance by a particular individual, who was subsequently fired from the company. Thereafter, convincing other workers that data collection is a helpful activity was, needless to say, a challenge.

Perhaps all the alternatives in this situation (like retraining or assignment to a different job) had already been exhausted. But the appropriateness of the firing is not the point here. Rather, the point is that circumstances were allowed to create an atmosphere that was not conducive to the collection and use of data.

Even where those who will gather data are convinced of its importance and are eager to cooperate, care must be exercised. Personal biases (whether conscious or subconscious) must not be allowed to enter the data collection process. Sometimes in a statistical study, hoped-for or predicted best conditions are deliberately or unwittingly given preference over others. If this is a concern, measurements can be made **blind** (i.e., without personnel knowing what set of conditions led to an item being measured). Other techniques for ensuring fair play, having less to do with human behavior, will be discussed in the next two sections.

Where should data be collected?

The "Where?" question of engineering data collection can be answered in general terms: "As close as possible in time and space to the phenomenon being studied." The importance of this principle is most obvious in the routine monitoring of complex manufacturing processes. The performance of one operation in such a process is most effectively monitored at the operation rather than at some later point. If items being produced turn out to be unsatisfactory at the end of the line, it is rarely easy to backtrack and locate the operation responsible. Even if that is accomplished, unnecessary waste has occurred during the time lag between the onset of operation malfunction and its later discovery.

Example 3

IC Chip Manufacturing Process Improvement

The preceding point was illustrated during a visit to a "clean room" where integrated circuit chips are manufactured. These are produced in groups of 50 or so on so-called wafers. Wafers are made by successively putting down a number of appropriately patterned, very thin layers of material on an inert background disk. The person conducting the tour said that at one point, a huge fraction of wafers produced in the room had been nonconforming. After a number of false starts, it was discovered that by appropriate testing (data collection) at the point of application of the second layer, a majority of the eventually nonconforming

Example 3 (*continued*)

wafers could be identified and eliminated, thus saving the considerable extra expense of further processing. What's more, the need for adjustments to the process was signaled in a timely manner.

2.1.3 Recording

The object of engineering data collection is to get data used. How they are recorded has a major impact on whether this objective is met. A good data recording format can make the difference between success and failure.

Example 4

A Data Collection Disaster

A group of students worked with a maker of molded plastic business signs in an effort to learn what factors affect the shrinkage a sign undergoes as it cools. They considered factors such as Operator, Heating Time, Mold Temperature, Mold Size, Ambient Temperature, and Humidity. Then they planned a partially observational and partially experimental study of the molding process. After spending two days collecting data, they set about to analyze them. The students discovered to their dismay that although they had recorded many features of what went on, they had neglected to record either the size of the plastic sheets before molding or the size of the finished signs. Their considerable effort was entirely wasted. It is likely that this mistake could have been prevented by careful precollection development of a data collection form.

When data are collected in a routine, ongoing, process-monitoring context (as opposed to a one-shot study of limited duration), it is important that they be used to provide effective, timely feedback of information. Increasingly, computer-made graphical displays of data, in real time, are used for this purpose. But it is often possible to achieve this much more cheaply through clever design of a manual data collection form, if the goal of making data recording convenient and immediately useful is kept in sight.

Example 5

Recording Bivariate Data on PVC Bottles

Table 2.1 presents some bivariate data on bottle mass and width of bottom piece resulting from blow molding of PVC plastic bottles (taken from *Modern Methods for Quality Control and Improvement* by Wadsworth, Stephens, and Godfrey). Six consecutive samples of size 3 are represented.

Such bivariate data could be recorded in much the same way as they are listed in Table 2.1. But if it is important to have immediate feedback of information (say, to the operator of a machine), it would be much more effective to use a well-thought-out bivariate **check sheet** like the one in Figure 2.1. On such a sheet, it

Table 2.1
Mass and Bottom Piece Widths of PVC Bottles

Sample	Item	Mass (g)	Width (mm)	Sample	Item	Mass (g)	Width (mm)
1	1	33.01	25.0	4	10	32.80	26.5
1	2	33.08	24.0	4	11	32.86	28.5
1	3	33.24	23.5	4	12	32.89	25.5
2	4	32.93	26.0	5	13	32.73	27.0
2	5	33.17	23.0	5	14	32.57	28.0
2	6	33.07	25.0	5	15	32.65	26.5
3	7	33.01	25.5	6	16	32.43	30.0
3	8	32.82	27.0	6	17	32.54	28.0
3	9	32.91	26.0	6	18	32.61	26.0

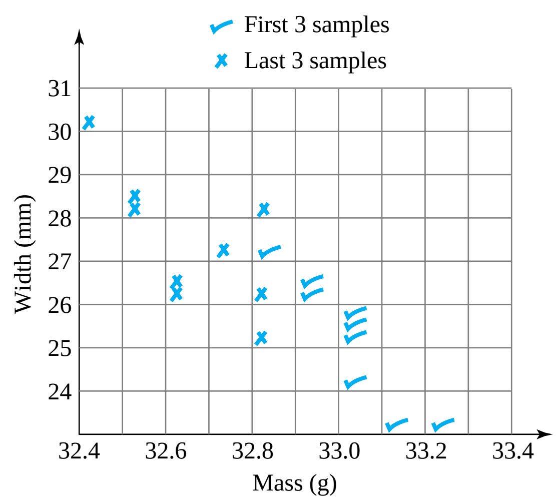

Figure 2.1 Check sheet for the PVC bottle data

is easy to see how the two variables are related. If, as in the figure, the recording symbol is varied over time, it is also easy to track changes in the characteristics over time. In the present case, width seems to be inversely related to mass, which appears to be decreasing over time.

To be useful (regardless of whether data are recorded on a routine basis or in a one-shot mode, automatically or by hand), the recording must carry enough **documentation** that the important circumstances surrounding the study can be reconstructed. In a one-shot experimental study, someone must record responses

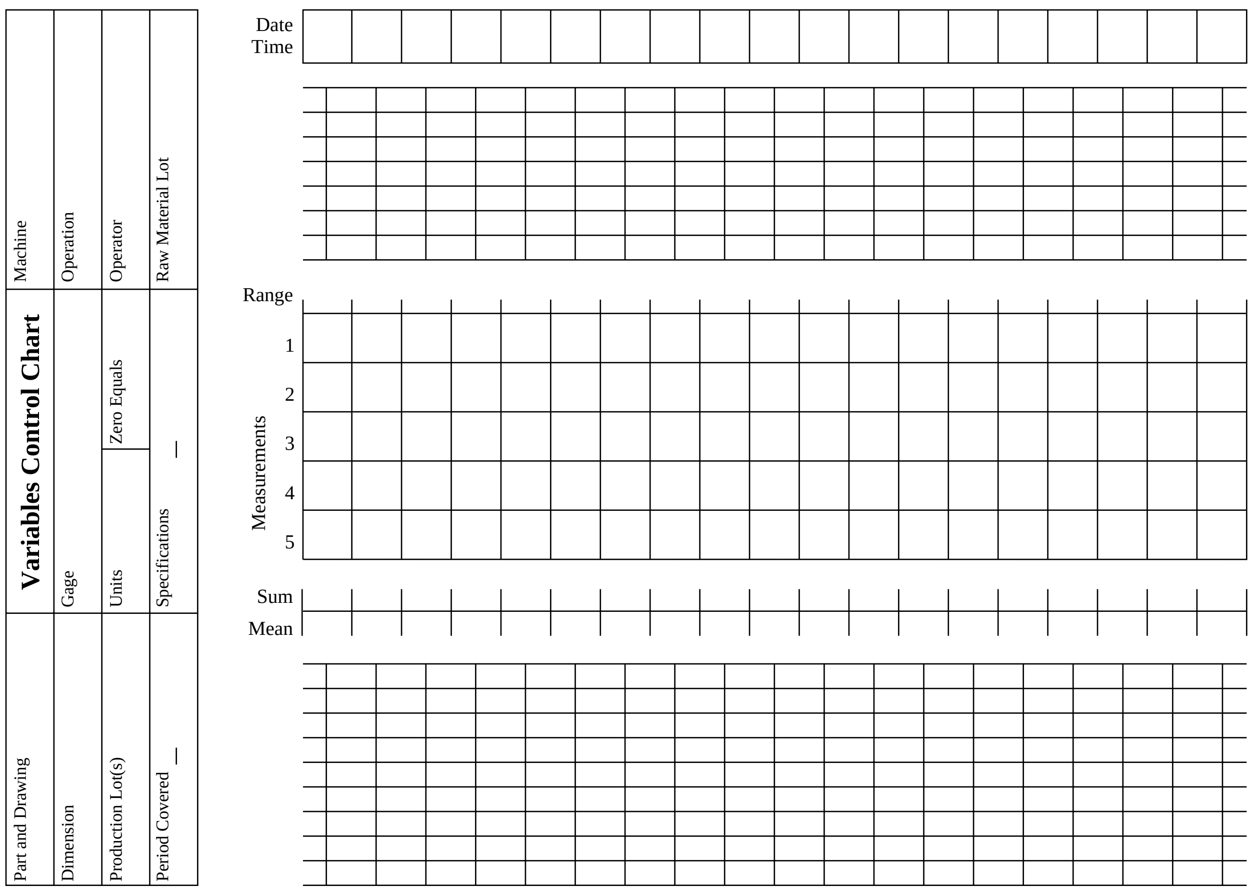

Figure 2.2 Variables control chart form

and values of the experimental variables, and it is also wise to keep track of other variables that might later prove to be of interest. In the context of routine process monitoring, data records will be useful in discovering differences in raw material lots, machines, operators, etc., only if information on these is recorded along with the responses of interest. Figure 2.2 shows a form commonly used for the routine collection of measurements for process monitoring. Notice how thoroughly the user is invited to document the data collection.

Section 1 Exercises

1. Consider the context of a study on making paper airplanes where two different Designs (say delta versus t wing), two different Papers (say construction versus typing), and two different Loading Conditions (with a paper clip versus without a paper clip) are of interest with regard to their impact on flight distance. Give an operational definition of *flight distance* that you might use in such a study.

2. Explain how training operators in the proper use of measurement equipment might affect both the repeatability and the reproducibility of measurements made by an organization.
3. What would be your response to another engineer's comment, "We have great information on our product—we take 5% samples of every outgoing order, regardless of order size!"?
4. State briefly why it is critical to make careful operational definitions for response variables in statistical engineering studies.

2.2 Sampling in Enumerative Studies

An enumerative study has an identifiable, concrete population of items. This section discusses selecting a sample of the items to include in a statistical investigation.

Using a sample to represent a (typically much larger) population has obvious advantages. Measuring some characteristics of a sample of 30 electrical components from an incoming lot of 10,000 can often be feasible in cases where it would not be feasible to perform a **census** (a study that attempts to include every member of the population). Sometimes testing is destructive, and studying an item renders it unsuitable for subsequent use. Sometimes the timeliness and data quality of a sampling investigation far surpass anything that could be achieved in a census. Data collection technique can become lax or sloppy in a lengthy study. A moderate amount of data, collected under close supervision and put to immediate use, can be very valuable—often more valuable than data from a study that might appear more complete but in fact takes too long.

If a sample is to be used to stand for a population, how that sample is chosen becomes very important. The sample should somehow be representative of the population. The question addressed here is how to achieve this.

Systematic and **judgment-based** methods can in some circumstances yield samples that faithfully portray the important features of a population. If a lot of items is manufactured in a known order, it may be reasonable to select, say, every 20th one for inclusion in a statistical engineering study. Or it may be effective to force the sample to be balanced—in the sense that every operator, machine, and raw material lot (for example) appears in the sample. Or an old hand may be able to look at a physical population and fairly accurately pick out a representative sample.

But there are potential problems with such methods of sample selection. Humans are subject to conscious and subconscious preconceptions and biases. Accordingly, judgment-based samples can produce distorted pictures of populations. Systematic methods can fail badly when unexpected cyclical patterns are present. (For example, suppose one examines every 20th item in a lot according to the order in which the items come off a production line. Suppose further that the items are at one point processed on a machine having five similar heads, each performing the same operation on every fifth item. Examining every 20th item only gives a picture of how one of the heads is behaving. The other four heads could be terribly misadjusted, and there would be no way to find this out.)

Even beyond these problems with judgment-based and systematic methods of sampling, there is the additional difficulty that it is not possible to quantify their properties in any useful way. There is no good way to take information from samples drawn via these methods and make reliable statements of likely margins of error. The method introduced next avoids the deficiencies of systematic and judgment-based sampling.

Definition 1
Simple random sampling

A **simple random sample of size** n from a population is a sample selected in such a manner that every collection of n items in the population is a priori equally likely to compose the sample.

Probably the easiest way to think of simple random sampling is that it is conceptually equivalent to drawing n slips of paper out of a hat containing one for each member of the population.

Example 6

Random Sampling Dorm Residents

C. Black did a partially enumerative and partially experimental study comparing student reaction times under two different lighting conditions. He decided to recruit subjects from his coed dorm floor, selecting a simple random sample of 20 of these students to recruit. In fact, the selection method he used involved a table of so-called random digits. But he could have just as well written the names of all those living on his floor on standard-sized slips of paper, put them in a bowl, mixed thoroughly, closed his eyes, and selected 20 different slips from the bowl.

Mechanical methods and simple random sampling

Methods for actually carrying out the selection of a simple random sample include **mechanical methods** and **methods using "random digits**." Mechanical methods rely for their effectiveness on symmetry and/or thorough mixing in a physical randomizing device. So to speak, the slips of paper in the hat need to be of the same size and well scrambled before sample selection begins.

The first Vietnam-era U.S. draft lottery was a famous case in which adequate care was not taken to ensure appropriate operation of a mechanical randomizing device. Birthdays were supposed to be assigned priority numbers 1 through 366 in a "random" way. However, it was clear after the fact that balls representing birth dates were placed into a bin by months, and the bin was poorly mixed. When the balls were drawn out, birth dates near the end of the year received a disproportionately large share of the low draft numbers. In the present terminology, the first five dates out of the bin should *not* have been thought of as a simple random sample of size 5. Those who operate games of chance more routinely make it their business to know (via the collection of appropriate data) that their mechanical devices are operating in a more random manner.

Using random digits to do sampling implicitly relies for "randomness" on the appropriateness of the method used to generate those digits. *Physical random processes* like radioactive decay and *pseudorandom number generators* (complicated recursive numerical algorithms) are the most common sources of random digits. Until fairly recently, it was common to record such digits in printed tables. Table B.1 consists of random digits (originally generated by a physical random process). The first five rows of this table are reproduced in Table 2.2 for use in this section.

In making a random digit table, the intention is to use a method guaranteeing that a priori

1. each digit 0 through 9 has the same chance of appearing at any particular location in the table one wants to consider, and
2. knowledge of which digit will occur at a given location provides no help in predicting which one will appear at another.

In a random digit table, condition 1 should typically be reflected in roughly equal representation of the 10 digits, and condition 2 in the lack of obvious internal patterns in the table.

Random digit tables and simple random sampling

For populations that can easily be labeled with consecutive numbers, the following steps can be used to synthetically draw items out of a hat one at a time—to draw a simple random sample using a table like Table 2.2.

Step 1 For a population of N objects, determine the number of digits in N (for example, $N = 1291$ is a four-digit number). Call this number M and assign each item in the population a different M-digit label.

Step 2 Move through the table left to right, top to bottom, M digits at a time, beginning from where you left off in last using the table, and choose objects from the population by means of their associated labels until n have been selected.

Step 3 In moving through the table according to step 2, ignore labels that have not been assigned to items in the population and any that would indicate repeat selection of an item.

Table 2.2
Random Digits

12159	66144	05091	13446	45653	13684	66024	91410	51351	22772
30156	90519	95785	47544	66735	35754	11088	67310	19720	08379
59069	01722	53338	41942	65118	71236	01932	70343	25812	62275
54107	58081	82470	59407	13475	95872	16268	78436	39251	64247
99681	81295	06315	28212	45029	57701	96327	85436	33614	29070

12159	66144	05091	13446	45653	13684	66024	91410	51351	22772
30156	90519	95785	47544	66735	35754	11088	67310	19720	08379

Figure 2.3 Use of a random digit table

As an example of how this works, consider selecting a simple random sample of 25 members of a hypothetical population of 80 objects. One first determines that 80 is an $M = 2$-digit number and therefore labels items in the population as 01, 02, 03, 04, . . . , 77, 78, 79, 80 (labels 00 and 81 through 99 are not assigned). Then, if Table 2.2 is being used for the first time, begin in the upper left corner and proceed as indicated in Figure 2.3. Circled numbers represent selected labels, Xs indicate that the corresponding label has not been assigned, and slash marks indicate that the corresponding item has already entered the sample. As the final item enters the sample, the stopping point is marked with a penciled hash mark. Movement through the table is resumed at that point the next time the table is used.

Any predetermined systematic method of moving through the table could be substituted in place of step 2. One could move down columns instead of across rows, for example. It is useful to make the somewhat arbitrary choice of method in step 2 for the sake of classroom consistency.

Statistical or spreadsheet software and simple random sampling

With the wide availability of personal computers, random digit tables have become largely obsolete. That is, random numbers can be generated "on the spot" using statistical or spreadsheet software. In fact, it is even easy to have such software automatically do something equivalent to steps 1 through 3 above, selecting a simple random sample of n of the numbers 1 to N. For example, Printout 1 was produced using the MINITAB™ statistical package. It illustrates the selection of $n = 25$ members of a population of $N = 80$ objects. The numbers 1 through 80 are placed into the first column of a worksheet (using the routine under the "Calc/Make Patterned Data/Simple Set of Numbers" menu). Then 25 of them are selected using MINITAB's pseudorandom number generation capability (located under the "Calc/Random Data/Sample from Columns" menu). Finally, those 25 values (the results beginning with 56 and ending with 72) are printed out (using the routine under the "Manip/Display Data" menu).

Printout 1 Random Selection of 25 Objects from a Population of 80 Objects

```
MTB > Set C1
DATA>   1( 1 : 80 / 1 )1
DATA>   End.
MTB > Sample 25 C1 C2.
MTB > Print C2.

Data Display

C2
       56      74      43      61      80      22      30      67      35       7
       10      69      19      49       8      45       3      37      21      17
        2      12       9      14      72
```

Regardless of how Definition 1 is implemented, several comments about the method are in order. First, it must be admitted that simple random sampling meets the original objective of providing representative samples only in some average or long-run sense. It is possible for the method to produce particular realizations that are horribly unrepresentative of the corresponding population. A simple random sample of 20 out of 80 axles could turn out to consist of those with the smallest diameters. But this doesn't happen often. On the average, a simple random sample will faithfully portray the population. Definition 1 is a statement about a method, not a guarantee of success on a particular application of the method.

Second, it must also be admitted that there is no guarantee that it will be an easy task to make the physical selection of a simple random sample. Imagine the pain of retrieving 5 out of a production run of 1,000 microwave ovens stored in a warehouse. It would probably be a most unpleasant job to locate and gather 5 ovens corresponding to randomly chosen serial numbers to, for example, carry to a testing lab.

But the virtues of simple random sampling usually outweigh its drawbacks. For one thing, it is an **objective method** of sample selection. An engineer using it is protected from conscious and subconscious human bias. In addition, the method **interjects probability** into the selection process in what turns out to be a manageable fashion. As a result, the quality of information from a simple random sample can be quantified. Methods of formal statistical inference, with their resulting conclusions ("I am 95% sure that . . ."), can be applied when simple random sampling is used.

It should be clear from this discussion that there is nothing mysterious or magical about simple random sampling. We sometimes get the feeling while reading student projects (and even some textbooks) that the phrase *random sampling* is used (even in analytical rather than enumerative contexts) to mean "magically OK sampling" or "sampling with magically universally applicable results." Instead, simple random sampling is a concrete methodology for enumerative studies. It is generally about the best one available without a priori having intimate knowledge of the population.

Section 2 Exercises

1. For the sake of exercise, treat the runout values for 38 laid gears (given in Table 1.1) as a population of interest, and using the random digit table (Table B.1), select a simple random sample of 5 of these runouts. Repeat this selection process a total of four different times. (Begin the selection of the first sample at the upper left of the table and proceed left to right and top to bottom.) Are the four samples identical? Are they each what you would call "representative" of the population?

2. Repeat Exercise 1 using statistical or spreadsheet software to do the random sampling.

3. Explain briefly why in an enumerative study, a simple random sample is or is not guaranteed to be representative of the population from which it is drawn.

2.3 Principles for Effective Experimentation

Purposely introducing changes into an engineering system and observing what happens as a result (i.e., experimentation) is a principal way of learning how the system works. Engineers meet such a variety of experimental situations that it is impossible to give advice that will be completely relevant in all cases. But it is possible to raise some general issues, which we do here. The discussion in this section is organized under the headings of

1. taxonomy of variables,
2. handling extraneous variables,
3. comparative study,
4. replication, and
5. allocation of resources.

Then Section 2.4 discusses a few generic experimental frameworks for planning a specific experiment.

2.3.1 Taxonomy of Variables

One of the hard realities of experiment planning is the multidimensional nature of the world. There are typically many characteristics of system performance that the engineer would like to improve and many variables that might influence them. Some terminology is needed to facilitate clear thinking and discussion in light of this complexity.

Definition 2

A **response variable** in an experiment is one that is monitored as characterizing system performance/behavior.

A response variable is a system output. Some variables that potentially affect a response of interest are **managed** by the experimenter.

Definition 3

A **supervised (or managed) variable** in an experiment is one over which an investigator exercises power, choosing a setting or settings for use in the study. When a supervised variable is held constant (has only one setting), it is called a **controlled variable**. And when a supervised variable is given several different settings in a study, it is called an **experimental variable**.

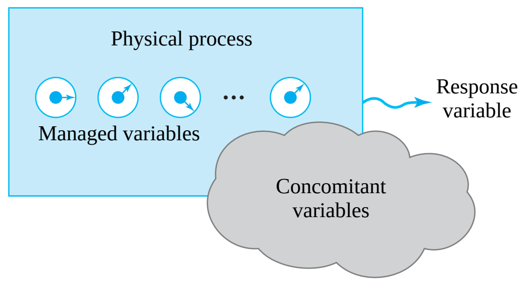

Figure 2.4 Variables in an experiment

Some of the variables that are neither primary responses nor managed in an experiment will nevertheless be observed.

Definition 4

A **concomitant (or accompanying) variable** in an experiment is one that is observed but is neither a primary response variable nor a managed variable. Such a variable can change in reaction to either experimental or unobserved causes and may or may not itself have an impact on a response variable.

Figure 2.4 is an attempt to picture Definitions 2 through 4. In it, the physical process somehow produces values of a response. "Knobs" on the process represent managed variables. Concomitant variables are floating about as part of the experimental environment without being its main focus.

Example 7
(*Example 6, Chapter 1, revisited—p. 15*)

Variables in a Wood Joint Strength Experiment

Dimond and Dix experimented with three different woods and three different glues, investigating joint strength properties. Their primary interest was in the effects of experimental variables Wood Type and Glue Type on two observed response variables, joint strength in a tension test and joint strength in a shear test.

In addition, they recognized that strengths were probably related to the variables Drying Time and Pressure applied to the joints while drying. Their method of treating the nine wood/glue combinations fairly with respect to the Time and Pressure variables was to manage them as controlled variables, trying to hold them essentially constant for all the joints produced.

Some of the variation the students observed in strengths could also have originated in properties of the particular specimens glued, such as moisture content. In fact, this variable was not observed in the study. But if the students had had some way of measuring it, moisture content might have provided extra insight into how the wood/glue combinations behave. It would have been a potentially informative concomitant variable.

2.3.2 Handling Extraneous Variables

In planning an experiment, there are always variables that could influence the responses but which are not of practical interest to the experimenter. The investigator may recognize some of them as influential but not even think of others. Those that are recognized may fail to be of primary interest because there is no realistic way of exercising control over them or compensating for their effects outside of the experimental environment. So it is of little practical use to know exactly how changes in them affect the system.

But completely ignoring the existence of such **extraneous variables** in experiment planning can needlessly cloud the perception of the effects of factors that *are* of interest. Several methods can be used in an active attempt to avoid this loss of information. These are to manage them (for experimental purposes) as **controlled variables** (recall Definition 3) or as **blocking variables**, or to attempt to balance their effects among process conditions of interest through **randomization**.

Control of extraneous variables

When choosing to control an extraneous variable in an experiment, both the pluses and minuses of that choice should be recognized. On the one hand, the control produces a homogeneous environment in which to study the effects of the primary experimental variables. In some sense, a portion of the background noise has been eliminated, allowing a clearer view of how the system reacts to changes in factors of interest. On the other hand, system behavior at other values of the controlled variable cannot be projected on the firm basis of data. Instead, projections must be based on the basis of *expert opinion* that what is seen experimentally will prove true more generally. Engineering experience is replete with examples where what worked fine in a laboratory (or even a pilot plant) was much less dependable in subsequent experience with a full-scale facility.

Example 7 *(continued)*

The choice Dimond and Dix made to control Drying Time and the Pressure provided a uniform environment for comparing the nine wood/glue combinations. But strictly speaking, they learned only about joint behavior under their particular experimental Time and Pressure conditions.

To make projections for other conditions, they had to rely on their experience and knowledge of material science to decide how far the patterns they observed were likely to extend. For example, it may have been reasonable to expect what they observed to also hold up for any drying time at least as long as the experimental one, because of expert knowledge that the experimental time was sufficient for the joints to fully set. But such extrapolation is based on other than statistical grounds.

An alternative to controlling extraneous variables is to handle them as experimental variables, including them in study planning at several different levels. Notice that this really amounts to applying the notion of control *locally*, by creating not one but several (possibly quite different) homogeneous environments in which to compare levels of the primary experimental variables. The term *blocking* is often used to refer to this technique.

Blocking extraneous variables

Definition 5

A **block** of experimental units, experimental times of observation, experimental conditions, etc. is a homogeneous group within which different levels of primary experimental variables can be applied and compared in a relatively uniform environment.

Example 7 (*continued*)

Consider embellishing a bit on the gluing study of Dimond and Dix. Imagine that the students were uneasy about two issues, the first being the possibility that surface roughness differences in the pieces to be glued might mask the wood/glue combination differences of interest. Suppose also that because of constraints on schedules, the strength testing was going to have to be done in two different sessions a day apart. Measuring techniques or variables like ambient humidity might vary somewhat between such periods. How might such potential problems have been handled?

Blocking is one way. If the specimens of each wood type were separated into relatively rough and relatively smooth groups, the factor Roughness could have then served as an experimental factor. Each of the glues could have been used the same number of times to join both rough and smooth specimens of each species. This would set up comparison of wood/glue combinations separately for rough and for smooth surfaces.

In a similar way, half the testing for each wood/glue/roughness combination might have been done in each testing session. Then, any consistent differences between sessions could be identified and prevented from clouding the comparison of levels of the primary experimental variables. Thus, Testing Period could have also served as a blocking variable in the study.

Randomization and extraneous variables

Experimenters usually hope that by careful planning they can account for the most important extraneous variables via control and blocking. But not all extraneous variables can be supervised. There are an essentially infinite number, most of which cannot even be named. And there is a way to take out insurance against the possibility that major extraneous variables get overlooked and then produce effects that are mistaken for those of the primary experimental variables.

Definition 6

Randomization is the use of a randomizing device or table of random digits at some point where experimental protocol is not already dictated by the specification of values of the supervised variables. Often this means that experimental objects (or units) are divided up between the experimental conditions at random. It can also mean that the order of experimental testing is randomly determined.

The goal of randomization is to average between sets of experimental conditions the effects of all unsupervised extraneous variables. To put it differently, sets of experimental conditions are treated fairly, giving them equal opportunity to shine.

Example 8
(Example 1, Chapter 1, revisited—p. 2)

Randomization in a Heat Treating Study

P. Brezler, in his "Heat Treating" article, describes a very simple randomized experiment for comparing the effects on thrust face runout of laying versus hanging gears. The variable Loading Method was the primary experimental variable. Extraneous variables Steel Heat and Machining History were controlled by experimenting on 78 gears from the same heat code, machined as a lot. The 78 gears were broken at random into two groups of 39, one to be laid and the other to be hung. (Note that Table 1.1 gives only 38 data points for the laid group. For reasons not given in the article, one laid gear was dropped from the study.)

Although there is no explicit mention of this in the article, the principle of randomization could have been (and perhaps was) carried a step further by making the runout measurements in a random order. (This means choosing gears 01 through 78 one at a time at random to measure.) The effect of this randomization would have been to protect the investigator from clouding the comparison of heat treating methods with possible unexpected and unintended changes in measurement techniques. Failing to randomize and, for example, making all the laid measurements before the hung measurements, would allow unintended changes in measurement technique to appear in the data as differences between the two loading methods. (Practice with measurement equipment might, for example, increase precision and make later runouts appear to be more uniform than early ones.)

Example 7
(continued)

Dimond and Dix took the notion of randomization to heart in their gluing study and, so to speak, randomized everything in sight. In the tension strength testing for a given type of wood, they glued $.5'' \times .5'' \times 3''$ blocks to a $.75'' \times 3.5'' \times 31.5''$ board of the same wood type, as illustrated in Figure 2.5.

Each glue was used for three joints on each type of wood. In order to deal with any unpredicted differences in material properties (e.g., over the extent of the board) or unforeseen differences in loading by the steel strap used to provide pressure on the joints, etc., the students randomized the order in which glue was applied and the blocks placed along the base board. In addition, when it came time to do the strength testing, that was carried out in a randomly determined order.

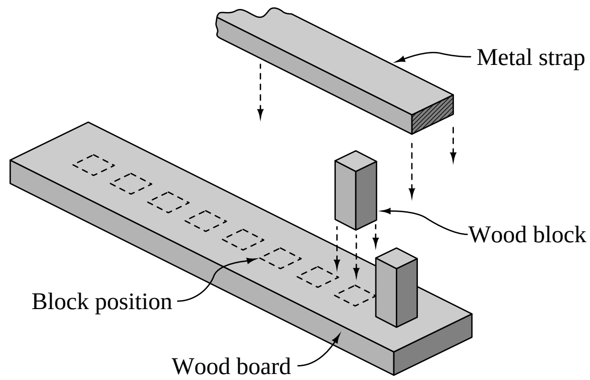

Figure 2.5 Gluing method for a single wood type

Simple random sampling in enumerative studies is only guaranteed to be effective in an average or long-run sense. Similarly, randomization in experiments will not prove effective in averaging the effects of extraneous variables between settings of experimental variables every time it is used. Sometimes an experimenter will be unlucky. But the methodology is objective, effective on the average, and about the best one can do in accounting for those extraneous variables that will not be managed.

2.3.3 Comparative Study

Statistical engineering studies often involve more than a single sample. They usually involve comparison of a number of settings of process variables. This is true not only because there may be many options open to an engineer in a given situation, but for other reasons as well.

Even in experiments where there is only a single new idea or variation on standard practice to be tried out, it is a good idea to make the study **comparative** (and therefore to involve more than one sample). Unless this is done, there is no really firm basis on which to say that any effects observed come from the new conditions under study rather than from unexpected extraneous sources. If standard yield for a chemical process is 63.2% and a few runs of the process with a supposedly improved catalyst produce a mean yield of 64.8%, it is not completely safe to attribute the difference to the catalyst. It could be caused by a number of things, including miscalibration of the measurement system. But suppose a few experimental runs are taken for both the standard and the new catalysts. If these produce two samples with small internal variation and (for example) a difference of 1.6% in mean yields, that difference is more safely attributed to a difference in the catalysts.

Example 8 *(continued)*

In the gear loading study, hanging was the standard method in use at the time of the study. From its records, the company could probably have located some values for thrust face runout to use as a baseline for evaluating the laying method. But the choice to run a comparative study, including both laid and hung gears, put the engineer on firm ground for drawing conclusions about the new method.

A second usage of "control"

In a potentially confusing use of language, the word *control* is sometimes used to mean the practice of including a standard or no-change sample in an experiment for comparison purposes. (Notice that this is not the usage in Definition 3.) When a *control group* is included in a medical study to verify the effectiveness of a new drug, that group is either a standard-treatment or no-treatment group, included to provide a solid basis of comparison for the new treatment.

2.3.4 Replication

In much of what has been said so far, it has been implicit that having more than one observation for a given setting of experimental variables is a good idea.

Definition 7

Replication of a setting of experimental variables means carrying through the whole process of adjusting values for supervised variables, making an experimental "run," and observing the results of that run—more than once. Values of the responses from replications of a setting form the (single) sample corresponding to the setting, which one hopes represents typical process behavior at that setting.

Purposes of replication

The idea of replication is fundamental in experimentation. **Reproducibility of results** is important in both science and engineering practice. Replication helps establish this, protecting the investigator from unconscious blunders and validating or confirming experimental conclusions.

But replication is not only important for establishing that experimental results are reproducible. It is also essential to quantifying the *limits* of that reproducibility—that is, for getting an idea of the size of experimental error. Even under a fixed setting of supervised variables, repeated experimental runs typically will not produce exactly the same observations. The effects of unsupervised variables and measurement errors produce a kind of **baseline variation**, or background noise. Establishing the magnitude of this variation is important. It is only against this background that one can judge whether an apparent effect of an experimental variable is big enough to establish it as clearly real, rather than explainable in terms of background noise.

When planning an experiment, the engineer must think carefully about what kind of repetition will be included. Definition 7 was written specifically to suggest that simply remeasuring an experimental unit does not amount to real replication. Such repetition will capture measurement error, but it ignores the effects of (potentially changing) unsupervised variables. It is a common mistake in logic to seriously underestimate the size of experimental error by failing to adopt a broad enough view of what should be involved in replication, settling instead for what amounts to remeasurement.

Example 9

Replication and Steel Making

A former colleague once related a consulting experience that went approximately as follows. In studying the possible usefulness of a new additive in a type of steel, a metallurgical engineer had one heat (batch) of steel made with the additive and one without. Each of these was poured into ingots. The metallurgist then selected some ingots from both heats, had them cut into pieces, and selected some pieces from the ingots, ultimately measuring a property of interest on these pieces and ending up with a reasonably large amount of data. The data from the heat with additive showed it to be clearly superior to the no-additive heat. As a result, the existing production process was altered (at significant expense) and the new additive incorporated. Unfortunately, it soon became apparent that the alteration to the process had actually degraded the properties of the steel.

The statistician was (only at this point) called in to help figure out what had gone wrong. After all, the experimental results, based on a large amount of data, had been quite convincing, hadn't they?

The key to understanding what had gone wrong was the issue of replication. In a sense, there was none. The metallurgist had essentially just remeasured the same two physical objects (the heats) many times. In the process, he had learned quite a bit about the two particular heats in the study but very little about all heats of the two types. Apparently, extraneous and uncontrolled foundry variables were producing large heat-to-heat variability. The metallurgist had mistaken an effect of this fluctuation for an improvement due to the new additive. The metallurgist had no notion of this possibility because he had not replicated the with-additive and without-additive settings of the experimental variable.

Example 10

Replication and Paper Airplane Testing

Beer, Dusek, and Ehlers completed a project comparing the Kline-Fogelman and Polish Frisbee paper airplane designs on the basis of flight distance under a number of different conditions. In general, it was a carefully done project. However, replication was a point on which their experimental plan was extremely weak. They made a number of trials for each plane under each set of experimental conditions, but only one Kline-Fogelman prototype and one Polish Frisbee prototype were used throughout the study. The students learned quite a bit about the prototypes in hand but possibly much less about the two designs. If their purpose was to pick a winner between the two prototypes, then perhaps the design of their study was appropriate. But if the purpose was to make conclusions about planes

Example 10 (*continued*)

"like" the two used in the study, they needed to make and test several prototypes for each design.

ISU Professor Emeritus L. Wolins calls the problem of identifying what constitutes replication in an experiment the **unit of analysis problem**. There must be replication of the basic experimental unit or object. The agriculturalist who, in order to study pig blood chemistry, takes hundreds of measurements per hour on one pig, has a (highly multivariate) sample of size 1. The pig is the unit of analysis.

Without proper replication, one can only hope to be lucky. If experimental error is small, then accepting conclusions suggested by samples of size 1 will lead to correct conclusions. But the problem is that without replication, one usually has little idea of the size of that experimental error.

2.3.5 Allocation of Resources

Experiments are done by people and organizations that have finite time and money. Allocating those resources and living within the constraints they impose is part of experiment planning. The rest of this section makes several points in this regard.

First, real-world investigations are often most effective when approached **sequentially**, the planning for each stage building upon what has been learned before. The classroom model of planning and/or executing a single experiment is more a result of constraints inherent in our methods of teaching than a realistic representation of how engineering problems are solved. The reality is most often iterative in nature, involving a series of related experiments.

This being the case, one can not use an entire experimental budget on the first pass of a statistical engineering study. Conventional wisdom on this matter is that no more than 20–25% of an experimental budget should be allocated to the first stage of an investigation. This leaves adequate resources for follow-up work built on what is learned initially.

Second, what is easy to do (and therefore usually cheap to do) should not dictate completely what is done in an experiment. In the context of the steel formula development study of Example 9, it seems almost certain that one reason the metallurgist chose to get his "large sample sizes" from pieces of ingots rather than from heats is that it was easy and cheap to get many measurements in that way. But in addition to failing to get absolutely crucial replication and thus botching the study, he probably also grossly overmeasured the two heats.

A final remark is an amplification of the discussion of sample size in Section 2.1. That is, minimum experimental resource requirements are dictated in large part by the magnitude of effects of engineering importance in comparison to the magnitude of experimental error. The larger the effects in comparison to the error (the larger the signal-to-noise ratio), the smaller the sample sizes required, and thus the fewer the resources needed.

Section 3 Exercises

1. Consider again the paper airplane study from Exercise 1 of Section 2.1. Describe some variables that you would want to control in such a study. What are the response and experimental variables that would be appropriate in this context? Name a potential concomitant variable here.
2. In general terms, what is the trade-off that must be weighed in deciding whether or not to control a variable in a statistical engineering study?
3. In the paper airplane scenario of Exercise 1 of Section 2.1, if (because of schedule limitations, for example) two different team members will make the flight distance measurements, discuss how the notion of blocking might be used.
4. Again using the paper airplane scenario of Exercise 1 of Section 2.1, suppose that two students are each going to make and fly one airplane of each of the $2^3 = 8$ possible types once. Employ the notion of randomization and Table B.1 and develop schedules for Tom and Juanita to use in their flight testing. Explain how the table was used.
5. Continuing the paper airplane scenario of Exercise 1 of Section 2.1, discuss the pros and cons of Tom and Juanita flying each of their own eight planes twice, as opposed to making and flying two planes of each of the eight types, one time each.
6. Random number tables are sometimes used in the planning of both enumerative and analytical/experimental studies. What are the two different terminologies employed in these different contexts, and what are the different purposes behind the use of the tables?
7. What is blocking supposed to accomplish in an engineering experiment?
8. What are some purposes of replication in a statistical engineering study?
9. Comment briefly on the notion that in order for a statistical engineering study to be statistically proper, one should know before beginning data collection exactly how an entire experimental budget is to be spent. (Is this, in fact, a correct idea?)

2.4 Some Common Experimental Plans

In previous sections, experimentation has been discussed in general terms, and the subtlety of considerations that enter the planning of an effective experiment has been illustrated. It should be obvious that any exposition of standard experimental "plans" can amount only to a discussion of standard "skeletons" around which real plans can be built. Nevertheless, it is useful to know something about such skeletons. In this section, so-called completely randomized, randomized complete block, and incomplete block experimental plans are considered.

2.4.1 Completely Randomized Experiments

Definition 8

A **completely randomized experiment** is one in which all experimental variables are of primary interest (i.e., none are included only for purposes of blocking), and randomization is used at every possible point of choosing the experimental protocol.

Notice that this definition says nothing about how the combinations of settings of experimental variables included in the study are structured. In fact, they may be essentially unstructured or produce data with any of the structures discussed in Section 1.2. That is, there are completely randomized one-factor, factorial, and fractional factorial experiments. The essential point in Definition 8 is that all else is randomized except what is restricted by choice of which combinations of levels of experimental variables are to be used in the study.

Paraphrase of the definition of complete randomization

Although it doesn't really fit every situation (or perhaps even most) in which the term *complete randomization* is appropriate, language like the following is commonly used to capture the intent of Definition 8. "Experimental units (objects) are allocated at random to the treatment combinations (settings of experimental variables). Experimental runs are made in a randomly determined order. And any post-facto measuring of experimental outcomes is also carried out in a random order."

Example 11

Complete Randomization in a Glass Restrengthening Study

Bloyer, Millis, and Schibur studied the restrengthening of damaged glass through etching. They investigated the effects of two experimental factors—the Concentration of hydrofluoric acid in an etching bath and the Time spent in the etching bath—on the resulting strength of damaged glass rods. (The rods had been purposely scratched in a 1″ region near their centers by sandblasting.) Strengths were measured using a three-point bending method on a 20 kip MTS machine.

The students decided to run a 3×3 factorial experiment. The experimental levels of Concentration were 50%, 75%, and 100% HF, and the levels of Time employed were 30 sec, 60 sec, and 120 sec. There were thus nine treatment combinations, as illustrated in Figure 2.6.

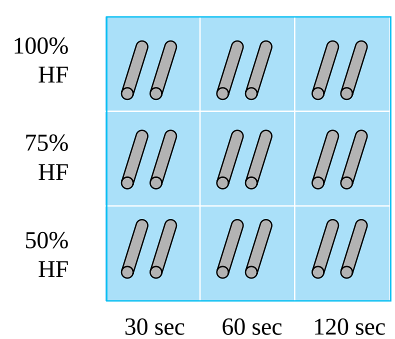

Figure 2.6 Nine combinations of three levels of concentration and three levels of time

The students decided that 18 scratched rods would be allocated—two apiece to each of the nine treatment combinations—for testing. Notice that this could be done at random by labeling the rods 01–18, placing numbered slips of paper in a hat, mixing, drawing out two for 30 sec and 50% concentration, then drawing out two for 30 sec and 75% concentration, etc.

Having determined at random which rods would receive which experimental conditions, the students could again have used the slips of paper to randomly determine an etching order. And a third use of the slips of paper to determine an order of strength testing would have given the students what most people would call a completely randomized 3×3 factorial experiment.

Example 12

Complete Randomization and a Study of the Flight of Golf Balls

G. Gronberg studied drive flight distances for 80, 90, and 100 compression golf balls, using 10 balls of each type in his experiment. Consider what complete randomization would entail in such a study (involving the single factor Compression).

Notice that the paraphrase of Definition 8 is not particularly appropriate to this experimental situation. The levels of the experimental factor are an intrinsic property of the experimental units (balls). There is no way to randomly divide the 30 test balls into three groups and "apply" the treatment levels 80, 90, and 100 compression to them. In fact, about the only obvious point at which randomization could be employed in this scenario is in the choice of an order for hitting the 30 test balls. If one numbered the test balls 01 through 30 and used a table of random digits to pick a hitting order (by choosing balls one at a time without replacement), most people would be willing to call the resulting test a completely randomized one-factor experiment.

Randomization is a good idea. Its virtues have been discussed at some length. So it would be wise to point out that using it can sometimes lead to practically unworkable experimental plans. Dogmatic insistence on complete randomization can in some cases be quite foolish and unrealistic. Changing experimental variables according to a completely randomly determined schedule can sometimes be exceedingly inconvenient (and therefore expensive). If the inconvenience is great and the fear of being misled by the effects of extraneous variables is relatively small, then backing off from complete to partial randomization may be the only reasonable course of action. But when choosing not to randomize, the implications of that choice must be carefully considered.

Example 11
(*continued*)

Consider an embellishment on the glass strengthening scenario, where an experimenter might have access to only a single container to use for a bath and/or have only a limited amount of hydrofluoric acid.

Example 11 *(continued)*

From the discussion of replication in the previous section and present considerations of complete randomization, it would seem that the purest method of conducting the study would be to make a new dilution of HF for each of the rods as its turn comes for testing. But this would be time-consuming and might require more acid than was available.

If the investigator had three containers to use for baths but limited acid, an alternative possibility would be to prepare three different dilutions, one 100%, one 75%, and one 50% dilution. A given dilution could then be used in testing all rods assigned to that concentration. Notice that this alternative allows for a randomized order of testing, but it introduces some question as to whether there is "true" replication.

Taking the resource restriction idea one step further, notice that even if an investigator could afford only enough acid for making one bath, there is a way of proceeding. One could do all 100% concentration testing, then dilute the acid and do all 75% testing, then dilute the acid again and do all 50% testing. The resource restriction would not only affect the "purity" of replication but also prevent complete randomization of the experimental order. Thus, for example, any unintended effects of increased contamination of the acid (as more and more tests were made using it) would show up in the experimental data as indistinguishable from effects of differences in acid concentration.

To choose intelligently between complete randomization (with "true" replication) and the two plans just discussed, the real severity of resource limitations would have to be weighed against the likelihood that extraneous factors would jeopardize the usefulness of experimental results.

2.4.2 Randomized Complete Block Experiments

Definition 9

A **randomized complete block experiment** is one in which at least one experimental variable is a blocking factor (not of primary interest to the investigator); and within each block, every setting of the primary experimental variables appears at least once; and randomization is employed at all possible points where the exact experimental protocol is determined.

A helpful way to think of a randomized complete block experiment is as a collection of completely randomized studies. Each of the blocks yields one of the component studies. Blocking provides the simultaneous advantages of homogeneous environments for studying primary factors and breadth of applicability of the results.

Definition 9 (like Definition 8) says nothing about the structure of the settings of primary experimental variables included in the experiment. Nor does it say anything about the structure of the blocks. It is possible to design experiments where experimental combinations of primary variables have one-factor, factorial, or fractional factorial structure, and at the same time the experimental combinations of blocking variables also have one of these standard structures. The essential points of Definition 9 are the *completeness* of each block (in the sense that it contains each setting of the primary variables) and the *randomization* within each block. The following two examples illustrate that depending upon the specifics of a scenario, Definition 9 can describe a variety of experimental plans.

Example 12 (*continued*)

As actually run, Gronberg's golf ball flight study amounted to a randomized complete block experiment. This is because he hit and recorded flight distances for all 30 balls on six different evenings (over a six-week period). Note that this allowed him to have (six different) homogeneous conditions under which to compare the flight distances of balls having 80, 90, and 100 compression. (The blocks account for possible changes over time in his physical condition and skill level as well as varied environmental conditions.)

Notice the structure of the data set that resulted from the study. The settings of the single primary experimental variable Compression combined with the levels of the single blocking factor Day to produce a 3×6 factorial structure for 18 samples of size 10, as pictured in Figure 2.7.

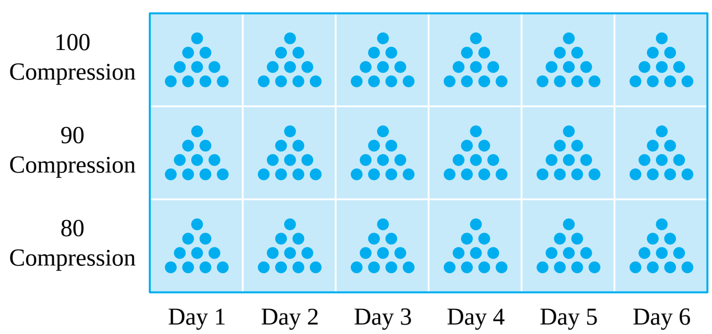

Figure 2.7 18 combinations of compression and day

Example 13 (*Example 2, Chapter 1, revisited—pp. 6, 13*)

Blocking in a Pelletizing Experiment

Near the end of Section 1.2, the notion of a fractional factorial study was illustrated in the context of a hypothetical experiment on a pelletizing machine. The factors Volume, Flow, and Mixture were of primary interest. Table 1.3 is reproduced here as Table 2.3, listing four (out of eight possible) combinations of two levels each of the primary experimental variables, forming a fractional factorial arrangement.

Consider a situation where two different operators can make four experimental runs each on two consecutive days. Suppose further that Operator and Day are blocking factors, their combinations giving four blocks, within which the four combinations listed in Table 2.3 are run in a random order. This ends

Example 13
(continued)

Table 2.3
Half of a 2^3 Factorial

Volume	Flow	Mixture
high	current	no binder
low	manual	no binder
low	current	binder
high	manual	binder

up as a randomized complete block experiment in which the blocks have 2×2 factorial structure and the four combinations of primary experimental factors have a fractional factorial structure.

There are several ways to think of this plan. For one, by temporarily ignoring the structure of the blocks and combinations of primary experimental factors, it can be considered a 4×4 factorial arrangement of samples of size 1, as is illustrated in Figure 2.8. But from another point of view, the combinations under discussion (listed in Table 2.4) have fractional factorial structure of their own, representing a (not particularly clever) choice of 16 out of $2^5 = 32$ different possible combinations of the two-level factors Operator, Day, Volume, Flow, and Mixture. (The lines in Table 2.4 separate the four blocks.) A better use of 16 experimental runs in this situation (at least from the perspective that the combinations in Table 2.4 have their own fractional factorial structure) will be discussed next.

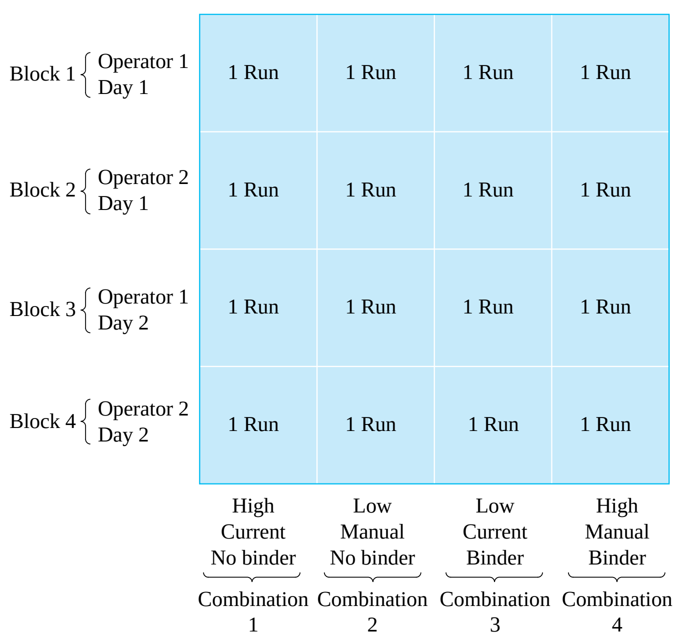

Figure 2.8 16 combinations of blocks and treatments

Table 2.4
Half of a 2^3 Factorial Run Once in Each of Four Blocks

Operator	Day	Volume	Flow	Mixture
1	1	high	current	no binder
1	1	low	manual	no binder
1	1	low	current	binder
1	1	high	manual	binder
2	1	high	current	no binder
2	1	low	manual	no binder
2	1	low	current	binder
2	1	high	manual	binder
1	2	high	current	no binder
1	2	low	manual	no binder
1	2	low	current	binder
1	2	high	manual	binder
2	2	high	current	no binder
2	2	low	manual	no binder
2	2	low	current	binder
2	2	high	manual	binder

2.4.3 Incomplete Block Experiments (*Optional*)

In many experimental situations where blocking seems attractive, physical constraints make it impossible to satisfy Definition 9. This leads to the notion of incomplete blocks.

Definition 10

An **incomplete** (usually randomized) **block experiment** is one in which at least one experimental variable is a blocking factor and the assignment of combinations of levels of primary experimental factors to blocks is such that not every combination appears in every block.

Example 13 (*continued*)

In Section 1.2, the pelletizing machine study examined all eight possible combinations of Volume, Flow, and Mixture. These are listed in Table 2.5. Imagine that only half of these eight combinations can be run on a given day, and there is some fear that daily environmental conditions might strongly affect process performance. How might one proceed?

There are then two blocks (days), each of which will accommodate four runs. Some possibilities for assigning runs to blocks would clearly be poor. For example, running combinations 1 through 4 on the first day and 5 through 8 on

Example 13
(continued)

Table 2.5
Combinations in a 2^3 Factorial Study

Combination Number	Volume	Flow	Mixture
1	low	current	no binder
2	high	current	no binder
3	low	manual	no binder
4	high	manual	no binder
5	low	current	binder
6	high	current	binder
7	low	manual	binder
8	high	manual	binder

the second would make it impossible to distinguish the effects of Mixture from any important environmental effects.

What turns out to be a far better possibility is to run, say, the four combinations listed in Table 2.3 (combinations 2, 3, 5, and 8) on one day and the others on the next. This is illustrated in Table 2.6. In a well-defined sense (explained in Chapter 8), this choice of an incomplete block plan minimizes the unavoidable clouding of inferences caused by the fact all eight combinations of levels of Volume, Flow, and Mixture cannot be run on a single day.

As one final variation on the pelletizing scenario, consider an alternative that is superior to the experimental plan outlined in Table 2.4: one that involves incomplete blocks. That is, once again suppose that the two-level primary factors Volume, Flow, and Mixture are to be studied in four blocks of four observations, created by combinations of the two-level blocking factors Operator and Day.

Since a total of 16 experimental runs can be made, all eight combinations of primary experimental factors can be included in the study twice (instead of

Table 2.6
A 2^3 Factorial Run in Two Incomplete Blocks

Day	Volume	Flow	Mixture
2	low	current	no binder
1	high	current	no binder
1	low	manual	no binder
2	high	manual	no binder
1	low	current	binder
2	high	current	binder
2	low	manual	binder
1	high	manual	binder

Table 2.7
A Once-Replicated 2^3 Factorial Run in Four Incomplete Blocks

Operator	Day	Volume	Flow	Mixture
1	1	high	current	no binder
1	1	low	manual	no binder
1	1	low	current	binder
1	1	high	manual	binder
2	1	low	current	no binder
2	1	high	manual	no binder
2	1	high	current	binder
2	1	low	manual	binder
1	2	low	current	no binder
1	2	high	manual	no binder
1	2	high	current	binder
1	2	low	manual	binder
2	2	high	current	no binder
2	2	low	manual	no binder
2	2	low	current	binder
2	2	high	manual	binder

including only four combinations four times apiece). To do this, incomplete blocks are required, but Table 2.7 shows a good incomplete block plan. (Again, blocks are separated by lines.)

Notice the symmetry present in this choice of half of the $2^5 = 32$ different possible combinations of the five experimental factors. For example, a full factorial in Volume, Flow, and Mixture is run on each day, and similarly, each operator runs a full factorial in the primary experimental variables.

It turns out that the study outlined in Table 2.7 gives far more potential for learning about the behavior of the pelletizing process than the one outlined in Table 2.4. But again, a complete discussion of this must wait until Chapter 8.

There may be some reader uneasiness and frustration with the "rabbit out of a hat" nature of the examples of incomplete block experiments, since there has been no discussion of how to go about making up a good incomplete block plan. Both the choosing of an incomplete block plan and corresponding techniques of data analysis are advanced topics that will not be developed until Chapter 8. The purpose here is to simply introduce the possibility of incomplete blocks as a useful option in experimental planning.

Section 4 Exercises

1. What standard name might be applied to the experimental plan you developed for Exercise 4 of Section 2.3?
2. Consider an experimental situation where the three factors A, B, and C each have two levels, and it is desirable to make three experimental runs for each of the possible combinations of levels of the factors.
   (a) Select a completely random order of experimentation. Carefully describe how you use Table B.1 or statistical software to do this. Make an ordered list of combinations of levels of the three factors, prescribing which combination should be run first, second, etc.
   (b) Suppose that because of physical constraints, only eight runs can be made on a given day. Carefully discuss how the concept of blocking could be used in this situation when planning which experimental runs to make on each of three consecutive days. What possible purpose would blocking serve?
   (c) Use Table B.1 or statistical software to randomize the order of experimentation within the blocks you described in part (b). (Make a list of what combinations of levels of the factors are to be run on each day, in what order.) How does the method you used here differ from what you did in part (a)?
3. Once more referring to the paper airplane scenario of Exercise 1 of Section 2.1, suppose that only the factors Design and Paper are of interest (all planes will be made without paper clips) but that Tom and Juanita can make and test only two planes apiece. Devise an incomplete block plan for this study that gives each student experience with both designs and both papers. (Which two planes will each make and test?)
4. Again in the paper airplane scenario of Exercise 1 of Section 2.1, suppose that Tom and Juanita each have time to make and test only four airplanes apiece, but that in toto they still wish to test all eight possible types of planes. Develop a sensible plan for doing this. (Which planes should each person test?) You will probably want to be careful to make sure that each person tests two delta wing planes, two construction paper planes, and two paper clip planes. Why is this? Can you arrange your plan so that each person tests each Design/Paper combination, each Design/Loading combination, and each Paper/Loading combination once?
5. What standard name might be applied to the plan you developed in Exercise 4?

2.5 Preparing to Collect Engineering Data

This chapter has raised many of the issues that engineers must consider when planning a statistical study. What is still lacking, however, is a discussion of how to get started. This section first lists and then briefly discusses a series of steps that can be followed in preparing for engineering data collection.

2.5.1 A Series of Steps to Follow

The following is a list of steps that can be used to organize the planning of a statistical engineering study.

PROBLEM DEFINITION

Step 1 Identify the problem to be addressed in general terms.

Step 2 Understand the context of the problem.

Step 3 State in precise terms the objective and scope of the study. (State the questions to be answered.)

STUDY DEFINITION

Step 4 Identify the response variable(s) and appropriate instrumentation.

Step 5 Identify possible factors influencing responses.

Step 6 Decide whether (and if so how) to manage factors that are likely to have effects on the response(s).

Step 7 Develop a detailed data collection protocol and timetable for the first phase of the study.

PHYSICAL PREPARATION

Step 8 Assign responsibility for careful supervision.

Step 9 Identify technicians and provide necessary instruction in the study objectives and methods to be used.

Step 10 Prepare data collection forms and/or equipment.

Step 11 Do a dry run analysis on fictitious data.

Step 12 Write up a "best guess" prediction of the results of the actual study.

These 12 points are listed in a reasonably rational order, but planning any real study may involve departures from the listed order as well as a fair amount of iterating among the steps before they are all accomplished. The need for other steps (like finding funds to pay for a proposed study) will also be apparent in some contexts. Nevertheless, steps 1 through 12 form a framework for getting started.

2.5.2 Problem Definition

Step 1 Identifying the general problem to work on is, for the working engineer, largely a matter of prioritization. An individual engineer's job description and place in an organization usually dictate what problem areas need attention. And far more things could always be done than resources of time and money will permit. So some choice has to be made among the different possibilities.

It is only natural to choose a general topic on the basis of the perceived importance of a problem and the likelihood of solving it (given the available resources). These criteria are somewhat subjective. So, particularly when a project team or other working group must come to consensus before proceeding, even this initial planning step is a nontrivial task. Sometimes it is possible to remove part of the subjectivity and reliance on personal impressions by either examining existing data or commissioning a statistical study of the current state of affairs. For example, suppose members of an engineering project team can name several types of flaws that occur in a mechanical part but disagree about the frequencies or dollar impacts of the flaws. The natural place to begin is to search company records or collect some new data aimed at determining the occurrence rates and/or dollar impacts.

An effective and popular way of summarizing the findings of such a preliminary look at the current situation is through a **Pareto diagram**. This is a bar chart whose vertical axis delineates frequency (or some other measure of impact of system misbehavior) and whose bars, representing problems of various types, have been placed left to right in decreasing order of importance.

Example 14

Maintenance Hours for a Flexible Manufacturing System

Figure 2.9 is an example of a Pareto diagram that represents a breakdown (by craft classification) of the total maintenance hours required in one year on four particular machines in a company's flexible manufacturing system. (This information is excerpted from the ISU M.S. thesis work of M. Patel.) A diagram like Figure 2.9 can be an effective tool for helping to focus attention on the most important problems in an engineering system. Figure 2.9 highlights the fact that (in terms of maintenance hours required) mechanical problems required the most attention, followed by electrical problems.

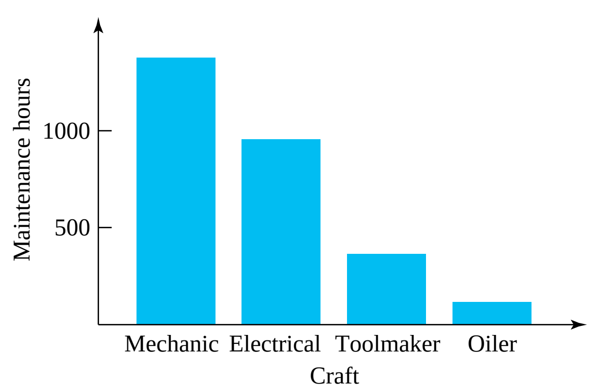

Figure 2.9 Pareto diagram of maintenance hours by craft classification

Step 2

In a statistical engineering study, it is essential to understand the context of the problem. Statistics is no magic substitute for good, hard work learning how a process is configured; what its inputs and environment are; what applicable engineering, scientific, and mathematical theory has to say about its likely behavior; etc. A statistical study is an engineering tool, not a crystal ball. Only when an engineer has studied and asked questions in order to gain expert knowledge about a system is he or she then in a position to decide intelligently what is not known about the system—and thus what data will be of help.

It is often helpful at step 2 to make **flowcharts** describing an ideal process and/or the process as it is currently operating. (Sometimes the comparison of the two is enough in itself to show an engineer how a process should be modified.) During the construction of such a chart, data needs and variables of potential interest can be identified in an organized manner.

Example 15

Work Flow in a Printing Shop

Drake, Lach, and Shadle worked with a printing shop. Before collecting any data, they set about to understand the flow of work through the shop. They made a flowchart similar to Figure 2.10. The flowchart facilitated clear thinking about

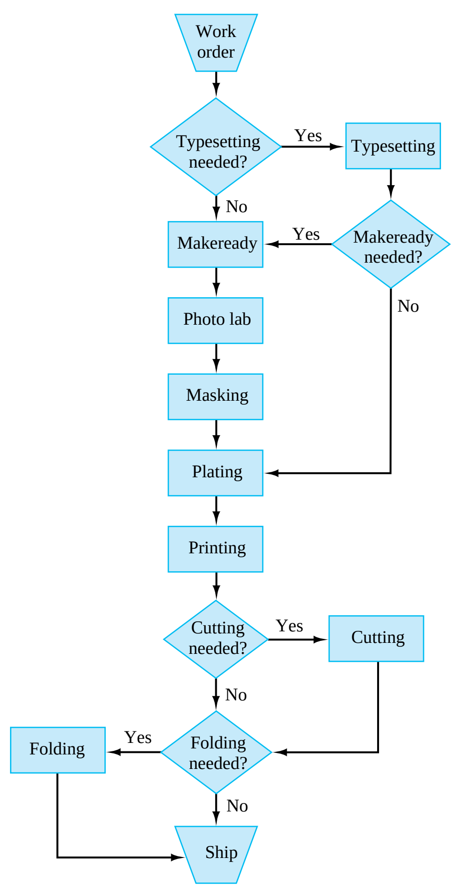

Figure 2.10 Flowchart of a printing process

Example 15 (*continued*)

what might go wrong in the printing process and at what points what data could be gathered in order to monitor and improve process performance.

Step 3 After determining the general arena and physical context of a statistical engineering study, it is necessary to agree on a statement of purpose and scope for the study. An engineering project team assigned to work on a wave soldering process for printed circuit boards must understand the steps in that process and then begin to define what part(s) of the process will be included in the study and what the goal(s) of the study will be. Will flux formulation and application, the actual soldering, subsequent cleaning and inspection, and touch-up all be studied? Or will only some part of this list be investigated? Is system throughput the primary concern, or is it instead some aspect of quality or cost? The sharper a statement of purpose and scope can be made at this point, the easier subsequent planning steps will be.

2.5.3 Study Definition

Step 4 Once one has defined in qualitative terms what it is about an engineering system that is of interest, one must decide how to represent that property (or those properties) in precise terms. That is, one must choose a well-defined response variable (or variables) and decide how to measure it (or them). For example, in a manufacturing context, if "throughput" of a system is of interest, should it be measured in pieces/hour, or conforming pieces/hour, or net profit/hour, or net profit/hour/machine, or in some other way?

Sections 1.3 and 2.1 have already discussed issues that arise in measurement and the formation of operational definitions. All that needs to be added here is that these issues must be faced early in the planning of a statistical engineering study. It does little good to carefully plan a study assuming the existence of an adequate piece of measuring equipment, only to later determine that the organization doesn't own a device with adequate precision and that the purchase of one would cost more than the entire project budget.

Step 5 Identification of variables that may affect system response requires expert knowledge of the process under study. Engineers who do not have hands-on experience with a system can sometimes contribute insights gained from experience with similar systems and from basic theory. But it is also wise (in most cases, essential) to include on a project team several people who have first-hand knowledge of the particular process and to talk extensively with those who work with the system on a regular basis.

Typically, the job of identifying factors of potential importance in a statistical engineering study is a group activity, carried out in brainstorming sessions. It is therefore helpful to have tools for lending order to what might otherwise be an inefficient and disorganized process. One tool that has proved effective is variously known as a **cause-and-effect diagram**, or **fishbone diagram**, or **Ishikawa diagram**.

Example 16

Identifying Potentially Important Variables in a Molding Process

Figure 2.11 shows a cause-and-effect diagram from a study of a molding process for polyurethane automobile steering wheels. It is taken from the paper "Fine Tuning of the Foam System and Optimization of the Process Parameters for the Manufacturing of Polyurethane Steering Wheels Using Reaction Injection Molding by Applying Dr. Taguchi's Method of Design of Experiments" by Vimal Khanna, which appeared in 1985 in the *Third Supplier Symposium on Taguchi Methods*, published by the American Supplier Institute, Inc. Notice how the diagram in Figure 2.11 organizes the huge number of factors possibly affecting

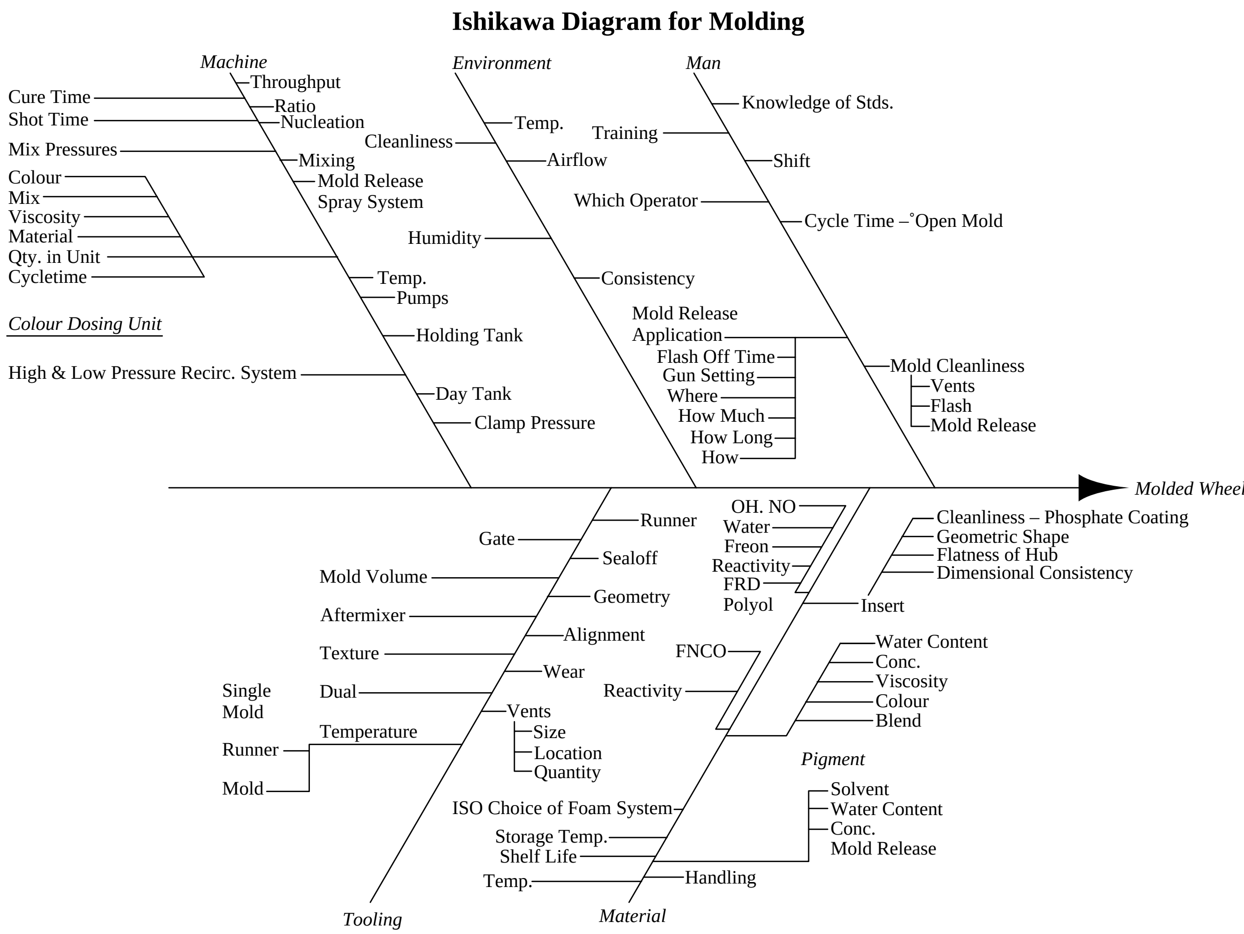

Figure 2.11 Cause and effect diagram for a molding process. From the *Third Symposium on Taguchi Methods*. © Copyright, American Supplier Institute, Dearborn, Michigan (U.S.A.). Reproduced by permission under License No. 930403.

Example 16 (*continued*)

wheel quality. Without some kind of organization, it would be all but impossible to develop anything like a complete list of important factors in a complex situation like this.

Step 6

Armed with (1) a list of variables that might influence the response(s) of interest and some guesses at their relative importance, (2) a solid understanding of the issues raised in Section 2.3, and (3) knowledge of resource and physical constraints and time-frame requirements, one can begin to make decisions about which (if any) variables are to be managed. Experiments have some real advantages over purely observational studies (see Section 1.2). Those must be weighed against possible extra costs and difficulties associated with managing both variables that are of interest and those that are not. The hope is to choose a physically and financially workable set of managed variables in such a way that the aggregate effects of variables not of interest and not managed are not so large as to mask the effects of those variables that *are* of interest.

Step 7

Choosing experimental levels and then combinations for managed variables is part of the task of deciding on a detailed data collection protocol. Levels of controlled and block variables should usually be chosen to be representative of the values that will be met in routine system operation. For example, suppose the amount of contamination in a transmission's hydraulic fluid is thought to affect time to failure when the transmission is subjected to stress testing, where Operating Speed and Pressure are the primary experimental variables. It only makes sense to see that the contamination level(s) during testing are representative of the level(s) that will be typical when the transmission is used in the field.

With regard to primary experimental variables, one should also choose typical levels—with a couple of provisos. Sometimes the goal in an engineering experiment is to compare an innovative, nonstandard way of doing things to current practice. In such cases, it is not good enough simply to look at system behavior with typical settings for primary experimental variables. Also, where primary experimental variables are believed to have relatively small effects on a response, it may be necessary to choose ranges for the primary variables that are wider than normal, to see clearly how they act on the response.

Other physical realities and constraints on data collection may also make it appropriate to use atypical values of managed variables and subsequently extrapolate experimental results to "standard" circumstances. For example, it is costly enough to run studies on **pilot plants** using small quantities of chemical reagents and miniature equipment but much cheaper than experimentation on a full-scale facility. Another kind of engineering study in which levels of primary experimental variables are purposely chosen outside normal ranges is the **accelerated life test**. Such studies are done to predict the life-length properties of products that in normal usage would far outlast any study of feasible length. All that can then be done is to turn up the stress on sample units beyond normal levels, observe performance, and try to extrapolate back to a prediction for behavior under normal usage. (For example, if sensitive electronic equipment performs well under abnormally high temperature and humidity, this could well be expected to imply long useful life under normal temperature and humidity conditions.)

After the experimental levels of individual manipulated variables are chosen, they must be combined to form the experimental patterns (combinations) of managed variables. The range of choices is wide: factorial structures, fractional factorial structures, other standard structures, and patterns tailor-made for a particular problem. (Tailor-made plans will, for example, be needed in situations where particular combinations of factor levels prescribed by standard structures are a priori clearly unsafe or destructive of company property.)

But developing a detailed data collection protocol requires more than even choices of experimental combinations. Experimental order must be decided. Explicit instructions for actually carrying out the testing must be agreed upon and written down in such a way that someone who was not involved in study planning can carry out the data collection. A timetable for initial data collection must be developed. In all of this, it must be remembered that several iterations of data collection and analysis (all within given budget constraints) may be required in order to find a solution to the original engineering problem.

2.5.4 Physical Preparation

Step 8

After a project team has agreed on exactly what is to be done in a statistical study, it can address the details of how to accomplish it and assign responsibility for completion. One team member should be given responsibility for the direct oversight of actual data collection. It is all too common for people who collect the data to say, after the fact, "Oh, I did it the other way . . . I couldn't figure out exactly what you meant here . . . and besides, it was easier the way I did it."

Step 9

Again, technicians who carry out a study planned by an engineering project group often need training in the study objectives and the methods to be used. As discussed in Section 2.1, when people know why they are collecting data and have been carefully shown how to collect them, they will produce better information. Overseeing the data collection process includes making sure that this necessary training takes place.

Steps 10 & 11

The discipline involved in carefully preparing complete data collection forms and doing a dry run data analysis on fictitious values provides opportunities to refine (and even salvage) a study before the expense of data collection is incurred. When carrying out steps 10 and 11, each individual on the team gets a chance to ask, "Will the data be adequate to answer the question at hand? Or are other data needed?" The students referred to in Example 4 (page 30), who failed to measure their primary response variables, learned the importance of these steps the hard way.

Step 12

The final step in this list is writing up a best guess at what the study will show. We first came across this idea in *Statistics for Experimenters* by Box, Hunter, and Hunter. The motivation for it is sound. After a study is complete, it is only human to say, "Of course that's the way things are. We knew that all along." When a careful before-data statement is available to compare to an after-data summarization of findings, it is much easier to see what has been learned and appreciate the value of that learning.

Section 5 Exercises

1. Either take an engineering system and response variable that you are familiar with from your field or consider, for example, the United Airlines passenger flight system and the response variable Customer Satisfaction and make a cause-and-effect diagram showing a variety of variables that may potentially affect the response. How might such a diagram be practically useful?

Chapter 2 Exercises

1. Use Table B.1 and choose a simple random sample of $n = 8$ out of $N = 491$ widgets. Describe carefully how you label the widgets. Begin in the upper left corner of the table. Then use spreadsheet or statistical software to redo the selection.

2. Consider a potential student project concerning the making of popcorn. Possible factors affecting the outcome of popcorn making include at least the following: Brand of corn, Temperature of corn at beginning of cooking, Popping Method (e.g., frying versus hot air popping), Type of Oil used (if frying), Amount of Oil used (if frying), Batch Size, initial Moisture Content of corn, and Person doing the evaluation of a single batch. Using these factors and/or any others that you can think of, answer the following questions about such a project:
   (a) What is a possible response variable in a popcorn project?
   (b) Pick two possible experimental factors in this context and describe a 2×2 factorial data structure in those variables that might arise in such a study.
   (c) Describe how the concept of randomization might be employed.
   (d) Describe how the concept of blocking might be employed.

3. An experiment is to be performed to compare the effects of two different methods for loading gears in a carburizing furnace on the amount of distortion produced in a heat treating process. Thrust face runout will be measured for gears laid and for gears hung while treating.
   (a) 20 gears are to be used in the study. Randomly divide the gears into a group (of 10) to be laid and a group (of 10) to be hung, using either Table B.1 or statistical software. Describe carefully how you do this. If you use the table, begin in the upper left corner.
   (b) What are some purposes of the randomization used in part (a)?

4. A sanitary engineer wishes to compare two methods for determining chlorine content of Cl_2-demand-free water. To do this, eight quite different water samples are split in half, and one determination is made using the MSI method and another using the SIB method. Explain why it could be said that the principle of blocking was used in the engineer's study. Also argue that the resulting data set could be described as consisting of paired measurement data.

5. A research group is testing three different methods of electroplating widgets (say, methods A, B, and C). On a particular day, 18 widgets are available for testing. The effectiveness of electroplating may be strongly affected by the surface texture of the widgets. The engineer running the experiment is able to divide the 18 available widgets into three groups of 6 on the basis of surface texture. (Assume that widgets 1–6 are rough, widgets 7–12 are normal, and widgets 13–18 are smooth.)
   (a) Use Table B.1 or statistical software in an appropriate way and assign each of the treatments to 6 widgets. Carefully explain exactly how you do the assignment of levels of treatments A, B, and C to the widgets.
   (b) If equipment limitations are such that only one widget can be electroplated at once, but it is possible to complete the plating of all 18 widgets on a single day, in exactly what order would you have the widgets plated? Explain where you got this order.
   (c) If, in contrast to the situation in part (b), it is

possible to plate only 9 widgets in a single day, make up an appropriate plan for plating 9 on each of two consecutive days.

(d) If measurements of plating effectiveness are made on each of the 18 widgets, what kind of data structure will result from the scenario in part (b)? From the scenario in part (c)?

6. A company wishes to increase the light intensity of its photoflash cartridge. Two wall thicknesses ($\frac{1}{16}''$ and $\frac{1}{8}''$) and two ignition point placements are under study. Two batches of the basic formulation used in the cartridge are to be made up, each batch large enough to make 12 cartridges. Discuss how you would recommend running this initial phase of experimentation if all cartridges can be made and tested in a short time period by a single technician. Be explicit about any randomization and/or blocking you would employ. Say exactly what kinds of cartridges you would make and test, in what order. Describe the structure of the data that would result from your study.

7. Use Table B.1 or statistical software and

(a) Select a simple random sample of 5 widgets from a production run of 354 such widgets. (If you use the table, begin at the upper left corner and move left to right, top to bottom.)

(b) Select a random order of experimentation for a context where an experimental factor A has two levels; a second factor, B, has three levels; and two experimental runs are going to be made for each of the $2 \times 3 = 6$ different possible combinations of levels of the factors. Carefully describe how you do this.

8. Return to the situation of Exercise 8 of the Chapter 1 Exercises.

(a) Name factors and levels that might be used in a three-factor, full factorial study in this situation. Also name two response variables for the study. Suppose that in accord with good engineering data collection practice, you wish to include some replication in the study. Make up a data collection sheet, listing all the combinations of levels of the factors to be studied, and include blanks where the corresponding observed values of the two responses could be entered for each experimental run.

(b) Suppose that it is feasible to make the runs listed in your answer to part (a) in a completely randomized order. Use a mechanical method (like slips of paper in a hat) to arrive at a random order of experimentation for your study. Carefully describe the physical steps you follow in developing this order for data collection.

9. Use Table B.1 and

(a) Select a simple random sample of 7 widgets from a production run of 619 widgets (begin at the upper left corner of the table and move left to right, top to bottom). Tell how you labeled the widgets and name which ones make up your sample.

(b) Beginning in the table where you left off in (a), select a second simple random sample of 7 widgets. Is this sample the same as the first? Is there any overlap at all?

10. Redo Exercise 9 using spreadsheet or statistical software.

11. Consider a study comparing the lifetimes (measured in terms of numbers of holes drilled before failure) of two different brands of 8-mm drills in drilling 1045 steel. Suppose that steel bars from three different heats (batches) of steel are available for use in the study, and it is possible that the different heats have differing physical properties. The lifetimes of a total of 15 drills of each brand will be measured, and each of the bars available is large enough to accommodate as much drilling as will be done in the entire study.

(a) Describe how the concept of control could be used to deal with the possibility that different heats might have different physical properties (such as hardnesses).

(b) Name one advantage and one drawback to controlling the heat.

(c) Describe how one might use the concept of blocking to deal with the possibility that different heats might have different physical properties.