3 Introduction to Probability: The Binomial and Normal Probability Distributions

The theory of statistics is based on probability. Some basic definitions and theorems are introduced in this chapter. This elementary discussion leads to the concept of a probability distribution, a mathematical function that assigns probabilities for outcomes in its domain. The properties of (a) the binomial distribution, a discrete distribution, and (b) the normal distribution, a continuous distribution, are presented. The normal distribution is the basis of modern statistical theory and methodology. One of the chief reasons for the pervasion of the normal distribution in statistics is the central limit theorem, which shows that means of samples from virtually all probability distributions tend to be normal for large sample sizes. Also, many of the probability distributions used in statistical analyses are based on the normal distribution. These include the *t*, *F*, and chi-square distributions. The binomial distribution is applicable to experimental results that have two possible outcomes, such as pass or fail in quality control, or cured or not cured in a clinical drug study. With a minimal understanding of probability, one can apply statistical methods intelligently to the simple but prevalent problems that crop up in the analysis of experimental data.

3.1 INTRODUCTION

Most of us have an intuitive idea of the meaning of probability. The meaning and use of probability in everyday life is a subconscious integration of experience and knowledge that allows us to say, for example: "If I purchase this car at my local dealer, the convenience and good service will *probably* make it worthwhile despite the greater initial cost of the car." From a statistical point of view, we will try to be more precise in the definition of probability. The *Random House Dictionary of the English Language* defines probability as "The likelihood of an occurrence expressed by the ratio of the actual occurrences to that of all possible occurrences; the relative frequency with which an event occurs, or is likely to occur." Therefore, the probability of observing an event can be defined as the proportion of such events that will occur in a large number of observations or experimental trials.

The approach to probability is often associated with odds in gambling or games of chance, and picturing probability in this context will help its understanding. When placing a bet on the outcome of a coin toss, the game of "heads and tails," one could reasonably *guess* that the probability of a head or tail is one-half (1/2) or 50%. One-half of the outcomes will be heads and one-half will be tails. Do you think that the probability of observing a head (or tail) on a single toss of the coin is exactly 0.5 (50%)? Probably not, a probability of 50% would result only if the coin is absolutely balanced. The only way to verify the probability is to carry out an extensive experiment, tossing a coin a million times or more and counting the proportion of heads or tails that result.

The gambler who knows that the odds in a game of craps favor the "house" will lose in the long run. Why should a knowledgeable person play a losing game? Other than for psychological reasons, the gambler may feel that a reasonably good chance of winning on any single bet is worth the chance, and maybe "Lady Luck" will be on his side. Probability is a measure of uncertainty. We may be able to predict accurately some average result in the long run, but the outcome of a single experiment cannot be anticipated with certainty.

3.2 SOME BASIC PROBABILITY

The concept of probability is "probably" best understood when discussing discontinuous or *discrete* variables. These variables have a countable number of outcomes. Consider an experiment in which only one of two possible outcomes can occur. For example, the result of treatment with an antibiotic is that an infection is either *cured* or *not cured* within five days. Although this situation is conceptually analogous to the coin-tossing example, it differs in the following respect. For the coin-tossing example, the probability can be determined by a rational examination of the nature of the experiment. If the coin is balanced, heads and tails are equally likely; the probability of a head is equal to the probability of a tail = 0.5. In the case of the antibiotic cure, however, the probability of a cure is not easily ascertained a priori, that is, prior to performing an experiment. If the antibiotic were widely used, based on his or her own experience, a physician prescriber of the product might be able to give a good estimate of the probability of a cure for patients treated with the drug. For example, in the physician's practice, he or she may have observed that approximately three of four patients treated with the antibiotic are cured. For this physician, the probability that a patient will be cured when treated with the antibiotic is approximately 75%.

A large multicenter clinical trial would give a better estimate of the probability of success after treatment. A study of 1000 patients might show 786 patients cured; the *probability of a cure* is estimated as 0.786 or 78.6%. This does not mean that the exact probability is 0.786. The exact probability can be determined only by treating the total population and observing the proportion cured, a practical impossibility in this case. In this context, it would be fair to say that exact probabilities are nearly always unknown.

3.2.1 Some Elementary Definitions and Theorems

1. $$0 \le P(A) \le 1 \tag{3.1}$$

where $P(A)$ is the probability of observing event A. The probability of any event or experimental outcome, $P(A)$, cannot be less than 0 or greater than 1. An impossible event has a probability of 0. A certain event has a probability of 1.

2. If events $A, B, C, \ldots$ are *mutually exclusive*, the probability of observing A or B or $C\ldots$ is the sum of the probabilities of each event, $A, B, C, \ldots$. If two or more events are "mutually exclusive," the events cannot occur simultaneously, that is, if one event is observed, the other event(s) cannot occur. For example, we cannot observe both a head and a tail on a single toss of a coin.

$$P(A \text{ or } B \text{ or } C \ldots) = P(A) + P(B) + P(C) + \cdots \tag{3.2}$$

An example frequently encountered in quality control illustrates this theorem. Among 1,000,000 tablets in a batch, 50,000 are known to be flawed, perhaps containing *specks* of grease. The probability of finding a *randomly* chosen tablet with specks is 50,000/1,000,000 = 0.05 or 5%. The process of *randomly* choosing a tablet is akin to a lottery. The tablets are well mixed, ensuring that each tablet has an equal chance of being chosen. While blindfolded, one figuratively chooses a single tablet from a container containing the 1,000,000 tablets (see chapter 4 for a detailed discussion of random sampling). A gambler making an equitable bet would give odds of 19 to 1 against a specked tablet being chosen (1 of 20 tablets is specked). Odds are defined as

$$\frac{P(A)}{1 - P(A)}.$$

There are other defects among the 1,000,000 tablets. Thirty thousand, or 3%, have *chipped* edges and 40,000 (4%) are *discolored*. If these defects are mutually exclusive, the probability of observing any one of these events for a single tablet is 0.03 and 0.04, respectively [Fig. 3.1(A)]. According to Eq. (3.2), the probability of choosing an unacceptable tablet (specked, chipped, or discolored) at random is 0.05 + 0.03 + 0.04 = 0.12, or 12%. (The probability of choosing an acceptable tablet is 1 − 0.12 = 0.88.)

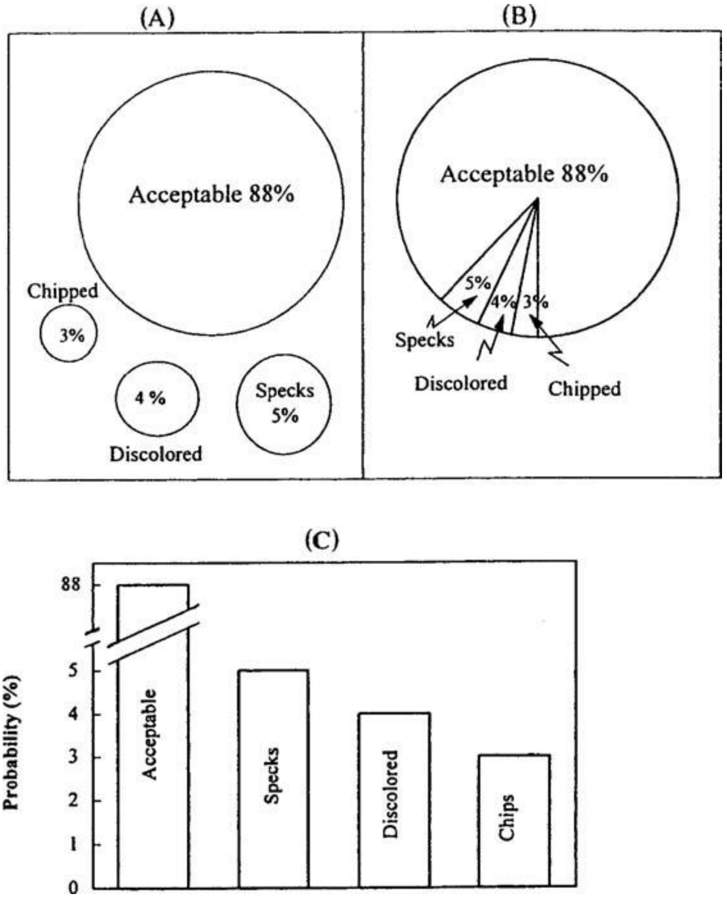


Figure 3.1 Probability distribution for tablet attributes.

3. $P\,(A) + P\,(B) + P\,(C) + \cdots = 1$ (3.3)

where $A, B, C, \ldots$ are mutually exclusive and exhaust all possible outcomes.

If the set of all possible experimental outcomes are mutually exclusive, the sum of the probabilities of all possible outcomes is equal to 1. This is equivalent to saying that we are certain that one of the mutually exclusive outcomes will occur.

All the four events in Figure 3.1 do not have to be mutually exclusive. In general:

4. If two events are not mutually exclusive,

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B). \tag{3.4}$$

Note that if A and B are mutually exclusive, $P(A$ and $B) = 0$, and for two events, A and B, Eqs. (3.2) and (3.4) are identical. (A and B) means the simultaneous occurrence of A and B. (A or B) means that A or B or both A and B occur. For example, some tablets with chips may also be specked. If 20,000 tablets are both chipped *and* specked in the example above, one can verify that 60,000 tablets are specked *or* chipped.

$$\begin{aligned} P(\text{specked or chipped}) &= P(\text{specked}) + P(\text{chipped}) - P(\text{specked or chipped}) \\ &= 0.05 + 0.03 - 0.02 = 0.06 \end{aligned}$$

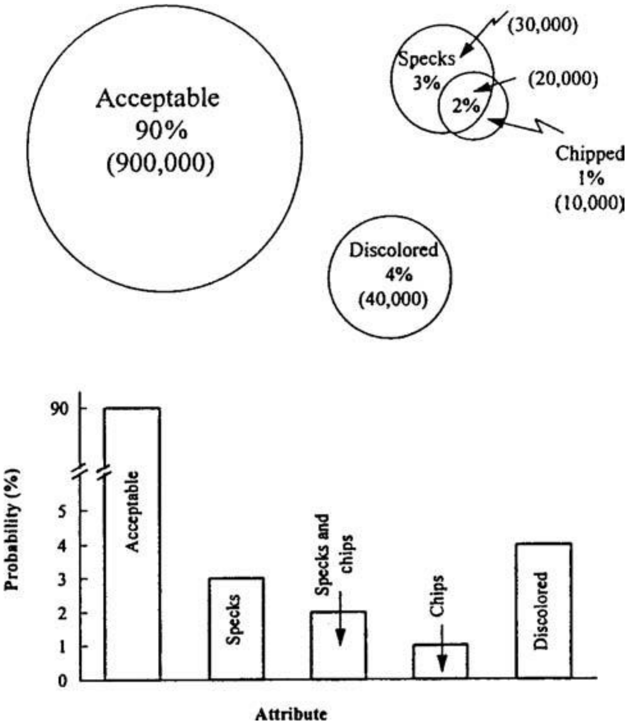


Figure 3.2 Distribution of tablet attributes where attributes are not all mutually exclusive.

The probability of finding a specked or chipped tablet is 0.06. Thirty thousand tablets are *only* specked, 10,000 tablets are *only* chipped, and 20,000 tablets are both specked and chipped; a total of 60,000 tablets specked or chipped. The distribution of tablet attributes under these conditions is shown in Figure 3.2. (Also, see Exercise Problem 23.)

With reference to this example of tablet attributes, we can enumerate all possible mutually exclusive events. In the former case, where each tablet was acceptable or had only a single defect, there are four possible outcomes (specked, chipped edges, discolored, and acceptable tablets). These four outcomes and their associated probabilities make up a *probability distribution*, which can be represented in several ways, as shown in Figure 3.1. The distribution of attributes where some tablets may be both specked and chipped is shown in Figure 3.2. The notion of a probability distribution is discussed further later in this chapter (sect. 3.3).

5. The multiplicative law of probability states that

$$P(A \text{ and } B) = P(A|B)\,P(B), \tag{3.5}$$

where $P(A|B)$ is known as the conditional probability of A given that B occurs. In the present example, the probability that a tablet will be specked given that the tablet is chipped is [from Eq. (3.5)]

$$\begin{aligned} P(\text{specked} \mid \text{chipped}) &= \frac{P(\text{specked and chipped})}{P(\text{chipped})} \\ &= \frac{0.02}{0.03} = \frac{2}{3}. \end{aligned}$$

Referring to Figure 3.2, it is clear that 2/3 of the chipped tablets are also specked. Thus, the probability of a tablet being specked given that it is also chipped is 2/3.

3.2.2 Independent Events

In games of chance, such as roulette, the probability of winning (or losing) is theoretically the same on each turn of the wheel, irrespective of prior outcomes. Each turn of the wheel results in an independent outcome. The events, A and B, are said to be independent if a knowledge of B does not affect the probability of A. Mathematically, two events are independent if

$$P(A \mid B) = P(A). \tag{3.6}$$

Substituting Eq. (3.6) into Eq. (3.5), we can say that if

$$P(A \text{ and } B) = P(A)P(B), \tag{3.7}$$

then A *and* B *are independent*. When sampling tablets for defects, if each tablet is selected at random and the batch size is very large, the sample observations may be considered independent. Thus, in the example of tablet attributes shown in Figure 3.4, the probability of selecting an acceptable tablet (A) followed by a defective tablet (B) is

$$(0.88)(0.12) = 0.106.$$

The probability of selecting two tablets, both of which are acceptable, is $0.88 \times 0.88 = 0.7744$.

3.3 PROBABILITY DISTRIBUTIONS—THE BINOMIAL DISTRIBUTION

To understand probability further, one should have a notion of the concept of a probability distribution, introduced in section 3.2. A probability distribution is a mathematical representation (function) of the probabilities associated with the values of a random variable.

For discrete data, the concept can be illustrated by using the simple example of the outcome of antibiotic therapy introduced earlier in this chapter. In this example, the outcome of a patient following treatment can take on one of two possibilities: a cure with a probability of 0.75 or a failure with a probability of 0.25. Assigning the value 1 for a cure and 0 for a failure, the probability distribution is simply

$$f(1) = 0.75$$
$$f(0) = 0.25.$$

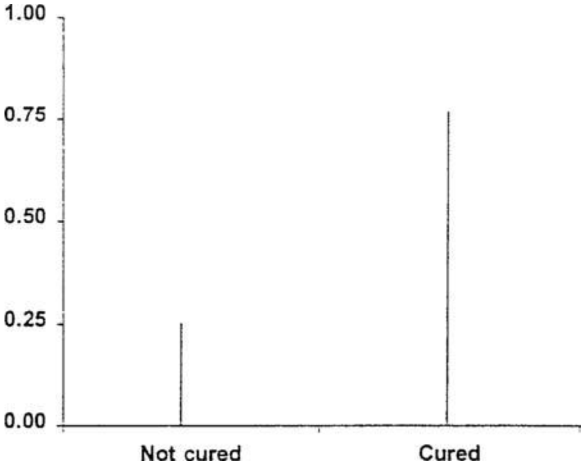


Figure 3.3 Probability distribution of a binomial outcome based on a single observation.

Table 3.1 Some Examples of Binomial Data in Pharmaceutical Research

Experiment or process	Dichotomous data
LD_{50} determination	Animals *live* or *die* after dosing. Determine dose that kills 50% of animals
ED_{50} determination	Drug is *effective* or *not effective.* Determine dose that is effective in 50% of animals
Sampling for defects	In quality control, product is sampled for defects. Tablets are *acceptable* or *unacceptable*
Clinical trials	Treatment is *successful* or *not successful*
Formulation modification	A. Palatability preference of *old* and *new* formulation B. New formulation is *more* or *less available* in crossover design

Figure 3.3 shows the probability distribution for this example, the random variable being the outcome of a patient treated with the antibiotic. This is an example of a binomial distribution. Another example of a binomial distribution is the coin-tossing game, heads or tails where the two outcomes have equal probability, 0.5. This binomial distribution ($p = 0.5$) has application in statistical methods, for example, the Sign test (sect. 15.2).

When a single observation can be dichotomized, that is, the observation can be placed into one of two possible categories, the binomial distribution can be used to define the probability characteristics of one or more such observations. The binomial distribution is a very important probability distribution in applications in pharmaceutical research. The few examples noted in Table 3.1 reveal its pervading presence in pharmaceutical processes.

3.3.1 Some Definitions

A *binomial trial* is a single binomial experiment or observation. The treatment of a single patient with the antibiotic is a binomial trial. The trial must result in only one of two outcomes, where the two outcomes are *mutually exclusive*. In the antibiotic example, the only possible outcomes are that a patient is either cured or not cured. In addition, only one of these outcomes is possible after treatment. A patient cannot be both cured and not cured after treatment. Each binomial trial must be *independent*. The result of a patient's treatment does not influence the outcome of the treatment for a different patient. In another example, when randomly sampling tablets for a

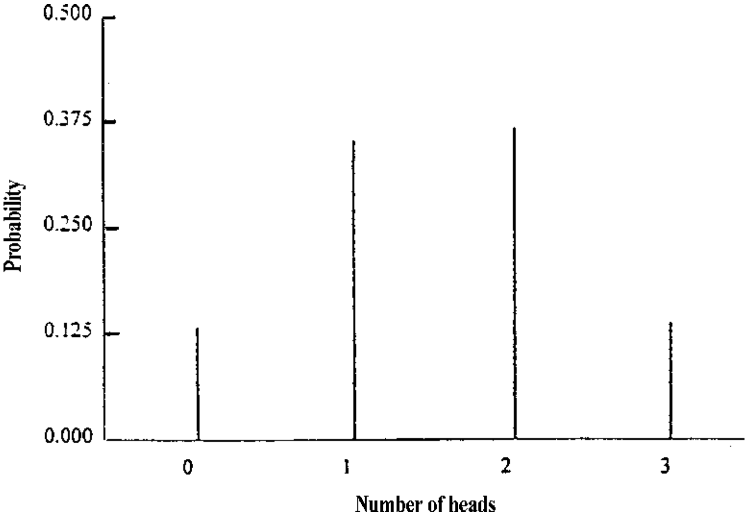


Figure 3.4 Probability distribution of binomial with $p = 0.5$ and $N = 3$.

binomial attribute, chipped or not chipped, the observation of a chipped tablet does not depend on or influence the outcome observed for any other tablet.

The binomial distribution is completely defined by two parameters: (a) the probability of one or the other outcome, and (b) the number of trials or observations, N. Given these two parameters, we can calculate the probability of any specified number of successes in N trials. For the antibiotic example, the probability of success is 0.75. With this information, we can calculate the probability that three of four patients will be cured ($N = 4$). We could also calculate this result, given the probability of failure (0.25). The probability of three of four patients being cured is exactly the same as the probability of one of four patients not being cured.

The probability of success (or failure) lies between 0 and 1. The probability of failure (the complement of a success) is 1 minus the probability of success $[1 - P(\text{success})]$.

Since the outcome of a binomial trial must be either success or failure, $P(\text{success}) + P(\text{failure}) = 1$ [see Eq. (3.3)].

The standard deviation of a binomial distribution with probability of success, p, and N trials is $\sqrt{pq/N}$, where $q = 1 - p$. The s.d. of the proportion of successes of antibiotic treatment in 16 trials is $\sqrt{0.75 \times 0.25/16} = 0.108$ (also see sect. 3.3.2).

The probability of the outcome of a binomial experiment consisting of N trials can be computed from the expansion of the expression

$$(p+q)^N, \tag{3.8}$$

where p is defined as the probability of success and q is the probability of failure. For example, consider the outcomes that are possible after three tosses of a coin. There are four ($N + 1$) possible results

1. three heads;
2. two heads and one tail;
3. two tails and one head;
4. three tails.

For the outcome of the treatment of three patients in the antibiotic example, the four possible results are

1. three cures;
2. two cures and one failure;
3. two failures and one cure;
4. three failures.

The probabilities of these events can be calculated from the individual terms from the expansion of $(p + q)^N$, where $N = 3$, the number of binomial trials.

$$(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$$

If $p = q = 1/2$, as is the case in coin tossing, then

$$p^3 = (1/2)^3 = 1/8 = P(\text{three heads})$$

$$3p^2q = 3/8 = P(\text{two heads and one tail})$$

$$3pq^2 = 3/8 = P(\text{two tails and one head })$$

$$q^3 = 1/8 = P(\text{three tails})$$

If $p = 0.75$ and $q = 0.25$, as is the case for the antibiotic example, then

$$p^3 = (0.75)^3 = 0.422 = P(3\,\text{cures})$$

$$3p^2q = 3(0.75)^2(0.25) = 0.422\ P(2\,\text{cures and 1 failure})$$

$$3pq^2 = 3(0.75)(0.25)^2 = 0.141\ P(1\,\text{cure and 2 failures})$$

$$q^3 = (0.25)^3 = 0.016 = P(3\,\text{failures})$$

The sum of the probabilities of all possible outcomes of three patients being treated or three sequential coin tosses is equal to 1 (e.g., $1/8 + 3/8 + 3/8 + 1/8 = 1$).

This is true of any binomial experiment because $(p + q)^N$ must equal 1 by definition (i.e., $p + q = 1$). The probability distribution of the coin-tossing experiment with $N = 3$ is shown in Figure 3.4. Note that this is a *discrete* distribution. The particular binomial distribution shown in the figure comprises only *four* possible outcomes (the four sticks).

A gambler looking for a fair game, one with equitable odds, would give odds of 7 to 1 on a bet that three heads would be observed in three tosses of a coin. The payoff would be eight dollars (including the dollar bet) for a one-dollar bet. A bet that either three heads or three tails would be observed would have odds of 3 to 1. (The probability of either *three heads* or *three tails* is $1/4 = 1/8 + 1/8$.)

To calculate exact probabilities in the binomial case, the expansion of the binomial, $(p + q)^N$ can be generalized by a single formula:

$$\text{Probability of } X \text{ successes in } N \text{ trials} = \binom{N}{X} p^x q^{N-X}. \tag{3.9}$$

$$\binom{N}{X} \text{ is defined as } \frac{N!}{X!(N-X)!}$$

(Remember that 0! is equal to 1.)

Consider the binomial distribution with $p = 0.75$ and $N = 4$ for the antibiotic example. This represents the distribution of outcomes after treating four patients. There are five possible outcomes

no patients are cured;
one patient is cured;
two patients are cured;
three patients are cured;
four patients are cured.

The probability that three of four patients are cured can be calculated from Eq. (3.9)

$$\binom{4}{3}(0.75)^3(0.25)^1 = \frac{4 \cdot 3 \cdot 2 \cdot 1}{1 \cdot 3 \cdot 2 \cdot 1}(0.42188)(0.25) = 0.42188.$$

The meaning of this particular calculation will be explained in detail in order to gain some insight into solving probability problems. There are four ways in which three patients can be cured and one patient not cured (Table 3.2). Denoting the four patients as A, B, C, and D, the probability that patients A, B, and C are cured and patient D is not cured is equal to

$$(0.75)(0.75)(0.75)(0.25) = 0.1055, \tag{3.10}$$

Table 3.2 Four Ways in Which Three of Four Patients Are Cured

	1	2	3	4
Patients cured	A, B, C	A, B, D	A, C, D	B, C, D
Patients not cured	D	C	B	A

where 0.25 is the probability that patient D will *not* be cured. There is no reason why any of the four possibilities shown in Table 3.2 should occur more or less frequently than any other (i.e., each possibility is equally likely). Therefore, the probability that the antibiotic will successfully cure exactly three patients is four times the probability calculated in Eq. (3.10)

$$4(0.1055) = 0.422.$$

The expression $\binom{4}{3}$represents a combination, a selection of three objects, disregarding order, from four distinct objects. The combination,$\binom{4}{3}$, is equal to 4, and, as we have just demonstrated, there are four ways in which three cures can be obtained from four patients. Each one of these possible outcomes has a probability of $(0.75)^3(0.25)^1$. Thus, the probability of three cures in four patients is $4(0.75)^3\ (0.25)^1$ as before.

The probability distribution based on the possible outcomes of an experiment in which four patients are treated with the antibiotic (the probability of a cure is 0.75) is shown in Table 3.3 and Figure 3.5. Note that the sum of the probabilities of the possible outcomes equals 1, as is also shown in the cumulative probability function plotted in Figure 3.5(B). The cumulative distribution is a nondecreasing function starting at a probability of zero and ending at a probability of 1. Figures 3.1 and 3.2, describing the distribution of tablet attributes in a batch of tablets, are examples of other discrete probability distributions.

Statistical hypothesis testing, a procedure for making decisions based on variable data is based on probability theory. In the following example, we use data observed in a coin-tossing game to decide whether or not we believe the coin to be loaded (biased).

You are an observer of a coin-tossing game and you are debating whether or not you should become an active participant. You note that only one head occurred among 10 tosses of the coin. You calculate the probability of such an event because it occurs to you that one head in 10 tosses of a coin is very unlikely; something is amiss (a "loaded" coin!). Thus, if the probability of a head is 0.5, the chances of observing one head in 10 tosses of a coin is less than 1 in 100 (Exercise Problem 18). This low probability suggests a coin that is not balanced. However, you properly note that the probability of any *single event* or outcome (such as one head in 10 trials) is apt to be small if N is sufficiently large. You decide to calculate the probability of this perhaps unusual result *plus* all other possible outcomes that are equally or less probable. In our example, this includes possibilities of no heads in 10 tosses, in addition to one or no tails in 10 tosses. These four probabilities (no heads, one head, no tails, and one tail) total approximately 2.2%. This is strong evidence in favor of a biased coin. Such a decision is based on the fact that the chance of obtaining an event as unlikely or less likely than one head in 10 tosses is about 1 in

Table 3.3 Probability Distribution for Outcomes of Treating Four Patients with an Antibiotic

Outcome	Probability
No cures	0.00391
One cure	0.04688
Two cures	0.21094
Three cures	0.42188
Four cures	0.31641

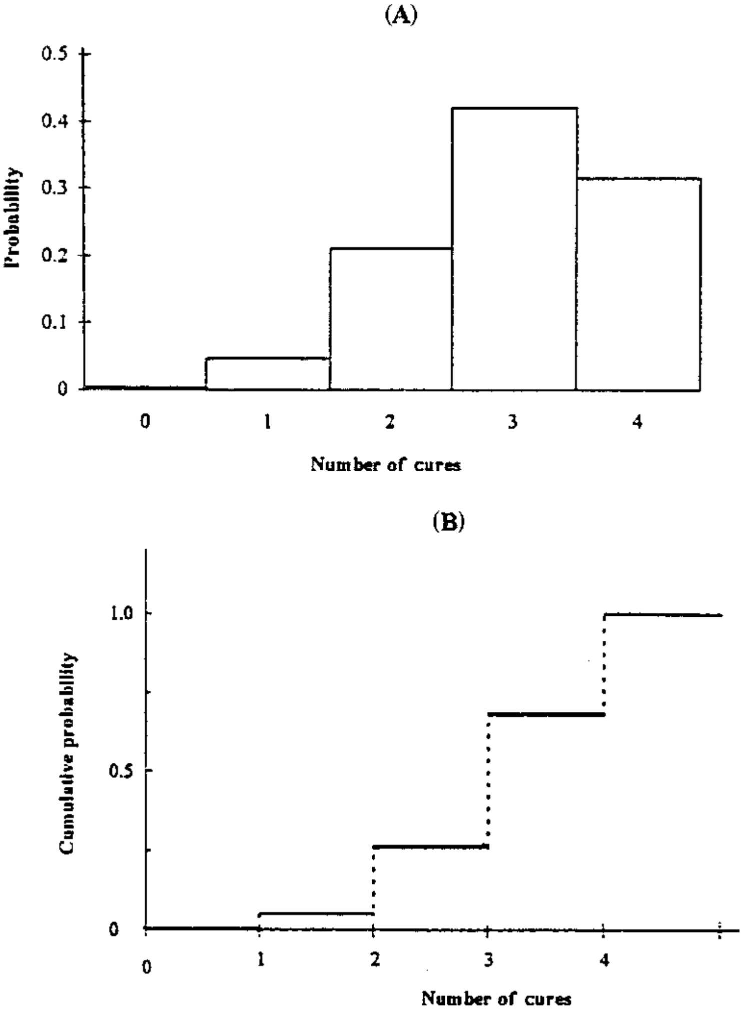


Figure 3.5 Probability distribution graph for outcomes of treating four patients with an antibiotic.

50 (2.2%) if the coin is *balanced.* You might wisely bet on tails on the next toss. You have made a decision: "The coin has a probability of less than 0.5 of showing heads on a single toss."

The probability distribution for the number of heads (or tails) in 10 tosses of a coin ($p = 0.5$ and $N = 10$) is shown in Figure 3.6. Note the symmetry of the distribution.

Although this is a discrete distribution, the "sticks" assume a symmetric shape similar to the normal curve. The two unlikely events in each "tail" (i.e., no heads or tails or one head or one tail) have a total probability of 0.022. The center and peak of the distribution is observed to be at $X = 5$, equal to NP, the number of trials times the probability of success. (See also Appendix Table IV.3, $p = 0.5$, $N = 10$.)

The application of binomial probabilities can be extended to more practical problems than gambling odds for the pharmaceutical scientist. When tablets are inspected for attributes or patients treated with a new antibiotic, we can apply a knowledge of the properties of the binomial distribution to estimate the true proportion or probability of success, and make appropriate decisions based on these estimates.

3.3.2 Summary of Properties of the Binomial Distribution

1. The binomial distribution is defined by N and p. With a knowledge of these parameters, the probability of any outcome of N binomial trials can be calculated from Eq. (3.9). We have

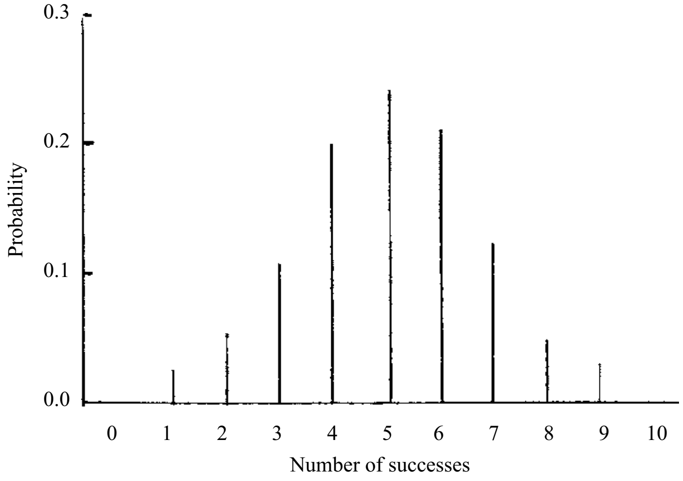


Figure 3.6 Probability distribution for $p = 0.5$ and $N = 10$.

noted that the sum of all possible outcomes of a binomial experiment with N trials is 1, which conforms to the notion of a probability distribution.

2. The results of a binomial experiment can be expressed either as the *number of successes* or as a *proportion*. Thus, if six heads are observed in 10 tosses of a coin, we can also say that 60% of the tosses are heads. If 16 defective tablets are observed in a random sample of 1000 tablets, we can say that 1.6% of the tablets sampled are defective. In terms of proportions, the *true mean* of the binomial population is equal to the probability of success, p. The sample proportion (0.6 in the coin-tossing example and 0.016 in the example of sampling for defective tablets) is an estimate of the true proportion.
3. The variability of the results of a binomial experiment is expressed as a standard deviation. For example, when inspecting tablets for the number of defectives, a different number of defective tablets will be observed depending on which 1000 tablets happen to be chosen. This variation, dependent on the particular sample inspected, is also known as *sampling error.* The s.d. of a binomial distribution can be expressed in two ways, depending on the manner in which the mean is presented (i.e., as a proportion or as the number of successes). The s.d. in terms of proportion of successes is

$$\sqrt{\frac{pq}{N}}. \tag{3.11}$$

In terms of number of successes, the s.d. is

$$\sqrt{Npq}, \tag{3.12}$$

where N is the sample size, the number of binomial trials. As shown in Eqs. (3.11) and (3.12), the s.d. is dependent on the value of P for binomial variables. The maximum s.d. occurs when $p = q = 0.5$, because Pq is maximized. The value of pq does not change very much with varying P and q until P or q reach low or high values, close to or more extreme than 0.2 and 0.8.

p	*q*	*pq*
0.5	0.5	0.25
0.4	0.6	0.24
0.3	0.7	0.21
0.2	0.8	0.16
0.1	0.9	0.09

4. When dealing with proportions, the variability of the observed proportion can be made as small as we wish by increasing the sample size [similar to the s.d. of the mean of samples of size N, Eq. (1.8)]. This means that we can estimate the proportion of "successes" in a population with very little error if we choose a sufficiently large sample. In the case of the tablet inspection example above, the variability (s.d.) of the proportion for samples of size 100 is

$$\sqrt{\frac{(0.016)(0.984)}{100}} = 0.0125.$$

By sampling 1000 tablets, we can reduce the variability by a factor of 3.16 ($\sqrt{100/1000} = 1/3.16$). The variability of the estimate of the true proportion (i.e., the sample estimate) is not dependent on the population size (the size of the entire batch of tablets in this example), but is dependent only on the size of the sample selected for observation. This interesting fact is true if the sample size is considerably smaller than the size of the population. Otherwise, a correction must be made in the calculation of the s.d. [4]. If the sample size is no more than 5% of the population size, the correction is negligible. In virtually all of the examples that concern us in pharmaceutical experimentation, the sample size is considerably less than the population size. Since binomial data are often easy to obtain, large sample sizes can often be accommodated to obtain very precise estimates of population parameters. An oft-quoted example is that a sample size of 6000 to 7000 randomly selected voters will be sufficient to estimate the outcome of a national election within 1% of the total popular vote. Similarly, when sampling tablets for defects, 6000 to 7000 tablets will estimate the proportion of a property of the tablets (e.g., defects) within, at most, 1% of the true value. (The least precise estimate occurs when $p = 0.5$.)

3.3.3 Confidence Limits with *N* Observations and Zero Successes or Failures

If one observes N independent binomial variables with zero successes (or failures), it is often of interest to place confidence limits on the true proportion of successes in the universe. As way of illustration, suppose we are testing an injectable product for sterility. There is no way of guaranteeing that all items in the batch will be sterile without 100% testing. Since the test may be destructive, a sample is taken. We expect to find all of the items tested to be sterile, that is, 100% sterile. One, then, may ask, what are the confidence limits for the true proportion of items in the batch that are sterile. The upper limit will be 100%. That is, if we see N items that are sterile, it is certainly possible that all of the items in the batch are sterile. The lower limit may be calculated as follows:

$$\text{Lower confidence limit} = p^N = P, \tag{3.12A}$$

Where, p is the lower confidence limit, $P = 1 -$ probability of the confidence interval (e.g (1 − 0.95 for a 95% confidence interval) and N is the sample size. Note that the upper limit is 1.00.

Example:

One thousand (1000) items in a batch of 100,000 are tested for sterility with no failures (100% successes). What is the 95% confidence interval for the true proportion of sterile items in the batch.

Eq. (3.12A) can be written as $\ln(p) = \ln(P)/N$

$\ln(p) = \ln(1 - 0.95)/1000 = (-2.996/1000)$

$p = 0.997.$

That is, the 95% confidence interval for the proportion of sterile items is 0.997 to 1.00.

The 99% confidence interval is:

$\ln(p) = \ln(1 - 0.99)/1000 = (-4.61/1000)$

$p = 0.995.$

The 99% confidence interval for the proportion of sterile items is 0.9954 to 1.00.

(Note that $0.9954^{1000} = 0.01$.)

3.3.4 The Negative Binomial Distribution [5,6]

The negative binomial distribution does not have wide use in the pharmaceutical sciences, but can be useful in special situations. In a clinical trial, we might ask, for example, "How many successive cures can we expect to observe before seeing a failure, with a knowledge of the cure rate?" In quality control, we may be interested in the expected number of consecutive successes before a failure is observed, given the rate of failure. Another question might be, "What is the average number of consecutive good tablets observed before a failure is observed?" In general, the probability function is

$$= \left(\frac{k+r-1}{k}\right)\{P^r\}\{1-P^r\}, \tag{3.12B}$$

where $0 < P < 1$, the probability of a success.

If $r = 1$, this is the probability distribution of failures before the first success. This can also be stated as the probability of success on the $(k+1)^{\text{th}}$ trial after k failures.

If $r = 1$, Eq. (3.11A) reduces to

$$\{P\}\left\{1-P^k\right\}. \tag{3.12C}$$

Consider the following question: What is the probability that we will observe 50 good tablets before observing a split tablet on the 51st tablet. The probability of a split tablet is 0.01. Here, $p = 0.99$. (A good tablet is considered a failure in this context.) From Eq. (3.12C), the probability is

$$0.99\left(1-0.99^{50}\right) = 0.599.$$

What is the average number of good tablets that a patient would take before he/she observes a split tablet. This can be calculated by using the negative binomial distribution.

The average number of good tablets before a split tablet is observed is P/q, where P is the probability of a good tablet (0.99) and q is $(1-P)$ equal to 0.01.

The average is $\frac{0.99}{0.01} = 99$ tablets.

Therefore, on the average, a patient would take 99 tablets before encountering a split tablet. If the tablet is taken once a day, it would take 99 days on the average before a split tablet was observed (5,6).

3.4 CONTINUOUS DATA DISTRIBUTIONS

Another view of probability concerns continuous data such as tablet dissolution time. The probability that any single tablet will have a particular specified dissolution result is 0, because the number of possible outcomes for continuous data is infinite. Probability can be conceived as the ratio of the number of times that an event occurs to the total number of possible outcomes. If the total number of outcomes is infinite, the probability of any single event is zero. This concept can be confusing. If one observes a large number of dissolution results, such as time to 90% dissolution, any particular observation might appear to have a finite probability of occurring. Analogous to the discussion for discrete data, could we not make an equitable bet that a result for dissolution of exactly 5 minutes 13 seconds, for example, would be observed? The apparent contradiction is due to the fact that data that are *continuous, in theory,* appear as *discrete* data *in practice* because of the limitations of measuring instruments, as discussed in chapter 1. For example, a sensitive clock could measure time to virtually any given precision (i.e., to small fractions of a second). It would be difficult to conceive of winning a bet that a 90% dissolution time would occur at a very specific time, where time can be measured to any specified degree of precision (e.g., 30 minutes 8.21683475... seconds).

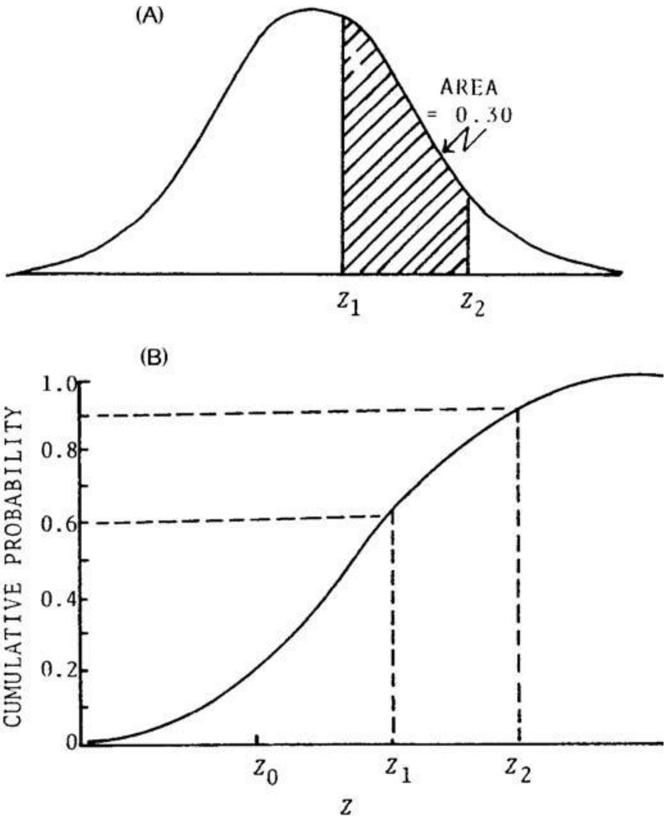


Figure 3.7 A normal distribution.

With *continuous* variables, we cannot express probabilities in as simple or intuitive a fashion as was done with discrete variables. Applications of calculus are necessary to describe concepts of probability with continuous distributions. Continuous cumulative probability distributions are represented by smooth curves (Fig. 3.7) rather than the step-like function shown in Figure 3.5(B). The area under the probability distribution curve (also known as the cumulative probability density) is equal to 1 for all probability functions. Thus the area under the normal distribution curve in Figure 3.7(A) is equal to 1.

3.4.1 The Normal Distribution

The normal distribution is an example of a continuous probability density function. The normal distribution is most familiar as the symmetrical, bell-shaped curve shown in Figure 3.8. A theoretical normal distribution is a continuous probability distribution and consists of an infinite number of values. In the theoretical normal distribution, the data points extend from positive infinity to negative infinity. It is clear that scientific data from pharmaceutical experiments cannot possibly fit this definition. Nevertheless, if real data conform reasonably well with the theoretical definition of the normal curve, adequate approximations, if not very accurate estimates of probability, can be computed based on normal curve theory.

The equation for the normal distribution (normal probability density) is

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(1/2)(X-\mu)^2/\sigma^2}, \tag{3.13}$$

where σ is the s.d., μ the mean, X the value of the observation, e the base of natural logarithms, 2.718 . . .; and Y the ordinate of normal curve, a function of X.

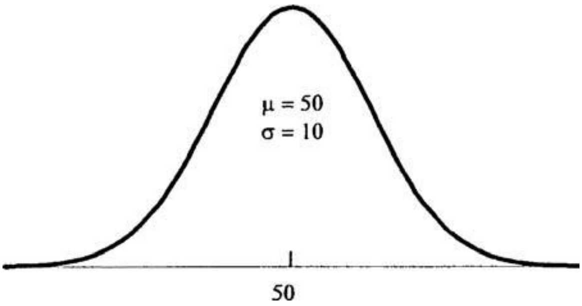


Figure 3.8 A typical normal curve.

The normal distribution is defined by its mean, μ, and its s.d., σ [see Eq. (3.13)]. This means that if these two parameters of the normal distribution are known, all the properties of the distribution are known. There are any number of different normal distributions. They all have the typical symmetrical, bell-shaped appearance. They are differentiated only by their means, a measure of location, and their s.d., a measure of spread. The normal curve shown in Figure 3.8 can be considered to define the distribution of the potencies of tablets in a batch of tablets. Most of the tablets have a potency close to the mean potency of 50 mg. The farther the assay values are from the mean, the fewer the number of tablets there will be with these more extreme values. As noted above, the spread or shape of the normal distribution is dependent on the s.d. A large s.d. means that the spread is large. In this example, a larger s.d. means that there are more tablets far removed from the mean, perhaps far enough to be out of specifications (Fig. 3.9).

In real-life situations, the distribution of a finite number of values often closely approximates a normal distribution. Weights of tablets taken from a single batch may be approximately normally distributed. For practical purposes, any continuous distribution can be visualized as being constructed by categorizing a large amount of data in small equilength intervals and constructing a histogram. Such a histogram can similarly be constructed for normally distributed variables.

Suppose that all the tablets from a large batch are weighed and categorized in small intervals or boxes (Fig. 3.10). The number of tablets in each box is counted and a histogram plotted as in Figure 3.11. As more boxes are added and the intervals made shorter, the intervals will eventually be so small that the distinction between the bars in the histogram is lost and a smooth curve results, as shown in Figure 3.12. In this example, the histogram of tablet weights looks like a normal curve.

Areas under the normal curve represent probabilities and are obtained by appropriate integration of Eq. (3.13). In Figure 3.7, the probability of observing a value between Z_1 and Z_2 is calculated by integrating the normal density function between Z_1 and Z_2.

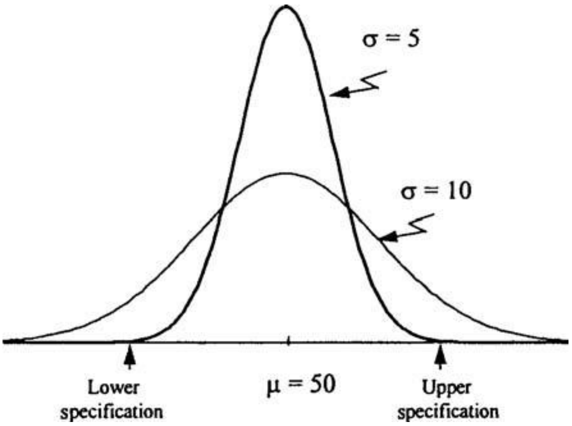


Figure 3.9 Two normal curves with different standard deviations.

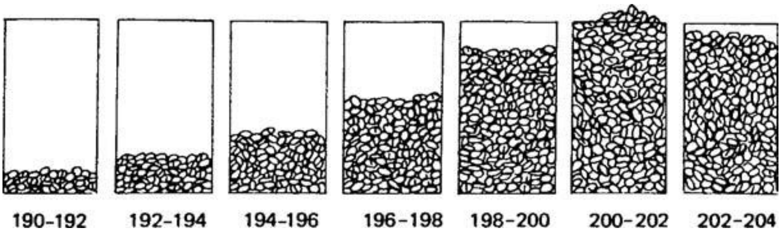


Figure 3.10 Categorization of tablets from a tablet batch by weight.

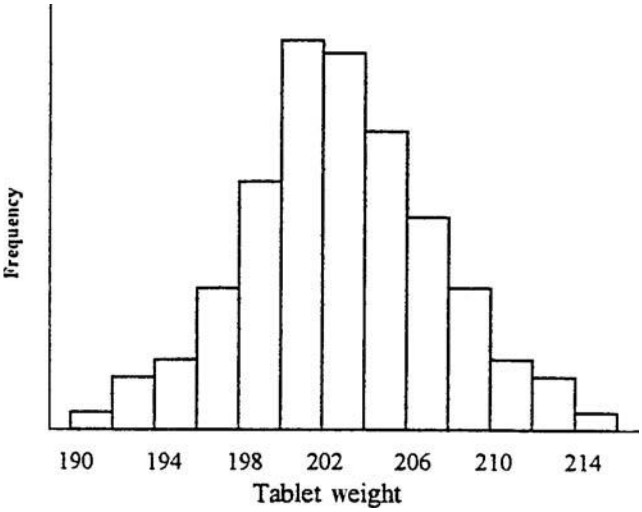


Figure 3.11 Histogram of tablet weights.

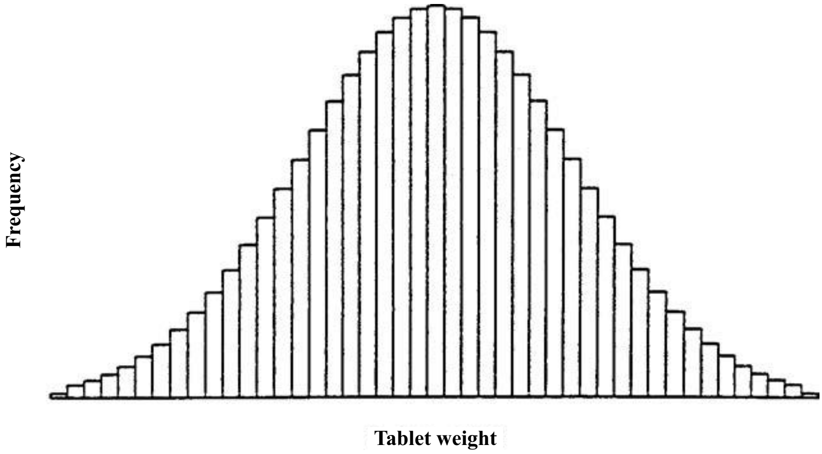


Figure 3.12 Histogram of tablet weights with small class intervals.

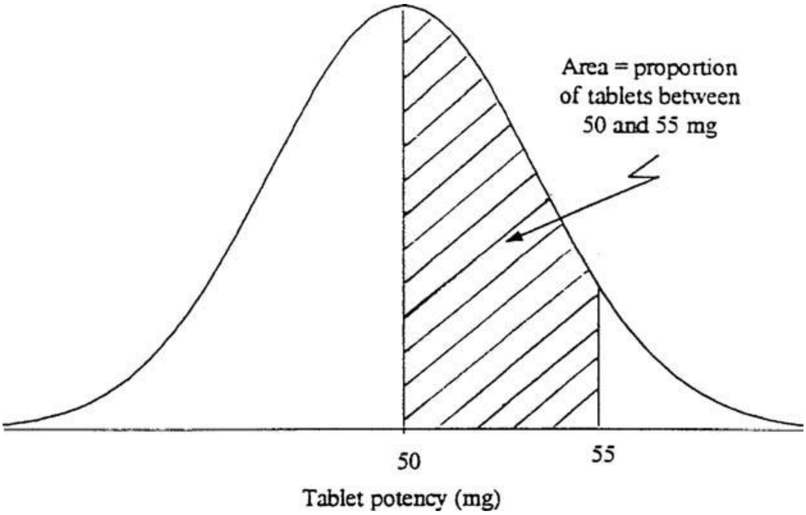


Figure 3.13 Area under normal curve as a representation of proportion of tablets in an interval.

This function is not easily integrated. However, tables are available that can be used to obtain the area between any two values of the variable, Z. Such an area is illustrated in Figure 3.7(A). If the area between Z_1 and Z_2 in Figure 3.7 is 0.3, the probability of observing a value between Z_1 and Z_2 is 3 in 10 or 0.3. In the case of the tablet potencies, the area in a specified interval can be thought of as the proportion of tablets in the batch contained in the interval. This concept is illustrated in Figure 3.13.

Probabilities can be determined directly from the cumulative distribution plot as shown in Figure 3.7(B) (see Exercise Problem 9). The probability of observing a value below Z_1 is 0.6. Therefore, the probability of observing a value between Z_1 and Z_2 is $0.9 - 0.6 = 0.3$.

There are an infinite number of normal curves depending on μ and σ. However, the area in any interval can be calculated from tables of cumulative areas under the *standard normal curve.* The standard normal curve has a mean of 0 and a s.d. of 1. Table IV.2 in App. IV is a table of cumulative areas under the standard normal curve, giving the area below Z (i.e., the area between $-\infty$ and Z). For example, for $Z = 1.96$, the area in Table IV.2 is 0.975. This means that 97.5% of the values comprising the standard normal curve are less than 1.96, lying between $-\infty$ and 1.96. The normal curve is symmetrical about its mean. Therefore, the area below -1.96 is 0.025 as depicted in Figure 3.14. The area between Z equal to -1.96 and $+1.96$ is 0.95. Referring to Table IV.2, the area below Z equal to $+2.58$ is 0.995, and the area below $Z = -2.58$ is 0.005. Thus the area between Z equal to -2.58 and $+2.58$ is 0.99. It would be very useful for the reader to memorize the Z values and the corresponding area between $\pm Z$ as shown in Table 3.4. These values of Z are commonly used in statistical analyses and tests.

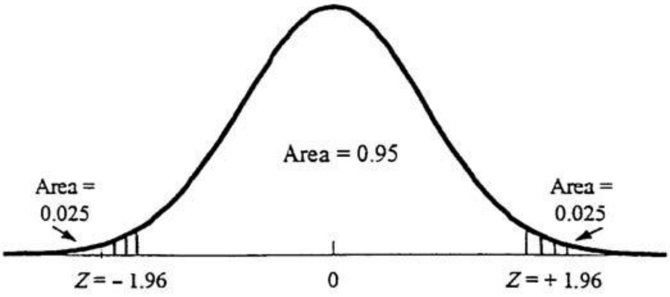


Figure 3.14 Symmetry of the normal curve.

Table 3.4 Area Between ±(for Some Commonly Used Values of *Z*

Z	Area between ±*Z*
0.84	0.60
1.00	0.68
1.28	0.80
1.65	0.90
1.96	0.95
2.32	0.98
2.58	0.99

The area in any interval of a normal curve with a mean and s.d. different from 0 and 1, respectively, can be computed from the standard normal curve table by using a transformation. The transformation changes a value from the normal curve with mean μ and s.d. σ, to the corresponding value, Z, in the standard normal curve. The transformation is

$$Z = \frac{X - \mu}{\sigma}. \tag{3.14}$$

The area (probability) between $-\infty$ and X (i.e., the area below X) corresponds to the value of the area below Z from the cumulative standard normal curve table. Note that if the normal curve that we are considering is the standard normal curve itself, transformation results in the identity

$$Z = \frac{X - 0}{1} = X.$$

Z is exactly equal to X, as expected. Effectively the transformation changes variables with a mean of μ and a s.d. of σ to variables with a mean of 0 and a s.d. of 1.

Suppose in the example of tablet potencies that the mean is 50 and the s.d. is 5 mg. Given these two parameters, what proportion of tablets in the batch would be expected to have more than 58.25 mg of drug? First we calculate the transformed value, Z. Then the desired proportion (equivalent to probability) can be obtained from Table IV.2. In this example, $X = 58.25$, $\mu = 50$, and $\sigma = 5$. Referring to Eq. (3.14), we have

$$\begin{aligned} Z &= \frac{X - \mu}{\sigma} \\ &= \frac{58.25 - 50}{5} = 1.65. \end{aligned}$$

According to Table IV.2, the area between $-\infty$ and 1.65 is 0.95. This represents the probability of a tablet having 58.25 mg or less of drug. Since the question was, "What proportion of tablets in the batch have a potency greater than 58.25 mg?", the area above 58.25 mg is the correct answer. The area under the entire curve is 1; the area above 58.25 mg is $1 - 0.95$, equal to 0.05. This is equivalent to saying that 5% of the tablets have at least 58.25 mg (58.25 mg or more) of drug in this particular batch or distribution of tablets. This transformation is illustrated in Figure 3.15.

One should appreciate that since the normal distribution is a perfectly symmetrical continuous distribution that extends from $-\infty$ to $+\infty$, real data never exactly fit this model. However, data from distributions reasonably similar to the normal can be treated as being normal, with the understanding that probabilities will be approximately correct. As the data are closer to normal, the probabilities will be more exact. Methods exist to test if data can reasonably be expected to be derived from a normally distributed population [1]. In this book, when applying the normal distribution to data we will either (a) assume that the data are close to

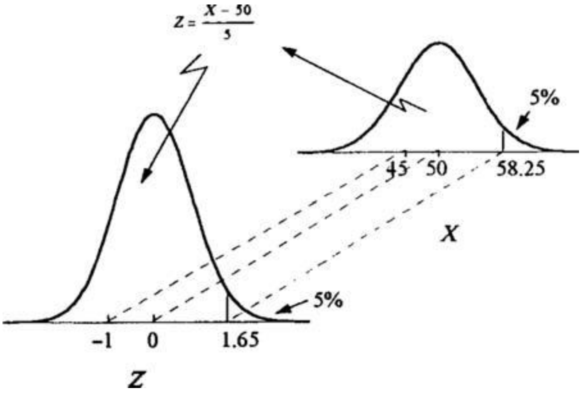


Figure 3.15 *Z* transformation for tablets with mean of 50 mg and s.d. of 5 mg.

normal according to previous experience or from an inspection of the data, or (b) that deviations from normality will not greatly distort the probabilities based on the normal distribution.

Several examples are presented below which further illustrate applications of the normal distribution.

Example 1: The U.S. Pharmacopia (USP) weight test for tablets states that for tablets weighing up to 100 mg, not more than 2 of 20 tablets may differ from the average weight by more than 10%, and no tablet may differ from the average weight by more than 20% [2]. To ensure that batches of a 100-mg tablet (labeled as 100 mg) will pass this test consistently, a statistician recommended that 98% of the tablets in the batch should weigh within 10% of the mean. One thousand tablets from a batch of 3,000,000 were weighed and the mean and s.d. were calculated as 101.2 ± 3.92 mg. Before performing the official USP test, the quality control supervisor wishes to know if this batch meets the statistician's recommendation. The calculation to answer this problem can be made by using areas under the standard normal curve if the tablet weights can be assumed to have a distribution that is approximately normal. For purposes of this example, the sample mean and s.d. will be considered equal to the true batch mean and s.d. Although not exactly true, the sample estimates will be close to the true values when a sample as large as 1000 is used. For this large sample size, the sample estimates are very close to the true parameters. However, one should clearly understand that to compute probabilities based on areas under the normal curve, both the mean and s.d. must be known. When these parameters are estimated from the sample statistics, other derived distributions can be used to calculate probabilities.

Figure 3.16 shows the region where tablet weights will be outside the limits, 10% from the mean ($\mu \pm 0.1\ \mu$), that is, 10.12 mg or more from the mean for an average tablet weight of 101.2 mg (101.2 ± 10.12 mg). The question to be answered is: What proportion of tablets are between

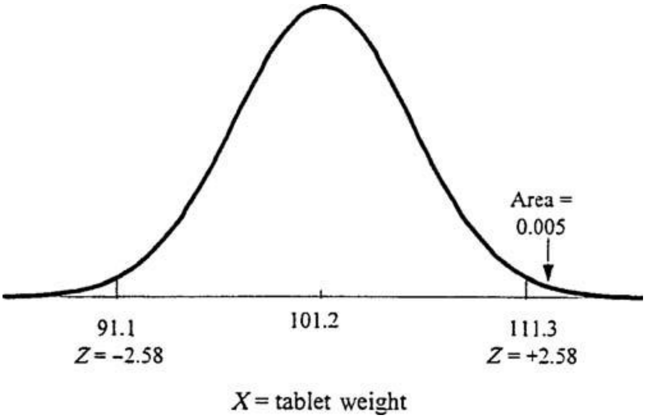


Figure 3.16 Distribution of tablets with mean weight 101.2 mg and s.d. equal to 3.92.

91.1 and 111.3 mg? If the answer is 98% or greater, the requirements are met. The proportion of tablets between 91.1 and 111.3 mg can be estimated by computing the area under the normal curve in the interval 91.1 to 111.3, the unshaded area in Figure 3.16. This can be accomplished by use of the Z transformation and the table of areas under the standard normal curve (Table IV.2). First, we calculate the areas below 111.3 by using the Z transformation

$$Z = \frac{X - \mu}{\sigma} = \frac{111.3 - 101.2}{3.92} = 2.58.$$

This corresponds to an area of 0.995 (see Table IV.2). The area above 111.3 is $(1 - 0.995) = 0.005$ or 1/200. Referring to Figure 3.16, this area represents the probability of finding a tablet that weighs 111.3 mg or more. The probability of a tablet weighing 91.1 mg or less is calculated in a similar manner

$$Z = \frac{91.1 - 101.2}{3.92} = -2.58.$$

Table IV.2 shows that this area is 0.005; that is, the probability of a tablet weighing between $-\infty$ and 91.1 mg is 0.005. The probability that a tablet will weigh more than 111.3 mg or less than 91.1 mg is $0.005 + 0.005$, equal to 0.01. Therefore, 99% $(1.00 - 0.01)$ of the tablets weigh between 91.1 and 111.3 mg and the statistician's recommendation is more than satisfied. The batch should have no trouble passing the USP test.

The fact that the normal distribution is symmetric around the mean simplifies calculations of areas under the normal curve. In the example above, the probability of values exceeding Z equal to 2.58 is exactly the same as the probability of values being less than Z equal to -2.58. This is a consequence of the symmetry of the normal curve, 2.58 and -2.58 being equidistant from the mean. This is easily seen from an examination of Figure 3.16.

Although this batch of tablets should pass the USP weight uniformity test, if *some* tablets in the batch are out of the 10% or 20% range, there is a chance that a random sample of 20 will fail the USP test. In our example, about 1% or 30,000 tablets will be more than 10% different from the mean (less than 91.1 or more than 111.3 mg). It would be of interest to know the chances, albeit small, that of 20 randomly chosen tablets, more than 2 would be "aberrant." When 1% of the tablets in a batch deviate from the batch mean by 10% or more, the chances of finding more than 2 such tablets in a sample of 20 is approximately 0.001 (1/1000). This calculation makes use of the binomial probability distribution.

Example 2: During clinical trials, serum cholesterol, among other serum components, is frequently monitored to ensure that a patient's cholesterol is within the normal range, as well as to observe possible drug effects on serum cholesterol levels. A question of concern is: What is an abnormal serum cholesterol value? One way to define "abnormal" is to tabulate cholesterol values for apparently normal healthy persons, and to consider values very remote from the average as abnormal. The distribution of measurements such as serum cholesterol often has an approximately normal distribution.

The results of the analysis of a large number of "normal" cholesterol values showed a mean of 215 mg% and a s.d. of 35 mg%. This data can be depicted as a normal distribution as shown in Figure 3.17. "Abnormal" can be defined in terms of the proportion of "normal" values that fall in the extremes of the distribution. This may be thought of in terms of a gamble. By choosing to say that extreme values observed in a new patient are abnormal, we are saying that persons observed to have very low or high cholesterol levels could be "normal," but the likelihood or probability that they come from the population of normal healthy persons is small. By defining an abnormal cholesterol value as one that has a 1 in 1000 chance of coming from the distribution of values from normal healthy persons, cutoff points can be defined for abnormality based on the parameters of the normal distribution. According to the cumulative standard normal curve, Table IV.2, a value of Z equal to approximately 3.3 leaves 0.05% of the area in the upper tail. Because of the symmetry of the normal curve, 0.05% of the area is below $Z = -3.3$. Therefore, 0.1% (1/1000) of the values will lie outside the values of Z equal to ± 3.3

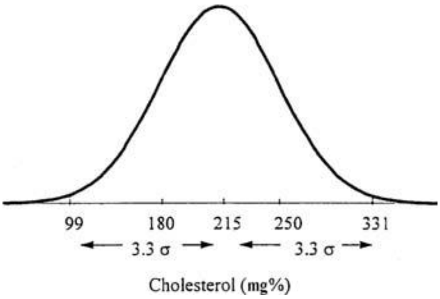


Figure 3.17 Distribution of "normal" cholesterol values.

in the standard normal curve. The values of X (cholesterol levels) corresponding to $Z = \pm 3.3$ can be calculated from the Z transformation.

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 215}{35} = \pm 3.3$$

$$X = 215 \pm (3.3)(35) = 99 \text{ and } 331.$$

This is equivalent to saying that cholesterol levels that deviate from the average of "normal" persons by 3.3 s.d. units or more are deemed to be abnormal. For example, the lower limit is the mean of the "normals" minus 3.3 times the s.d. or 215 − (3.3)(35) = 99. The cutoff points are illustrated in Figure 3.17.

Example 3: The standard normal distribution may be used to calculate the proportion of values in any interval from any normal distribution. As an example of this calculation, consider the data of cholesterol values in Example 2. We may wish to calculate the proportion of cholesterol values between 200 and 250 mg%.

Examination of Figure 3.18 shows that the area (probability) under the normal curve between 200 and 250 mg% is the probability of a value being less than 250 *minus* the probability of a value being less than 200. Referring to Table IV.2, we have

Probability of a value less than 250

$$\frac{250 - 215}{35} = 1 = Z \text{probability} = 0.841.$$

Probability of a value less than 200

$$\frac{200 - 215}{35} = -0.429 = Z \text{probability} = 0.334.$$

Therefore, the probability of a value falling between 250 and 200 is

$$0.841 - 0.334 = 0.507.$$

3.4.2 Central Limit Theorem

"Without doubt, the most important theorem in statistics is the central limit theorem" [3]. This theorem states that the distribution of sample means of size N taken from *any* distribution with a finite variance σ^2 and mean μ tends to be *normal* with variance σ^2/N and mean μ. We have previously discussed the fact that a sample mean of size N has a variance equal to σ^2/N. The new and important feature here is that if we are dealing with means of *sufficiently large sample size*, the means have a normal distribution, regardless of the form of the distribution from which the samples were selected.

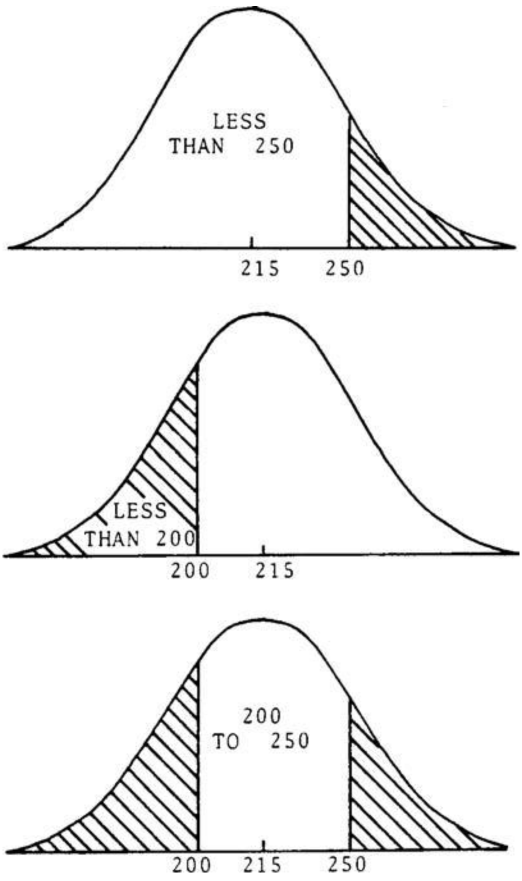


Figure 3.18 Illustration of the calculation of proportion of cholesterol values between 200 and 250 mg%.

How large is a "large" sample? The answer to this question depends on the form of the distribution from which the samples are taken. If the distribution is normal, any size sample will have a mean that is normally distributed. For distributions that deviate greatly from normality, larger samples will be needed to approximate normality than distributions that are more similar to the normal distributions (e.g., symmetrical distributions).

The power of this theorem is that the normal distribution can be used to describe most of the data with which we will be concerned, provided that the means come from samples of sufficient size. An example will be presented to illustrate how means of distributions far from normal tend to be normally distributed as the sample size increases. Later in this chapter, we will see that even the discrete binomial distribution, where only a very limited number of outcomes are possible, closely approximates the normal distribution with sample sizes as small as 10 in symmetrical cases (e.g. $P = q = 0.5$).

Consider a distribution that consists of outcomes 1, 2, and 3 with probabilities depicted in Figure 3.19. The probabilities of observing values of 1, 2, and 3 are 0.1, 0.3, and 0.6, respectively. This is an asymmetric distribution, with only three discrete outcomes. The mean is 2.5. Sampling from this population can be simulated by placing 600 tags marked with the number 3, 300 tags marked with the number 2, and 100 tags marked with the number 1 in a box. We will mix up the tags, select 10 (replacing each tag and mixing after each individual selection), and *compute the mean of the 10 samples.* A typical result might be five tags marked 3, four tags marked 2, and one tag marked 1, an average of 2.4. With a computer or programmable calculator, we can simulate this drawing of 10 tags. The distributions of 100 such means for samples of sizes 10 and 20 obtained from a computer simulation are shown in Figure 3.20. The distribution is closer to normal as the sample size is increased from 10 to 20. This is an empirical demonstration of

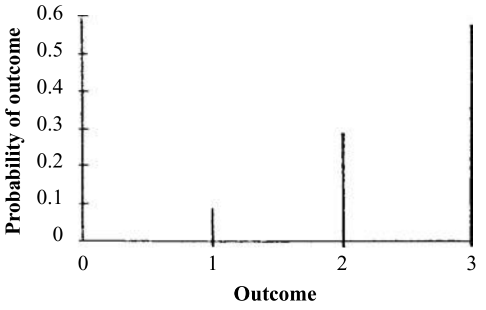


Figure 3.19 Probability distribution of outcomes 1, 2, and 3.

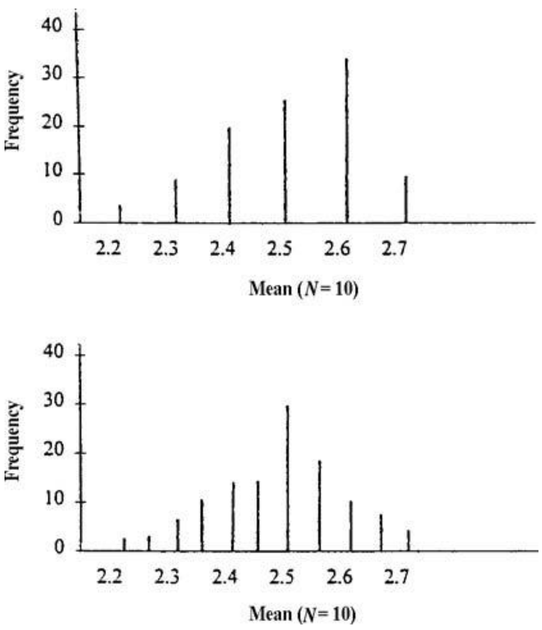


Figure 3.20 Distribution of means of sizes 10 and 20 from population shown in Figure 3.19.

the central limit theorem. Of course, under ordinary circumstances, we would not draw 100 samples each of size 10 (or 20) to demonstrate a result that can be proved mathematically.

3.4.3 Normal Approximation to the Binomial

A very important result in statistical theory is that the binomial probability distribution can be approximated by the normal distribution if the sample size is sufficiently large (see sect. 3.4.2). A conservative rule of thumb is that if *NP* (the product of the number of observations and the probability of success) and *Nq* are both greater than or equal to 5, we can use the normal distribution to approximate binomial probabilities. With symmetric binomial distributions, when $P = q = 0.5$, the approximation works well for *NP* less than 5.

To demonstrate the application of the normal approximation to the binomial, we will examine the binomial distribution described above, where $N = 10$ and $p = 0.5$. We can superimpose a normal curve over the binomial with $\mu = 5$ (number of successes) and standard deviation $\sqrt{NPq} = \sqrt{10(0.5)(0.5)} = 1.58$, as shown in Figure 3.21.

The probability of a discrete result can be calculated by using the binomial probability [Eq. (3.9)] or Table IV.3. The probability of seven successes, for example, is equal to 0.117. In a normal distribution, the probability of a single value cannot be calculated. We can only calculate the probability of a range of values within a specified interval. The area that approximately

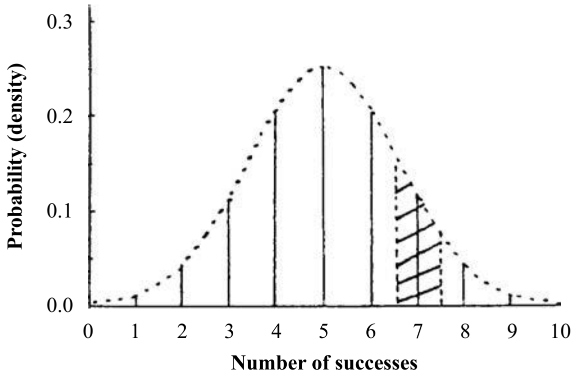


Figure 3.21 Normal approximation to binomial distribution: $NP = 5$ and s.d. $= 1.58$.

corresponds to the probability of observing seven successes in 10 trials is the area between 6.5 and 7.5, as illustrated in Figure 3.21. This area can be obtained by using the Z transformation discussed earlier in this chapter [Eq. (3.14)]. The area between 6.5 and 7.5 is equal to the area below 7.5 minus the area below 6.5.

Area below 6.5 $Z = \frac{6.5-5}{1.58} = 0.948$ from Table IV.2, area $= 0.828$.
Area below 7.5
$Z = \frac{7.5-5}{1.58} = 1.58$ from Table IV.2, area $= 0.943$.
Therefore, the area (probability) between 6.5 and 7.5 is

$$0.943 - 0.828 = 0.115.$$

This area is very close to the exact probability of 0.117.

The use of $X \pm 0.5$ to help estimate the probability of a discrete value, X, by using a continuous distribution (e.g., the normal distribution) is known as a continuity correction. We will see that the continuity correction is commonly used to improve the estimation of binomial probabilities by the normal approximation (chap. 5).

Most of our applications of the binomial distribution will involve data that allow for the use of the normal approximation to binomial probabilities. This is convenient because calculations using exact binomial probabilities are tedious and much more difficult than the calculations using the standard normal cumulative distribution (Table IV.2), particularly when the sample size is large.

3.5 OTHER COMMON PROBABILITY DISTRIBUTIONS

3.5.1 The Poisson Distribution

Although we will not discuss this distribution further in this book, the Poisson distribution deserves some mention. The Poisson distribution can be considered to be an approximation to the binomial distribution when the sample size is large and the probability of observing a specific event is small. In quality control, the probability of observing a defective item is often calculated by using the Poisson. The probability of observing X events of a given kind in N observations, where the probability of observing the event in a single observation is P, is

$$P(X) = \frac{\lambda^X e^{-\lambda}}{X!} \tag{3.15}$$

where $\lambda = NP$, e the base of natural logarithms (2.718...), and N the number of observations.

We may use the Poisson distribution to compute the probability of finding one defective tablet in a sample of 100 taken from a batch with 1% defective tablets. Applying Eq. (3.15), we have

$$N = 100 \quad P = 0.01 \quad NP = \lambda = (100)(0.01) = 1$$

$$P(1) = \frac{(1)^1(e^{-1})}{1!} = e^{-1} = 0.368.$$

The exact probability calculated from the binomial distribution is 0.370. (See Exercise Problem 8.)

3.5.2 The *t* Distribution ("Student's *t*")

The *t* distribution is an extremely important probability distribution. This distribution can be constructed by repeatedly taking samples of size N from a normal distribution and computing the statistic

$$t = \frac{\bar{X} - \mu}{S/\sqrt{N}},$$

where $\bar{X}$ is the sample mean, μ the true mean of the normal distribution, and S the sample standard deviation. The distribution of the t's thus obtained forms the t distribution. The exact shape of the t distribution depends on sample size (degrees of freedom), but the t distribution is symmetrically distributed about a mean of zero, as shown in Figure 3.22(A).

To elucidate further the concept of a sampling distribution obtained by repeated sampling, as discussed for the t distribution above, a simulated sampling of 100 samples each of size 4 ($N = 4$) was performed. These samples were selected from a normal distribution with mean 50 and standard deviation equal to 5, for this example. The mean and standard deviation of each sample of size 4 were calculated and a t ratio [Eq. (3.16)] constructed.

The distribution of the 100 t values thus obtained is shown in Table 3.5. The data are plotted (histogram) together with the theoretically derived t distribution with 3 degrees of freedom ($N - 1 = 4 - 1 = 3$) in Figure 3.23. Note that the distribution is symmetrically centered around a mean of 0, and that 5% of the t values are 3.18 or more units from the mean (theoretically).

3.5.3 The Chi-Square (χ^2) Distribution

Another important probability distribution in statistics is the chi-square distribution. The chi-square distribution may be derived from normally distributed variables, defined as the sum of squares of independent normal variables, each of which has mean 0 and standard deviation 1. Thus, if Z is normal with $\mu = 0$ and $\sigma = 1$,

$$\chi^2 = \sum Z_i^2. \tag{3.17}$$

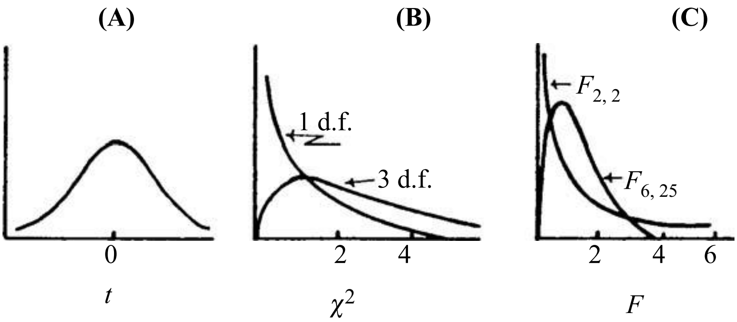


Figure 3.22 Examples of typical probability distributions.

Table 3.5 Frequency Distribution of 100 *t* Values Obtained by Simulated Repeat Sampling from a Normal Distribution with Mean 50 and Standard Deviation 5[a]

Class interval	Frequency
−5.5 to −4.5	1
−4.5 to −3.5	2
−3.5 to −2.5	2
−2.5 to −1.5	11
−1.5 to −0.5	18
−0.5 to +0.5	29
+0.5 to +1.5	21
+1.5 to +2.5	9
+2.5 to +3.5	4
+3.5 to +4.5	2
+4.5 to +5.5	1

[a] Sample size = 4.

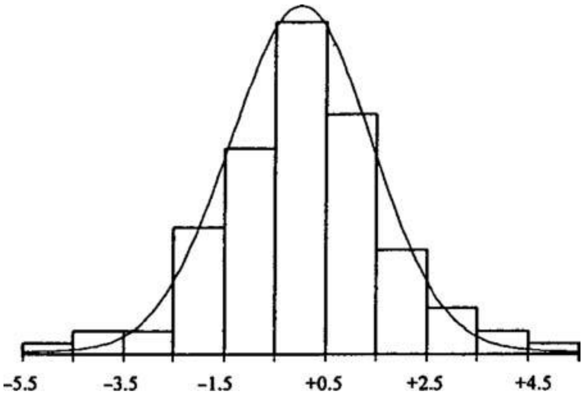


Figure 3.23 Simulated *t* distribution (d.f. = 3) compared to a theoretical *t* distribution.

Applications of the chi-square distribution are presented in chapters 5 and 15. The chi-square distribution is often used to assess probabilities when comparing discrete values from comparative groups, where the normal distribution can be used to approximate discrete probabilities.

As with the *t* distribution, the distribution of chi-square depends on degrees of freedom, equal to the number of independent normal variables as defined in Eq. (3.17). Figure 3.22(B) shows chi-square distributions with 1 and 3 degrees of freedom.

3.5.4 The *F* Distribution

After the normal distribution, the *F* distribution is probably the most important probability distribution used in statistics. This distribution results from the sampling distribution of the ratio of two independent variance estimates obtained from the same normal distribution. Thus, the first sample consists of N_1 observations and the second sample consists of N_2 observations

$$F = \frac{S_1^2}{S_2^2}. \tag{3.18}$$

The *F* distribution depends on two parameters, the degrees of freedom in the numerator $(N_1 - 1)$ and the degrees of freedom in the denominator $(N_2 - 1)$. This distribution is used to test for differences of means (analysis of variance) as well as to test for the equality of two variances. The *F* distribution is discussed in more detail in chapters 5 and 8 as applied to the comparison of two variances and testing of equality of means in the analysis of variance, respectively.

3.6 THE LOG-NORMAL DISTRIBUTION

The log-normal distribution results from a distribution, skewed to the right (see Fig. 10.1). Such data, when transformed into logs, exhibit the properties of a normal distribution. Thus, the logs of the original skewed data are normally distributed. The log-normal distribution is an important distribution in applications to the pharmaceutical sciences. Two important applications are in the analysis of bioequivalence data (see chap. 11) and particle size analysis. Some of the properties of a log-normal distribution are presented at this time as applied to the analysis of particle size of active drug substances and powders, such as excipients. This application is important, as the particle size of ingredients in common dosage forms may profoundly affect the therapeutic activity of the active ingredient.

3.6.1 Statistical Analysis of Particle Size

Typically, to characterize the particle size of a powdered substance, one defines the characteristics of the distribution of particles as previously described in this book (chap. 1), the mean, the standard deviation, and percentiles, particularly the 50th percentile or the median.

If the distribution of the particles followed an approximately normal distribution, the median and mean would be the "identical." However, experience shows that the particle size distribution is typically skewed to the right and follows an approximately log-normal distribution. We will not be concerned with the question of how to measure particle size, but for the present discussion, we will be measuring the diameter of the particles, assuming that the particles are perfect spheres. This measurement is also known as the spherical equivalent diameter. This is the diameter of a sphere that would have the equivalent volume of the irregular-shaped particle [7]. We will assume that the diameters have a log-normal distribution. Consider the data in Table 3.6 [7] that describe the distribution of diameters in the form of a frequency table.

The cumulative distribution is shown in Figure 3.24. Figure 3.24 is a special kind of graph that plots the cumulative distribution of the logs of the diameters on a probability scale, sometimes referred to as a probability plot. In this example, the X axis represents the area under a normal curve in terms of standard deviations (the standard normal curve). The Y axis represents the cumulative distribution of the logarithms of the diameters. Thus, the cumulative distribution of the logs of the diameters is conveniently shown on log-probability paper. If the distribution is log-normal, such a plot should show a straight line. That is, cumulative data plotted on probability paper will show a straight line if the data are normally distributed.

From Table 3.6 and Figure 3.24, the mean, median, and other percentiles, such as the 10th and 90th percentile, as well as the standard deviation can be ascertained for the log-transformed values. For example, referring to Figure 3.24, the median is represented by the 50% point on the probability scale, 10 μm for diameters (see below). Thus, the log-normal distribution can be conveniently characterized by these well-known parameters.

For example, inspection of Figure 3.24 shows that the 90th percentile is approximately 20 μm for the log-transformed diameters. Note, that for a log-normal distribution, the mean is larger than the median based on the original, untransformed numbers. The mean of the original, untransformed diameters is 12.25 mu. The median of the untransformed and transformed numbers does not change, 10 mu, when the median of the logs is back-transformed to the original numbers. The distribution of the logs of the diameters, however, will be normal, and shows a symmetric distribution where the mean and median are the same. Table 3.6 shows particle diameters in intervals (bins), and includes calculations based on the diameters and the mass or weight, as explained below.

As previously noted, typically, the distribution of diameters is considered to have a log-normal distribution. This distribution is based on the frequency of particles in the intervals describing the distribution (The first five columns in Table 3.6). The mean of the untransformed diameters is 12.25 mu as computed for data in the form of frequency tables (see sect. 1.2). Note that the mean of the log-transformed diameters is 2.300. The antilog of 2.300 is 9.97 or approximately 10, the same as the median.

Another way in which the particle distribution is described is by weight. If the density of the particles is considered constant, the weight will be proportional to the cube of the diameter. Note that the weight of a spherical particle is equal to volume x density. The volume of a

Table 3.6 Calculation of Some Average Diameters [7]

Interval	Diameter (*D*) (midsize)	Frequency (*N*)	Cumulative frequency (frequency)	Cum. % frequency (frequency)	Diameter × frequency (*D* × *N*)	ln(diameter)
1–1.4	1.2	2	2	0.2	2.4	0.182
1.4–2.0	1.7	5	7	0.7	8.5	0.531
2.0–2.8	2.4	14	21	2.1	33.6	0.875
2.8–3.6	3.2	60	81	8.1	192	1.163
3.6–6.0	4.8	100	181	18.1	480	1.569
6.0–7.6	6.8	190	371	37.1	1292	1.917
7.6–12.4	10	250	621	62.1	2500	2.303
12.4–15.6	14	160	781	78.1	2240	2.639
15.6–22.4	19	110	891	89.1	2090	2.944
22.4–29.6	26	70	961	96.1	1820	3.258
29.6–42.4	36	28	989	98.9	1008	3.584
42.4–59.6	51	10	999	99.9	510	3.932
59.6–84.4	72	1	1000	100	72	4.277
Total		1000			12248.5	
Average					12.2485	

Interval	Diameter (*D*) (midsize)	*N* × ln (diameter)	Cum. *N* × ln(*D*)	ND^3	Cum. ND^3	Cum. % (cum. ND^3)
1–1.4	1.2	0.365	0.365	3.456	3.5	6.01E-07
1.4–2.0	1.7	2.653	3.018	24.565	28.0	4.87E-06
2.0–2.8	2.4	12.257	15.274	193.536	221.6	3.85E-05
2.8–3.6	3.2	69.789	85.063	1966.08	2187.6	0.00038
3.6–6.0	4.8	156.862	241.925	11059.2	13246.8	0.002303
6.0–7.6	6.8	364.215	606.140	59742.08	72988.9	0.012687
7.6–12.4	10	575.646	1181.787	250000	322988.9	0.056143
12.4–15.6	14	422.249	1604.036	439040	762028.9	0.132458
15.6–22.4	19	323.888	1927.924	754490	1516518.9	0.263606
22.4–29.6	26	228.067	2155.991	1230320	2746838.9	0.477465
29.6–42.4	36	100.339	2256.329	1306368	4053206.9	0.704542
42.4–59.6	51	39.318	2295.648	1326510	5379716.9	0.935121
59.6–84.4	72	4.277	2299.924	373248	5752964.9	1
Total		2299.9242		5752964.917		
Average		2.2999242		5752.964917		

Average diameter = 12.2485; Standard deviation = 8.52766; CV = s.d./average diameter = 0.69622; GSD = Geometric s.d. from Figure 3.24 = ~1.9; CV = ~0.745; ln(GSD) = 0.642.

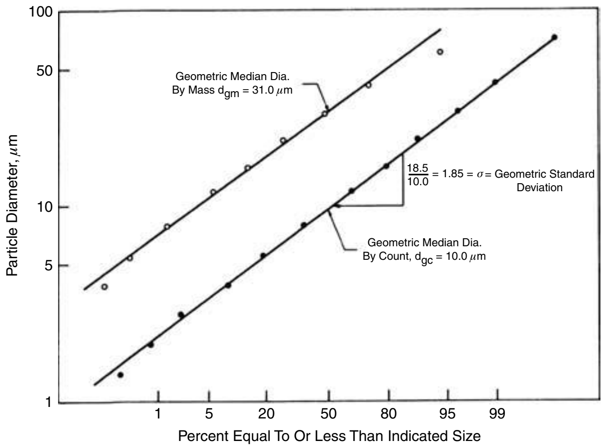


Figure 3.24 Log-probability plot of distribution from Table 3.6 [7].

sphere is 4/3 pi (radius)^3. The distribution can be described by the cube of the diameters (which are proportional to the weights of the particles) in each interval (see the last three columns in Table 3.6). The median weight is sometimes defined by particle size analysts as the particle *diameter* below which 50% of the weights lie. This is not the typical definition of the median. Using the usual definition, the median weight would be that weight that would have 50% of the weight below (and above) that value. For example, in Table 3.6, the usual definition would be the 50% cumulative point in the column labeled "Cum ND,^{3}" a value somewhat greater than 2,746,839, because that defines the weight that has 50% of the weight above and below this value. However, the median weight is often described in particle size analysis as *the diameter that corresponds to the median weight*, not the usual definition. From Figure 3.6 and Table 3.24, this value is 31 mu. Because this definition of the median weight, as often described by particle size analysts, is different from the usual definition, and there are several definitions of the median particle size [7,8], *one should clearly explain how the median is derived.*

Thus, there is some disconnect in these definitions. For example, the mean of the distribution would depend on whether we are talking about diameters or weights, and how the values are determined. For example, the determination of the mean of the particle distribution could be determined on a weight basis or diameter basis. Not only would the answers be different depending on the definition, but the distribution of particles will also depend on the definitions, whether we are talking about weight or diameters.

Other parameters can be determined from Figure 3.24 and Table 3.6. The coefficient of variation (CV = s.d./mean) for the diameters from Table 3.6 is 8.528/12.249 = 0.696. The geometric standard deviation (GSD) can be determined from the log-probability plot (Fig. 3.24). The standard deviation can be read from the plot as the distance of the cumulative diameters between the 50th percentile and 84th percentile. Note that one standard deviation encompasses 34% of the area above the mean or median. This GSD is approximately 1.9. It can be shown that the log (ln) of the GSD is equal to the CV of the untransformed data, equals approximately, 0.64. This is close to the observed CV of the untransformed diameters, 0.696. The CV based on the log-transformed diameters can be estimated by using the following equation:

$$\begin{aligned} \mathrm{CV} &= \surd\{e^{(\mathrm{GSD2}-1)}\} \\ &= \surd\{e^{(0.642-1)}\} = 0.74. \end{aligned}$$

Note that these are theoretical concepts, so that one does not expect the observed and theoretical values to be identical. Of course, we do not expect data to exactly conform to a log-normal distribution, just as we do not expect real data to exactly conform to a normal distribution. (For more detailed discussion of particle size analysis, see Refs. [7,8].)

KEY TERMS

Binomial distribution
Binomial formula
Binomial trial
Central limit theorem
Chi-square distribution
Combinations
Conditional probability
Continuous distribution
Cumulative distribution
Density function
Discontinuous variable
Discrete distribution
Distribution
Equally likely
Event
Failure
F distribution
Independent events
Log-normal distribution
Multiplicative probability
Mutually exclusive
Negative binomial distribution
Normal distribution
Outcome
Poisson distribution
Population
Probability distribution
Proportion
Random
Randomly chosen
Standard normal distribution
Success
t distribution
Variability
Factorial
Z transformation

EXERCISES

1. Explain why do you think that a controlled multicenter clinical study better estimates the probability of a patient responding to treatment than the observations of a single physician in daily practice.
2. Describe the population that represents the multicenter antibiotic clinical study described in section 3.3.
3. Give three examples of probability distributions that describe the probability of outcomes in terms of attributes.
4. Explain why 30,000 tablets are only specked if 20,000 tablets are both chipped and specked as described in section 3.2. What is the probability, in the example described in section 3.2, of finding a specked tablet *or* a chipped tablet? (Hint: Count all the tablets that have either a speck or a chip.) See Eq. (3.4).
5. In a survey of hospital patients, it was shown that the probability that a patient has high blood pressure given that he or she is diabetic was 0.85. If 10% of the patients are diabetic and 25% have high blood pressure:
   (a) What is the probability that a patient has both diabetes and high blood pressure?
   (b) Are the conditions of diabetes and high blood pressure independent? [Hint: See Eqs. (3.5), (3.6), and (3.7).]
6. Show how the result 0.21094 is obtained for the probability of two of four patients being cured if the probability of a cure is 0.75 for each patient and the outcomes are independent (Table 3.2). (Enumerate all ways in which two of four patients can be cured, and compute the probability associated with each of these ways.)
7. What is the probability that three of six patients will be cured if the probability of a cure is 60%?
8. Calculate the probability of one success in 100 trials if $p = 0.01$.
9. From the cumulative plot in Figure 3.7(B), estimate the probability that a value, selected at random, will be (a) greater than Z_0; (b) less than Z_0.
10. What is the probability that a normal patient has a cholesterol value below 170 ($\mu = 215$, $\sigma = 35$)?
11. If the mean and standard deviation of the potency of a batch of tablets are 50 and 5 mg, respectively, what proportion of the tablets have a potency between 40 and 60 mg?
12. If a patient has a serum cholesterol value outside normal limits, does this mean that the patient is abnormal in the sense of having a disease or illness?
13. Serum sodium values for normal persons have a mean of 140 mEq/L and a s.d. of 2.5. What is the probability that a person's serum sodium will be between 137 and 142 mEq/L?
14. Data were collected over many years on cholesterol levels of normal persons in a New York hospital with the following results based on 100,000 readings. The mean is 205 mg%; the s.d. is 45. Assuming that the data have a normal distribution, what is the probability that a normal patient has a value greater than 280 mg%?
15. In the game of craps, two dice are thrown, each dice having an equal probability of showing one of the numbers 1 to 6 inclusive. Explain why the probability of observing a point of 2 (the sum of the numbers on the two dice) is 1/36.
16. Is the probability of observing two heads and one tail the same under the two following conditions: (a) simultaneously throwing three coins; (b) tossing one coin three consecutive times? Explain your answer.
17. What odds would you give of finding either none *or* one defective tablet in a sample of size 20 if the batch of tablets has 1% defective? Answer the same question if the sample size is 100.

18. What is the probability of observing exactly one head in 10 tosses of a coin?

§§19. The chance of obtaining a cure using conventional treatment for a particular form of cancer is 1%. A new treatment being tested cures two of the first four patients tested. Would you announce to the world that a major breakthrough in the treatment of this cancer is imminent? Explain your answer.

20. What is the s.d. for the binomial experiments described in Problems 17 and 19? (Answer in terms of *NPq* and *Pq/N*.)

§§21. In screening new compounds for pharmacological activity, the compound is administered to 20 animals. For a standard drug, 50% of the animals show a response on the average. Fifteen of the twenty animals show the response after administration of a new drug. Is the new drug a promising candidate? Why? [Hint: Compute the s.d. of the response based on $p = 0.5$. See if the observed response is more than 2 s.d.'s greater than 0.5.]

22. Using the binomial formula, calculate the probability that a sample of 30 tablets will show 0 or 1 defect if there are 1% defects in the batch. (What is the probability that there will be more than one defect in the sample of 30?)

23. The following expression can be used to calculate the probability of observing *A* or *B* or *C* (or any combination of *A*, *B*, *C*)

$$P(A \text{ or } B \text{ or } C) = P(A) + P(B) + P(C) - P(A \text{ and } B)$$
$$-P(A \text{ and } C) - P(B \text{ and } C) + P(A \text{ and } B \text{ and } C).$$

A survey shows that 85% of people with colds have cough, rhinitis, pain, or a combination of these symptoms. Thirty-five percent have at least cough, 50% have at least rhinitis, and 50% have at least pain. Twenty percent have (at least) cough and rhinitis, 15% have cough and pain, and 25% have rhinitis and pain. What percentage have all three symptoms?

REFERENCES

1. Hald A. Statistical Theory and Engineering Applications. New York: Wiley, 1965.
2. United States Pharmacopeia, 23rd Rev., and National Formulary, 18th ed. Rockville, MD: USP Pharmacopeial Convention, Inc., 1995.
3. Ostle B. Statistics in Research, 3rd ed. Ames, IA: Iowa State University Press, 1981.
4. Dixon WJ, Massey FJ Jr. Introduction to Statistical Analysis, 3rd ed. New York: McGraw-Hill, 1969.
5. Feller W. Probability Theory and Its Applications. Wiley and Sons, NY: Wiley, 1950:218.
6. http://en.wikipedia.org/wiki/Negative_binomial_distribution.
7. Stockham JD, Fochtman EG, eds. Particle Size Analysis. Ann Arbor, MI: Ann Arbor Science, 1977.
8. Jillavenkatesa A, Dapkunas S, Lum LH. Particle Size Characterization. Washington, D.C.: Material Science and Engineering Laboratory, U.S. Department of Commerce, National Institute of Standards and Technology, 2001:960–961.

§§ Optional, more difficult problems.

4 Choosing Samples

The samples are the units that provide the experimental observations, such as tablets sampled for potency, patients sampled for plasma cholesterol levels, or tablets inspected for defects. The sampling procedure is an essential ingredient of a good experiment. An otherwise excellent experiment or investigation can be invalidated if proper attention is not given to choosing samples in a manner consistent with the experimental design or objectives. Statistical treatment of data and the inference based on experimental results depend on the sampling procedure. The way in which samples should be selected is not always obvious, and requires careful thought.

The implementation of the sampling procedure may be more or less difficult depending on the experimental situation, such as that which we may confront when choosing patients for a clinical trial, sampling blends, or choosing tablets for quality control tests. In this chapter, we discuss various ways of choosing samples and assigning treatments to experimental units (e.g., assigning different drug treatments to patients). We will briefly discuss various types of sampling schemes, such as simple random sampling, stratified sampling, systematic sampling, and cluster sampling. In addition, the use of random number tables to assign experimental units to treatments in designed experiments will be described.

4.1 INTRODUCTION

There are many different ways of selecting samples. We all take samples daily, although we usually do not think of this in a statistical sense. Cooks are always sampling their wares, tasting the soup to see if it needs a little more spice, or sampling a gravy or sauce to see if it needs more mixing. When buying a car, we take a test ride in a "sample" to determine if it meets our needs and desires.

The usual purpose of observing or measuring a property of a sample is to make some inference about the population from which the sample is drawn. In order to have reasonable assurance that we will not be deceived by the sample observations, we should take care that the samples are not biased. We would clearly be misled if the test car was not representative of the line, but had somehow been modified to entice us into a sale. We can never be sure that the sample we observe mirrors the entire population. If we could observe the entire population, we would then know its exact nature. However, 100% sampling is virtually never done. (One well-known exception is the U.S. census.) It is costly, time consuming, and may result in erroneous observations. For example, to inspect each and every one of 2 million tablets for specks, a tedious and time consuming task, would probably result in many errors due to fatigue of the inspectors.

Destructive testing precludes 100% sampling. To assay each tablet in a batch does not make sense. Under ordinary circumstances, no one would assay every last bit of bulk powder to ensure that it is not adulterated.

The sampling procedure used will probably depend on the experimental situation. Factors to be considered when devising a sampling scheme include

1. The nature of the population. For example, can we enumerate the individual units, such as packaged bottles of a product, or is the population less easily defined, as in the case of hypertensive patients?
2. The cost of sampling in terms of both time and money.
3. Convenience. Sometimes it may be virtually impossible to carry out a particular sampling procedure.
4. Desired precision. The accuracy and precision desired will be a function of the sampling procedure and sample size.

Sampling schemes may be roughly divided into *probability sampling* and *nonprobability sampling* (sometimes called authoritative sampling). Nonprobability sampling methods often are conceptually convenient and simple. These methods are considered as methods of *convenience* in many cases. Samples are chosen in a particular manner because alternatives are difficult. For example, when sampling powder from 10 drums of a shipment of 100 drums, those drums that are most easily accessible might be the ones chosen. Or, when sampling tablets from a large container, we may conveniently choose from those at the top. A "judgment" sample is chosen with possible knowledge that some samples are more "representative" than others, perhaps based on experience. A quality control inspector may decide to inspect a product during the middle of a run, feeling that the middle is more representative of the "average" product than samples obtained at the beginning or end of the run. The inspector may also choose particular containers for inspection based on knowledge of the manufacturing and bottling procedures. A "haphazard" sample is one taken without any predetermined plan, but one in which the sampler tries to avoid bias during the sampling procedure. Nonprobability samples often have a hidden bias, and it is not possible to apply typical statistical methods to estimate the population parameters (e.g., μ and σ) and the precision of the estimates. Nonprobability sampling methods should not be used unless probability sampling methods are too difficult or too expensive to implement.

We will discuss procedures and some properties of common *probability sampling methods.* Objects chosen to be included in probability samples have a known probability of being included in the sample and are chosen by some random device.

4.2 RANDOM SAMPLING

Simple *random sampling* is a common way of choosing samples. A random sample is one in which each individual (object) in the population to be sampled has an *equal chance of being selected.* The procedure of choosing a random sample can be likened to a bingo game or a lottery where the individuals (tags, balls, tablets, etc.) are thoroughly mixed, and the sample chosen at "random." This ensures that there is no bias; that is, *on the average*, the estimates of the population parameters (e.g., the mean) will be accurate. However, one should be aware, that in any single sample, random sampling does not ensure an accurate estimate of the mean and/or the standard deviation. An example of the lack of reliability of small samples can be shown based on the batting statistics of a 0.250 hitter in a given game. We would expect one hit in four times at bat, on the average. Suppose the batter comes up four times in the game. What is the probability that he will get exactly one hit? Applying the binomial theorem, the probability is 42%. See Problem number 12 at the end of this chapter.

Many statistical procedures are based on an assumption that samples are chosen at random. Simple random sampling is most effective when the variability is relatively small and uniform over the population [1].

In most situations, it is not possible to mix the objects that constitute the population and pick the samples out of a "box." But if all members of the population can be identified, a unique identification, such as a number, can be assigned to each individual unit. We can then choose the sample by picking numbers, randomly, from a box using a lottery-like technique. Usually, this procedure is more easily accomplished through the use of a table of random numbers. Random numbers have been tabulated extensively [2]. In addition to available tables, computer-generated random numbers may be used to select random samples or to assign experimental units randomly to treatments as described below.

4.2.1 Table of Random Numbers

Random numbers are frequently used as a device to choose samples to be included in a survey, a quality control inspection sample, or to assign experimental units to treatments such as assigning patients to drug treatments. The first step that is often necessary in the application of a table of random numbers is to assign a number to each of the experimental units in the population or to the units potentially available for inclusion in the sample. The numbers are assigned consecutively from 1 to N, where N is the number of units under consideration. The experimental units may be patients to be assigned to one of two treatments or bottles of tablets to be inspected for defects. We then choose a "starting point" in the table of random numbers, in some "random" manner. For example, we can close our eyes and point a finger on a page of the random number table, and this can be the starting point. Alternatively, the numbers thus

Choosing a random sample.

chosen can be thought of as the page, column, and row number of a new starting point. Using this random procedure, having observed the numbers 3674826, we would proceed to page 367, column 48, and row 26 in a book such as *A Million Random Digits* [2]. This would be the starting point for the random section. If the numbers designating the starting point do not correspond to an available page, row, or column, the next numbers in sequence (going down or across the page as is convenient) can be used, and so on.

Table IV.1 is a typical page from a table of random numbers. The exact use of the table will depend on the specific situation. Some examples should clarify applications of the random number table to randomization procedures.

1. A sample of 10 bottles is to be selected from a universe of 800 bottles. The bottles are numbered from 1 to 800 inclusive. A starting point is selected from the random number table and three-digit numbers are used to accommodate the 800 bottles. Suppose that the starting point is row 6 and column 21 in Table IV.1. (The first three-digit number is 177.) If a number greater than 800 appears or a number is chosen a second time (i.e., the same number appears twice or more in the table), skip the number and proceed to the next one. The first 10 numbers found in Table IV.1 with the starting point above and subject to the foregoing restraints are (reading down) 177, 703, 44, 127, 528, 43, 135, 104, 342, and 604 (Table 4.1). Note that we did not include 964 because there is no bottle with this number; only 800 bottles are available. These numbers correspond to the 10 bottles that will be chosen for inspection.
2. Random numbers may be used to assign patients randomly to treatments in clinical trials. Initially, the characteristics and source of the patients to be included in the trial should be carefully considered. If a drug for the treatment of asthma were to be compared to a placebo

Table 4.1 Excerpt from Table IV.1

	Column 21	
Row 6	17	7
	70	3
	04	4
	12	7
	52	8
	04	3
	13	5
	96	4
	10	4
	34	2
	60	4

treatment, the source (or population) of the samples to be chosen could be all asthmatics in this country. Clearly, even if we could identify all such persons, for obvious practical reasons it would not be possible to choose those to be included in the study using the simple random sampling procedure described previously.

In fact, in clinical studies of this kind, patients are usually recruited by an investigator (physician), and *all* patients who meet the protocol requirements and are willing to participate are included. Most of the time, patients in the study are randomly assigned to the two or more treatments by means of a table of random numbers or a similar "random" device. Consider a study with 20 patients designed to compare an active drug substance to an identically appearing placebo. As patients enter the study, they are assigned randomly to one of the treatment groups, 10 patients to be assigned to each group. One way to accomplish this is to "flip" a coin, assigning, for example, heads to the active drug product and tails to the placebo. After 10 patients have been assigned to one group, the remaining patients are assigned to the incomplete group.

A problem with a simple random assignment of this kind is that an undesirable allocation may result by chance. For example, although improbable, the first 10 patients could be assigned to the active treatment and the last 10 to the placebo, an assignment that the randomization procedure is intended to avoid. (Note that if the treatment outcome is associated with a time trend due to seasonal effects, physician learning, personnel changes, etc., such an assignment would bias the results.) In order to avoid this possibility, the randomization can be applied to subgroups of the sample, sometimes called a block randomization. For 20 patients, one possibility is to randomize in groups of 4, 2 actives and 2 placebos to be assigned to each group of 4. This procedure also ensures that if the study should be aborted at any time, approximately equal numbers of placebo and active treated patients will be included in the results. Another application of blocking is to adjust the randomization for baseline variables, such as sex, duration of disease, and so on.

If the randomization is performed in groups of 4 as recommended, the following patient allocation would result. (Use Table 4.1 for the random numbers as before, odd for placebo, even for active.)

Patient	Random no.	Drug	Comment
1	1	P	
2	7	P	
3	—	D	
4	—	D	Assign D to patient 3 and 4 to ensure equal allocation of D and P in the subgroup
5	0	D	
6	1	P	
7	5	P	
8	—	D	Assign D to patient 8 to ensure equal allocation of D and P in the subgroup
9	0	D	
10	1	P	
11	9	P	
12	—	D	Assign D to patient 12 to ensure equal allocation of D and P in the subgroup
13	1	P	
14	3	P	
15	—	D	
16	—	D	Assign D to patients 15 and 16 to ensure equal allocation of D and P in the subgroup
17	6	D	
18	7	P	
19	0	D	
20	—	P	Assign D to patient 20 to ensure equal allocation of D and P in the subgroup

Table 4.2 Excerpt from Table IV.1: Assignment of First 10 Numbers Between 1 and 20 to Placebo

		Column 11	
Row 11	_ _ _ _ _	44	22 78 84 26 04 33 46 09 52
	59 29 97 68 60	71	91 38 67 54 13 58 18 24 76
	48 55 90 65 72	96	57 69 36 10 96 46 92 42 45
	66 37 32 20 30	77	84 57 03 29 10 45 65 04 26
	68 49 69 10 82	53	75 91 93 30 34 25 20 57 27
	83 62 64 11 12	67	19 _ _ _ _ _ _ _ _

The source and methods of randomization schemes for experiments or clinical studies should be documented for U.S. Food and Drug Administration submissions or for legal purposes. Therefore, it is a good idea to use a table of random numbers or a computer-generated randomization scheme for documentation rather than the coin-flipping technique. One should recognize, however that the latter procedure is perfectly fair, the choice of treatment being due to chance alone. Using a table of random numbers, a patient may be assigned to one treatment if an odd number appears and to the other treatment if an even number appears. We use single numbers for this allocation. If even numbers are assigned to drug treatment, the numbers in Table 4.1 would result in the following assignment to drug and placebo (read numbers down each column, one number at a time; the first number is 1, the second number is 7, the third number is 0, etc.).

Patient		Patient		Patient		Patient	
1	1 P	6	0 D	11	6 D	16	2 D
2	7 P	7	1 P	12	7 P	17	4 D
3	0 D	8	9 P	13	0 D	18	3 P
4	1 P	9	1 P	14	4 D		
5	5 P	10	3 P	15	2 D		

Since 10 patients have been assigned to placebo (P), the remaining two patients are assigned to drug (D). Again, the randomization can be performed in subgroups as described in the previous paragraph. If the randomization is performed in subgroups of size 4, for example, the first 4 patients would be assigned as follows: patients 1 and 2 to placebo (random numbers 1 and 7), and patients 3 and 4 to drug to attain equal allocation of treatments in this sample of 4.

Another approach is to number the patients from 1 to 20 inclusive as they enter the study. The patients corresponding to the first 10 numbers from the random number table are assigned to one of the two treatment groups. The remaining patients are assigned to the second treatment. In our example, the first 10 numbers will be assigned to placebo and the remaining numbers to drug. In this case, two-digit numbers are used from the random number table. (The numbers 1–20 have at most two digits.) Starting at row 11, column 11 in Table IV.1 and reading across, the numbers in Table 4.2 represent patients to be assigned to the first treatment group, placebo. Reading across, the first 10 numbers to appear that are between 1 and 20 (disregarding repeats), underlined in Table 4.2, are 4, 9, 13, 18, 10, 20, 3, 11, 12, and 19. These patients are assigned to placebo. The remaining patients, 1, 2, 5, 6, 7, 8, 14, 15, 16, and 17, are assigned to drug.

Randomization in clinical trials is discussed further in section 11.2.6.

4.3 OTHER SAMPLING PROCEDURES: STRATIFIED, SYSTEMATIC, AND CLUSTER SAMPLING

4.3.1 Stratified Sampling

Stratified sampling is a procedure in which the population is divided into subsets or strata, and random samples are selected from each strata. Stratified sampling is a recommended way of sampling when the strata are very different from each other, but objects within each stratum are alike. The precision of the estimated population mean from this sampling procedure is based on the variability within the strata. Stratified sampling will be particularly advantageous when this within-object variability is small compared to the variability between objects in different strata. In quality control procedures, items are frequently selected for inspection at random within specified time intervals (strata) rather than in a completely random fashion (simple random sampling). Thus we might sample 10 tablets during each hour of a tablet run. Often, the sample size chosen from each stratum is proportional to the size of the stratum, but in some circumstances, disproportionate sampling may be optimal. The computation of the mean and variance based on stratified sampling can be complicated, and the analysis of the data should take stratification into account [1]. In the example of the clinical study on asthmatics (see sect. 4.2.1), the stratification could be accomplished by dividing the asthmatic patients into subsets (strata) depending on age, duration of illness, or severity of illness, for example. The patients are assigned to treatments randomly within each subset. (See randomization in blocks above.) Note in this example that patients within each stratum are more alike than patients from different strata.

Consider an example of sampling tablets for drug content (assay) during a tablet run. If we believe that samples taken close in time are more alike than those taken at widely differing times, stratification would be desirable. If the tableting run takes 10 hours to complete, and a sample of 100 tablets is desired, we could take 10 tablets randomly during each hour, a stratified sample. This procedure would result in a more precise estimate of the average tablet potency than a sample of 100 tablets taken randomly over the entire 10-hour run.

Although stratified sampling often results in better precision of the estimate of the population mean, in some instances the details of its implementation may be more difficult than those of simple random sampling.

4.3.2 Systematic Sampling

Systematic sampling is often used in quality control. In this kind of sampling, every *n*th item is selected (e.g., every 100th item). The initial sample is selected in a random manner. Thus, a quality control procedure may specify that 10 samples be taken at a particular time each hour during a production run. The time during the hour for each sampling may be chosen in a random manner. Systematic sampling is usually much more convenient, and much easier to accomplish than simple random sampling and stratified sampling. It also results in a uniform sampling over the production run, which may result in a more precise estimate of the mean. Care should be taken that the process does not show a cyclic or periodic behavior, because systematic sampling will then not be representative of the process. The correct variance for the mean of a systematic sample is less than that of a simple random sample if the variability of the systematic sample is greater than the variability of the entire set of data.

To illustrate the properties of a systematic sample, consider a tableting process in which tablet weights tend to decrease during the run, perhaps due to a gradual decrease in tableting pressure. The press operator adjusts the tablet pressure every hour to maintain the desired weight. The tablet weights during the run are illustrated in Figure 4.1. If tablets are sampled 45 minutes after each hour, the average result will be approximately 385 mg, a biased result.

If the data appear in a random manner, systematic sampling may be desirable because it is simple and convenient to implement. As noted above, "systematic sampling is more precise than random sampling if the variance within the systematic sample is larger than the population variance as a whole." Another way of saying this is that systematic sampling is precise when units within the same sample are heterogeneous, and imprecise when they are homogeneous [3]. In the tableting example noted in the previous paragraph, the units in the sample tend to be similar (precise) and systematic sampling is a poor choice. (See Exercise Problem 11 for an example of construction of a systematic sample.)

4.3.3 Cluster Sampling

In cluster sampling, the population is divided into groups or clusters each of which contain "subunits." In single-stage cluster sampling, clusters are selected at random and all elements of the clusters chosen are included in the sample (4).

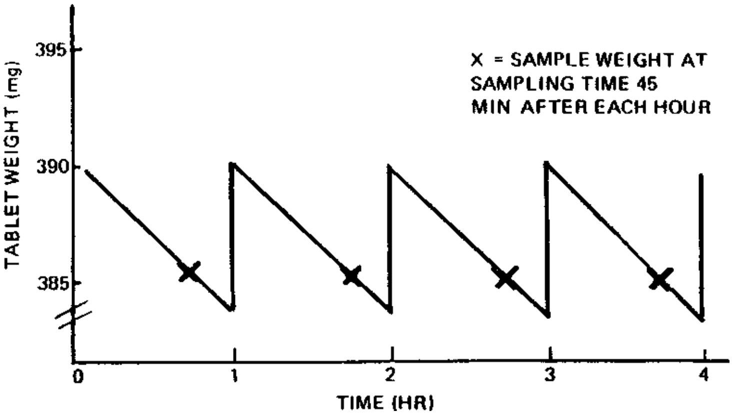


Figure 4.1 Illustration of problem with systematic sampling when process shows periodic behavior.

Two-stage cluster sampling may be used when there are many "primary" units, each of which can be "subsampled." For example, suppose that we wish to inspect tablets visually, packaged in the final labeled container. The batch consists of 10,000 bottles of 100 tablets each. The primary units are the bottles and the subsample units are the tablets within each bottle. Cluster sampling, in this example, might consist of randomly selecting a sample of 100 bottles, and then inspecting a random sample of 10 tablets from each of these bottles, thus the nomenclature, "two-stage" sampling. Often, cluster sampling is the most convenient way of choosing a sample. In the example above, it would be impractical to select 1000 tablets at random from the 1,000,000 packaged tablets (10,000 bottles × 100 tablets per bottle).

For a continuous variable such as tablet weights or potency, the estimate of the variance of the mean in two-stage cluster sampling is

$$(1 - f_1)S_1^2/n + [S_2^2/(\mathrm{nm})](f_1(1 - f_2)) \tag{4.1}$$

where S_1^2 is the estimate of the variance among the primary unit means (the means of bottles). S_2^2 is the estimate of the variance of the subsample units, that is, units within the primary units (between tablets within bottles). f_1 and f_2 are the sampling fractions of the primary and subsample units, respectively. These are the ratios of units sampled to the total units available. In the present example of bottled tablets,

f_1 = 100 bottles/10,000 bottles = 0.01 (100 bottles are randomly selected from 10,000)
f_2 = 10 tablets/100 tablets = 0.1 (10 tablets are randomly selected from 100 for each of the 100 bottles)
n = number of primary unit samples (100 in this example)
m = number of units sampled from each primary unit (10 in this example).

If, in this example, S_1^2 and S_2^2 are 2 and 20, respectively, from Eq. (4.1), the estimated variance of the mean of 1000 tablets sampled from 100 bottles (10 tablets per bottle) is

$$\frac{(1 - 0.01)(2)}{100} + \left[\frac{20}{(100 \times 10)}\right](0.01)(0.9) = 0.01998.$$

If 1000 tablets are sampled by taking 2 tablets from each of 500 bottles, the estimated variance of the mean is

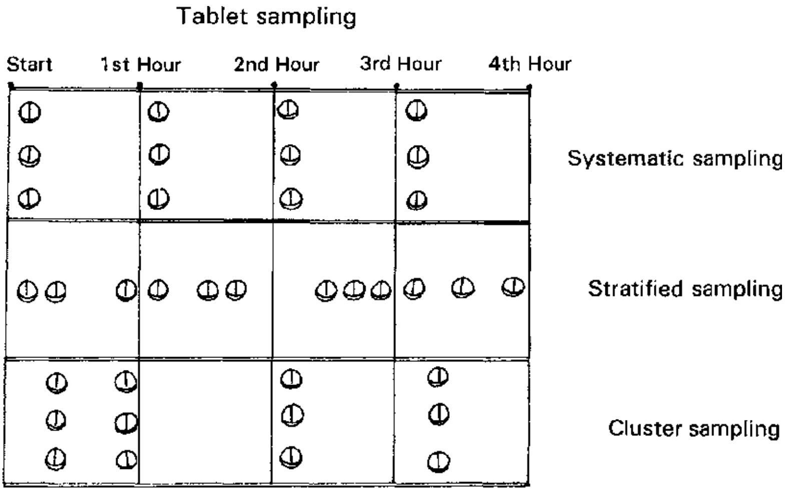
$$\frac{(1 - 0.05)(2)}{500} + \left[\frac{20}{(500 \times 2)}\right](0.05)(0.98) = 0.00478.$$

This example illustrates the increase in efficiency of sampling more primary units. The variance obtained by sampling 200 bottles is approximately one-half that of sampling 100 bottles.

If f_1 is small, the variance of the mean is related to the number of primary units sampled (n) equal to approximately S_1^2/n. Cost and time factors being equal, it is more efficient to sample more primary units and fewer subsample units given a fixed sample size. However, in many situations it is not practical or economical to sample a large number of primary units. The inspection of tablets in finished bottles is an example where inspection of many primary units (bottles) would be costly and inconvenient. See Exercise Problems 9 and 10 for further illustrations.

4.4 SAMPLING IN QUALITY CONTROL

Sampling of items for inspection, chemical, or physical analysis is a very important aspect of quality control procedures. For the moment, we will not discuss the important question: "What sample size should we take?" This will be discussed in chapter 6. What concerns us here is how to choose the samples. In this respect, the important points to keep in mind from a statistical point of view are as follows:

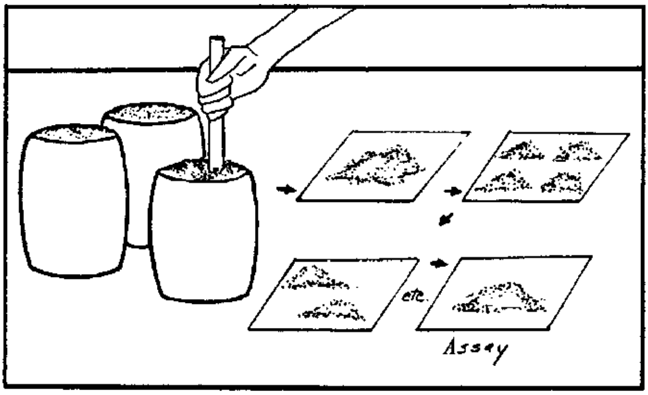


1. The sample should be "representative."
2. The sample should be chosen in a way that will be compatible with the objectives of the eventual data analysis.

For example, when sampling tablets, we may be interested in estimating the mean and standard deviation of the weight or potency of the tablet batch. If 20 tablets are chosen for a weight check during each hour for 10 hours from a tablet press (a stratified or systematic sample), the mean and standard deviation are computed in the usual manner if the production run is uniform resulting in random data. However, if warranted, the analysis should take into account the number of tablets produced each hour and the uniformity of production during the sampling scheme. For example, in a uniform process, an estimate of the average weight would be the average of the 200 tablets, or equivalently, the average of the averages of the 10 sets of 20 tablets sampled. However, if the rate of tablet production is doubled during the 9th and 10 hours, the averages obtained during these two hours should contribute twice the weight to the overall average as the average results obtained during the first eight hours. For further details of the statistical analysis of various sampling procedures, the reader is referred to Refs. [1,3].

Choosing a representative sample from a bulk powder, as an example, is often based on judgment and experience more than on scientific criteria (a "judgment" sample). Rules for sampling from containers and for preparing powdered material or granulations for assay are, strictly speaking, not "statistical" in nature. Bulk powder sampling schemes have been devised in an attempt to obtain a representative sample without having to sample an inordinately large amount of material. A common rule of thumb, taking samples from $\sqrt{N} + 1$ containers (N is the total number of containers), is a way to be reasonably sure that the material inspected is representative of the entire lot, based on tradition rather than on objective grounds. Using this rule, given a batch of 50 containers, we would sample $(\sqrt{50}+1=8)$ containers. The eight containers can be chosen using a random number table (see Exercise Problem 3).

Sampling plans for bulk powders and solid mixes such as granulations usually include the manner of sampling, the number of samples, and preparation for assay with an aim of obtaining a representative sample. One should bear in mind that a single assay will not yield information on variability. No matter what precautions we take to ensure that a single sample of a mix is representative of a batch, we can only estimate the degree of homogeneity by repeating the procedure one or more times on different portions of the mix. Repeat assays on the same sample gives an estimate of analytical error, not homogeneity of the mix. For a further discussion of this concept see chapter 13.



Sampling and assaying bulk powders.

An example of a procedure for sampling from large drums of a solid mixture is to insert a thief (a device for sampling bulk powders) and obtain a sample from the center of the container. A grain thief may be used to take samples from more than one part of the container. (If samples are to be taken for purposes of content uniformity, thieves that can sample small samples such as one or more tablets weights are recommended.) This procedure is repeated for an appropriate number of containers and the samples thoroughly mixed. The sample to be submitted for analysis is mixed further and quartered, rejecting two diagonal portions. The mixing and quartering is repeated until sufficient sample for analysis remains.

Other ideas on sampling for quality control and validation can be found in section 13.1.1.

KEY TERMS

Blocking
Cluster sample
Haphazard sample
Judgment sample
Multistage sample
Nonprobability sample
Probability sample
Representative sample
Sample
Sampling with replacement
Simple random sample
Stratified sample
Systematic sample
Table of random numbers
Two-stage cluster sample

EXERCISES

Use the table of random numbers (Tables IV.1) to answer the following questions.

1. Twenty-four patients are recruited for a clinical study, 12 patients to be randomly assigned to each of two groups, A and B. The patients come to the clinic and are entered into the

study chronologically, randomly assigned to treatment A or B. Devise a schedule showing to which treatment each of the 24 patients is assigned.

2. Devise a randomization scheme similar to that done in Problem 1 if 24 patients are to be assigned to three treatments.
3. Thirty drums of bulk material are to be sampled for analysis. How many drums would you sample? If the drums are numbered 1 to 30, explain how you chose drums and take the samples.
4. A batch of tablets is to be packaged in 5000 bottles each containing 1000 tablets. It takes four hours to complete the packaging operation. Ten bottles are to be chosen for quality control tests. Explain in detail how would you choose the 10 bottles.
5. Devise a randomization scheme to assign 20 patients to drug and placebo groups (10 patients in each group) using the numbers shown in Table 4.1 by using even numbers for assignment to drug and odd numbers for assignment to placebo.
6. Describe two different ways in which 20 tablets can be chosen during each hour of a tablet run.
7. One hundred bottles of a product, labeled 0 to 99 inclusive, are available to be analyzed. Analyze five bottles selected at random. Which five bottles would you choose to analyze?
8. A batch of tablets is produced over an eight-hour period. Each hour is divided into four 15-minute intervals for purposes of sampling. (Sampling can be done during 32 intervals, four per hour for eight hours.) Eight samples are to be taken during the run. Devise (a) a simple random sampling scheme, (b) a stratified sampling scheme, and (c) a systematic sampling scheme. Which sample would you expect to have the smallest variance? Explain.
9. The average potencies of tablets in 20 bottles labeled 1 to 20 are

Bottle number	Potency
1	312
2	311
3	309
4	309
5	310
6	308
7	307
8	305
9	306
10	307
11	305
12	301
13	303
14	300
15	299
16	300
17	300
18	297
19	296
20	294

(a) Choose a random sample of five bottles. Calculate the mean and standard deviation.
(b) Choose a systematic sample, choosing every 4th sample, starting randomly with one of the first four bottles. Calculate the mean and standard deviation of the sample.
(c) Compare the averages and standard deviations of the two samples and explain your results. Compare your results to those obtained by other class members.

10. Ten containers each contain four tablets. To estimate the mean potency, two tablets are to be randomly selected from three randomly chosen containers. Perform this sampling from the data shown below. Estimate the mean and variance of the mean. Repeat the sampling, taking three tablets from two containers. Explain your results. Compute the mean potency of all 40 tablets.

Container	Tablet potencies (mg)
1	290 289 305 313
2	317 300 285 327
3	288 322 306 299
4	281 305 309 289
5	292 295 327 283
6	286 327 297 314
7	311 286 281 288
8	306 282 282 285
9	313 301315 285
10	283 327 315 322

11. Twenty-four containers of a product are produced during eight minutes, three containers each minute. The drug content of each container is shown below

Minute	Container assay		
1	80	81	77
2	78	76	76
3	84	83	86
4	77	77	79
5	83	81	82
6	81	79	80
7	82	79	81
8	79	79	80

Eight containers are to be sampled and analyzed for quality control. Take a sample of eight as follows:

(a) Simple random sample.
(b) Stratified sample; take one sample at random each minute.
(c) Systematic sample; start with the first, second, or third container and then take every third sample thereafter.
Compute the mean and the variance of each of your three samples (a, b, and c). Discuss the results. Which sample gave the best estimate of the mean? Compare your results to those obtained from the other students in the class.

12. What is the probability that a batter with a 0.250 average will get exactly one hit in four times at bat? Answer: Probability of one hit in 4 times at bat is $4 \times (1/4) \times (3/4)^3 = 108/256 = 0.421875$, less than one-half of the time.

REFERENCES

1. Snedecor GW, Cochran WG. Statistical Methods, 8th ed. Ames, IA: Iowa State University Press, 1989.
2. The Rand Corporation. A Million Random Digits with 100,000 Normal Deviates. New York: The Free Press, 1966.
3. Cochran WG. Sampling Techniques, 3rd ed. New York: Wiley, 1967.
4. Stuart A. Basic Ideas of Scientific Sampling. London: Charles Griffith & Co., Ltd., 1976.