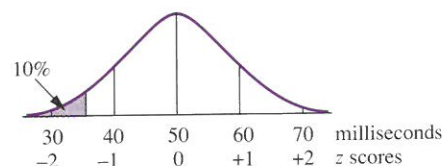
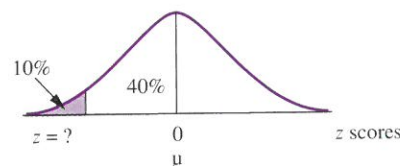


- b. Shade the area in question. Since the fastest times would be *less than* 50 ms, we shade the extreme left of the normal curve, estimating 10%.



- c. Next obtain the z score at the cutoff. To do this, we look up 40% (.4000) since the table reads data only from the center ($z = 0$) out.

Note in the table that the closest value to 40% (.4000) is .3997, which gives us a z score of $z = 1.28$. Since the z value is below μ , we must make the z value negative. Thus, $z = -1.28$.



Normal Curve Table				
z	.00	.0108
.0				
.				
.				
.				
1.2				.3997

- d. Now use the z formula to solve for x , the real data value at the cutoff. Essentially we know z (-1.28), and we wish to solve for x in the formula:

$$z = \frac{x - \mu}{\sigma} \quad \begin{cases} \mu = 50 \\ \sigma = 10 \end{cases}$$

$$-1.28 = \frac{x - 50}{10}$$

$$(10)(-1.28) = x - 50$$

$$-12.8 = x - 50$$

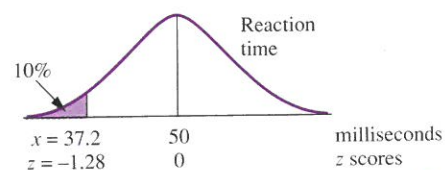
$$37.2 = x$$

$$\text{or } x = 37.2 \text{ ms}$$

This calculation required some algebraic manipulation. First, we multiplied both sides of the equation by 10 to obtain $-12.8 = x - 50$. Second, we added +50 to both sides of the equation to get $37.2 = x$. In other words, at the cutoff, $x = 37.2$ ms.

Answer

Below 37.2 ms you would expect to find the fastest 10% of the reaction times.



Practice 3 For the preceding problem, between what two values would you expect to find the *middle* 95% of the reaction times?

Answer Between 30.4 and 69.6 ms

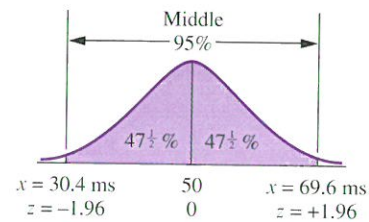
Note: We must look up $47\frac{1}{2}\%$ (half of 95%), or in decimal form .4750, to obtain $z = 1.96$ on both sides. When we substitute $z = -1.96$ and $z = +1.96$ in our z formula, we obtain the following:

$$z = \frac{x - \mu}{\sigma} \qquad z = \frac{x - \mu}{\sigma}$$

$$-1.96 = \frac{x - 50}{10} \qquad +1.96 = \frac{x - 50}{10}$$

$$x = 30.4 \text{ ms} \qquad x = 69.6 \text{ ms}$$

(To solve, multiply both sides by 10, then add 50 to both sides.)



Practice 4 For the above problem, *above* what value would you expect to find the *slowest* 70% of the reaction times?

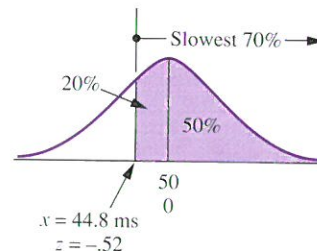
Answer Above 44.8 ms

Note: 50% of the data is above $\mu = 50$ ms so we must look up the remaining 20% (.2000). The closest value to .2000 is .1985, which is equivalent to $z = -.52$. Substituting $z = -.52$ in our z formula, we obtain the following:

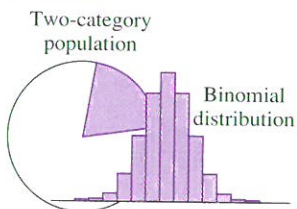
$$z = \frac{x - \mu}{\sigma}$$

$$-.52 = \frac{x - 50}{10}$$

$$x = 44.8 \text{ ms}$$



4.4 Binomial Distribution: An Introduction to Sampling



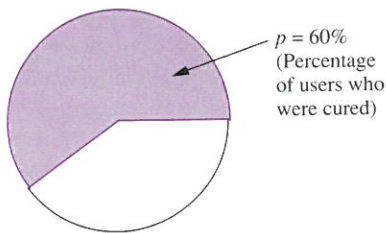
Although some natural populations have distributions that can be approximated with the normal curve, the normal curve's importance is derived more from its consistent and uncanny ability to predict the outcomes when we *sample* from a population. Although different "types" of populations exist (from which we may sample), one of the most important in research is the two-category population.

A **two-category population** is a population where every member is classified into exactly one of two categories.

Examples of two-category populations are as follows.

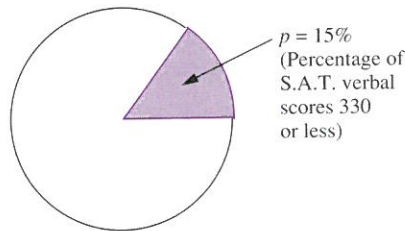
Medical Population

Many thousands of users of a new experimental drug designed to cure a specific form of bladder inflammation, classified into *users who were cured* and *users not cured*.



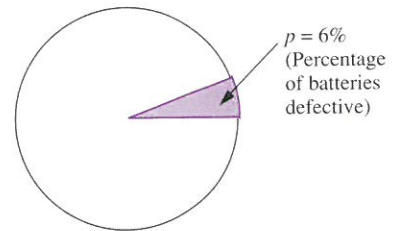
Educational Population

Hundreds of thousands of SAT verbal scores recorded over the past five years, classified into *scores 330 or less* and *scores above 330*.



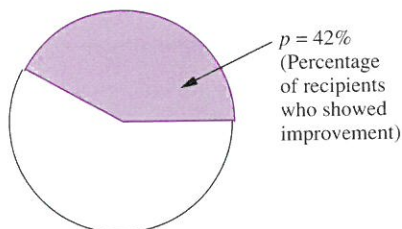
Manufacturing Population

Millions of assembly-line batteries produced last month by a large manufacturer, classified into *batteries defective* and *batteries not defective*.



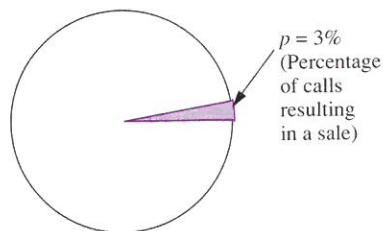
Psychological Population

Thousands of recipients of a new drug-free therapy, classified into *those who showed improvement* and *those with no improvement*.



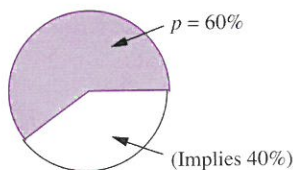
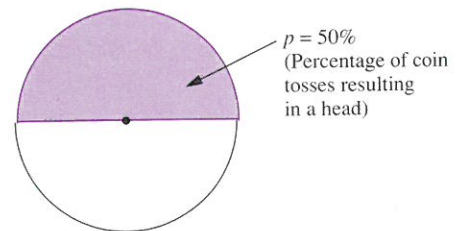
Marketing Population

Hundreds of thousands of phone calls made to New Jersey residents last year by the Fullins Co. selling magazine subscriptions, classified into *calls resulting in a sale* and *calls resulting in no sale*.



Gambling Population

Billions of coin flips, classified into *those resulting in heads* and *those resulting in tails*.



Notice that we may describe such two-category populations by the letter p , the proportion or percentage classified into *one* of the categories. Of course, once we know the percentage of the population in one category, we know the percentage in the other, since the sum of the two percentages must add to 100%. For instance, in the medical population, the first example above, if 60% of the users of this experimental drug were cured, indicated by the shaded region, this implies 40% were not cured. This 40% is represented by the *unshaded* region (note: $60\% + 40\% = 100\%$).

In case you were wondering, it doesn't matter which of the two categories in a two-category population we describe by p . For instance, in the manufacturing population, we described this population of assembly-line batteries as $p = 6\%$ defective, however a salesman for this company might describe this exact same population as $p = 94\%$ okay. Most often, we assign p to the particular category we are interested in.

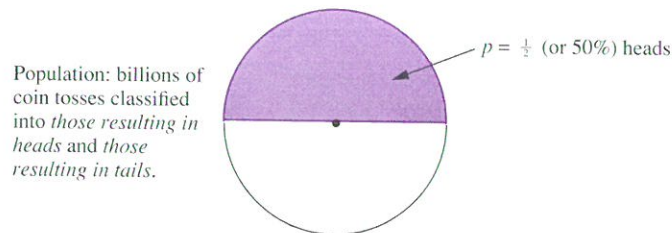
One last important point: every member of a two-category population must fall into one or the other category. In other words, each battery in the manufacturing population must be classified as either defective or not defective. Each user of the experimental drug must be classified as either cured or not cured. Each telephone call in the marketing population must be classified as sale or no sale. There can be no borderline cases. Each member of the population uniquely fits into one or the other category.

Sampling from a Two-Category Population

Once we determine we have a two-category population and describe this population by p (the percentage of values in one category), then we may wish to know what we can expect when we sample from such a population.

Since the methodology for determining such sampling evolved from early gambling experiments, usually involving the tossing of coins or dice, we offer the following.*

Suppose we have the following two-category population:



Now, if we were to randomly sample from this population, say for instance, we sample 12 coin flips, what may we expect to happen? We know from theory and a long history of experience, that if we were to *randomly* sample from any large two-category population,

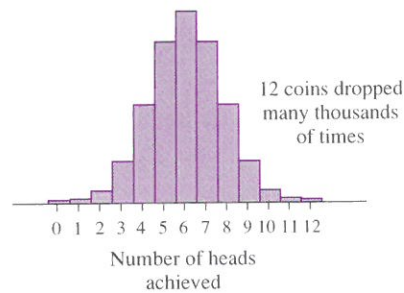
$$p_s \approx p$$

The sample proportion, p_s , will be approximately equal to the population proportion, p .

*This topic was introduced at the end of chapter 3, section 3.5.

That is, since the population consists of 50% heads, then a random sample should consist of *approximately* 50% heads. In the case of $n = 12$ coin flips, we should get *approximately* 6 heads (50% of $12 = 6$). However, we can also get 5 heads or perhaps even 9 heads. How can we determine the percentage of times we can expect each of these outcomes to occur?

One way is to actually perform this experiment a great many times, as follows.



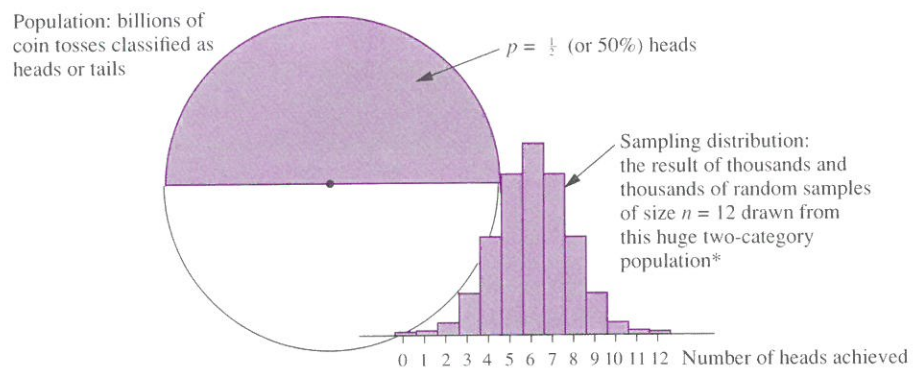
Say we drop $n = 12$ coins on a table thousands and thousands of times and record the number of heads achieved on each drop, we would get something like the histogram shown here. (See footnote for further discussion.)

This is called a sampling distribution, defined as follows:

A **sampling distribution** shows us what we can expect when we randomly select n values (a fixed number) repeatedly from a particular population.

In fact, the above sampling distribution shows us what we can expect when we randomly select $n = 12$ values repeatedly from a large two-category population described by $p = \frac{1}{2}$ (or 50%) heads.

These results can be summarized with the following diagrams:



*A sampling distribution technically is based on the concept of selecting all possible different samples of a fixed size from a given population. However, even small populations produce enormous numbers of different possible samples. Usually after randomly selecting several hundred samples, the characteristics of a sampling distribution become quite clear. Rudimentary sampling distributions in this text can be generated using Microsoft Excel (Tools, Data Analysis, Random Number Generator, Histogram; for the given histogram, fifteen thousand samples were randomly selected with the following random number input: 1, 15000, binomial distribution, $p = .5$, trials = 12). The obtained values of μ and σ , the mean and standard deviation of the sampling distribution, matched calculated values (formulas to be introduced on page 123) to approximately two decimal places. The technical concept of a sampling distribution is discussed at length in chapter 5 endnote 2.

Notice that this particular sampling distribution (the histogram) is symmetrical around the value we would most likely expect to occur. In the case of dropping $n = 12$ coins, we would most likely expect 50% heads or 6 heads. And indeed, in this instance, we do most often get 6 heads. This is called the **expected value** and can be calculated as follows.*

$$\text{Expected value} = np$$

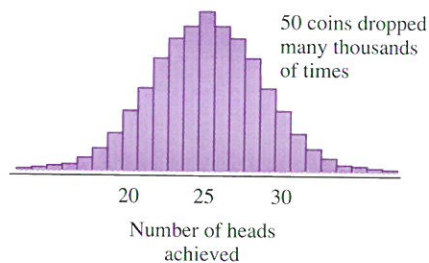
Since our sample size is $n = 12$, and the population proportion is $p = \frac{1}{2}$ heads,

$$\begin{aligned}\text{Expected value} &= np \\ &= (12)(1/2 \text{ heads}) \\ &= 6 \text{ heads}\end{aligned}$$

Although we indeed most often get 6 heads, on a great many occasions we get somewhat more than 6 heads, and on a great many occasions somewhat less, with the heights of the histogram bars falling off in a shape strongly resembling that of a normal distribution.

Actually, this should not come as a surprise, since the initial discovery of the normal curve evolved from these same early coin experiments; recall De Moivre's and Laplace's work discussed at the beginning of this chapter.

In fact, these bell-shaped sampling distributions appear repeatedly in gambling experiments when n is large. For instance, the following:



Suppose we drop $n = 50$ coins on a table thousands and thousands of times and record the number of heads achieved on each drop, we would get a distribution something like this.

*Expected value was defined in section 3.6 using the general formula, expected value $= \sum xp(x)$. It can be shown for binomial experiments such as these, after algebraic manipulation, expected value $= np$.

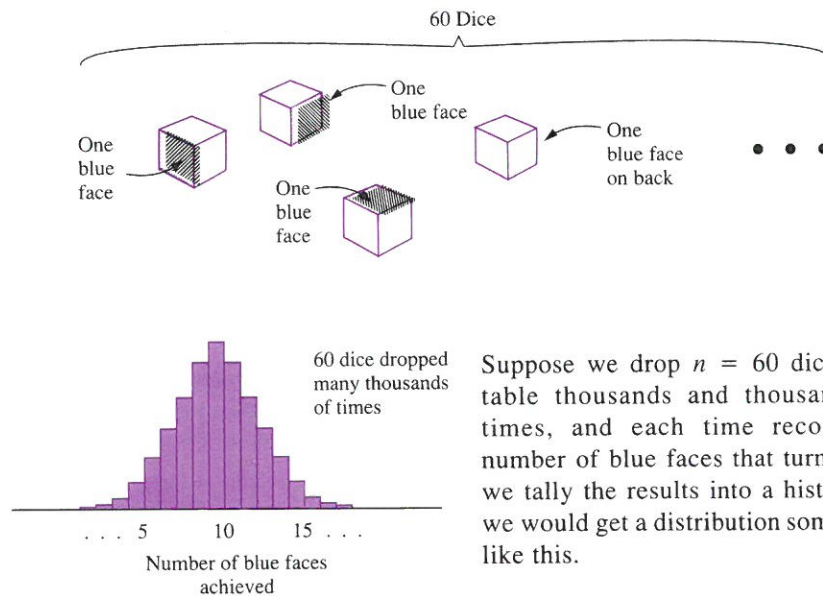
Notice the sampling distribution is again symmetrical around the value we would most likely expect, which is 25 heads (since the population consists of $\frac{1}{2}$ heads, then any random sample should consist of *approximately* $\frac{1}{2}$ heads; $\frac{1}{2}$ of $50 = 25$). Again, this may be calculated as follows:

$$\begin{aligned}\checkmark \text{ Expected value} &= np \\ &= (50)(1/2 \text{ heads}) \\ &= 25 \text{ heads}\end{aligned}$$

And indeed we do most often get 25 heads (see histogram above); however, on a great many occasions we get somewhat more than 25 heads and on a great many occasions somewhat less, with the heights of the histogram bars again falling off in a shape resembling that of a normal distribution.

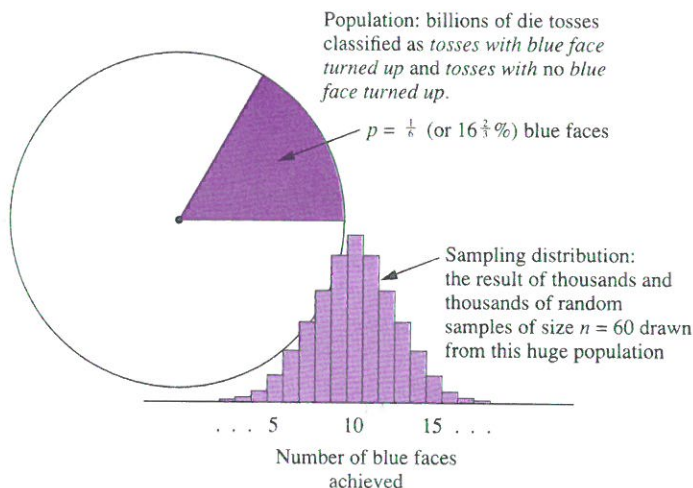
Okay, you might ask, this may happen with coin tosses, where the probability of a head for a coin toss is $\frac{1}{2}$ (50%), but what if we sampled a different population, say die tosses, where the probability of a particular face turning up is $\frac{1}{6}$ ($16\frac{2}{3}\%$). What happens then?

Well, let's take 60 dice and paint one face on each blue (for identification purposes).



Suppose we drop $n = 60$ dice on a table thousands and thousands of times, and each time record the number of blue faces that turn up. If we tally the results into a histogram, we would get a distribution something like this.

For a clearer picture of this, let's summarize the results with the following diagram.



Notice the shape of the sampling distribution (the histogram). It is symmetrical around the value we would most likely expect, in this case 10 blue faces. In other words, since the population consists of $\frac{1}{6}$ blue faces, any random sample should consist of *approximately* $\frac{1}{6}$ blue faces (note: $\frac{1}{6}$ of 60 = 10). Again, this expected value can be calculated as follows.

$$\begin{aligned}\text{Expected value} &= np \\ &= (60)(1/6 \text{ blue faces}) \\ &= 10 \text{ blue faces}\end{aligned}$$

And indeed 10 blue faces is our most frequently occurring value. However, on many occasions we get somewhat more than 10 blue faces and on many occasions somewhat less, again with the heights of the histogram bars falling off in a shape approximating that of a normal distribution.

Normal Curve Approximation to the Binomial Sampling Distribution

These bell-shaped sampling distributions kept occurring with amazing regularity in coin and dice experiments when n , the number of coins or dice dropped *was sufficiently large*. Of course, at this point you might ask, how large must n be to

be considered “sufficiently large”? Large enough, so when multiplied by p or $(1 - p)$, the result exceeds 5—which leads us to the following important rule.

In these types of sampling experiments, known as **binomial sampling experiments** (to be further defined in Section 4.5),

$$\text{if expected value } (np) > 5 \text{ and } n(1 - p) > 5$$

the sampling distribution (known as a **binomial sampling distribution**) will be approximately normally distributed with mean and standard deviation

$$\begin{aligned}\mu &= \text{expected value } (np) \\ \sigma &= \sqrt{np(1 - p)}\end{aligned}$$

and a normal curve with these dimensions can be fitted over the distribution and used to estimate probabilities.

Does this imply that if np or $n(1 - p)$ is 5 or less, the normal curve cannot be used to estimate probabilities? Yes, for np or $n(1 - p)$ of 5 or less, the sampling distribution is often skewed or sloping and generally the normal curve cannot be depended on to give reliable estimates. For these special cases, other techniques are available, which are discussed in chapter 11.

For the remainder of this chapter, we will demonstrate only those situations where the sampling distribution can be approximated with the normal curve, namely when

$$\text{Expected value } (np) > 5 \quad \text{and} \quad n(1 - p) > 5$$

Let's see how this works in an example.

Example

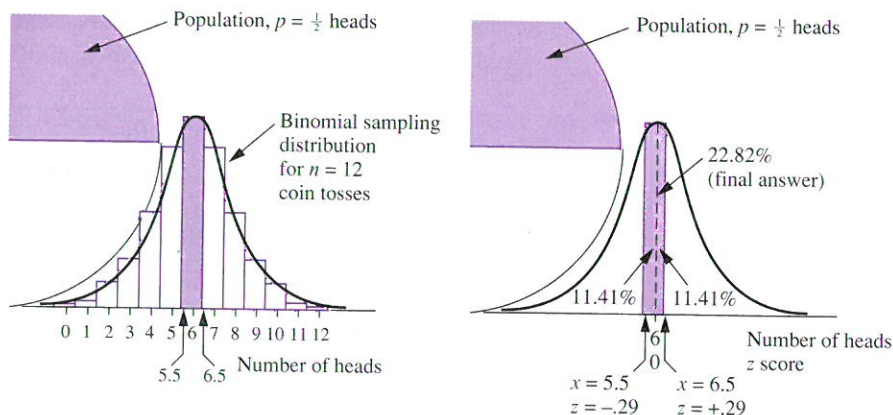
Out of 12 tosses of a coin, find the probability of achieving exactly 6 heads.

Solution

Since the expected value $(np) = (12)(\frac{1}{2}) = 6$, which is greater than 5, and $n(1 - p) = (12)(1 - \frac{1}{2}) = 6$, which is greater than 5, we now know repeated samples of $n = 12$ will produce a sampling distribution approximately normally distributed such that a normal curve with mean and standard deviation as follows can be used to estimate probabilities.

$$\begin{aligned}\mu &= \text{expected value} = np \\ &= 12(1/2) \\ &= 6 \text{ heads}\end{aligned}\qquad \begin{aligned}\sigma &= \sqrt{np(1 - p)} \\ &= \sqrt{12(1/2)(1/2)} \\ &= 1.73\end{aligned}$$

Now, to answer the question, what is the probability that out of 12 tosses we will achieve exactly 6 heads, we proceed as follows.



First, we shade the histogram bar representing exactly 6 heads. Note the shading must extend from 5.5 to 6.5 to include the entire width of the histogram bar representing 6 heads.

This $\frac{1}{2}$ -unit adjustment (referred to as a **continuity correction factor**) is necessary when the normal curve is used to estimate probabilities in the binomial sampling distribution. The term *continuity correction factor* is further defined at the end of the example. Now fit a normal curve over the histogram to estimate probabilities.

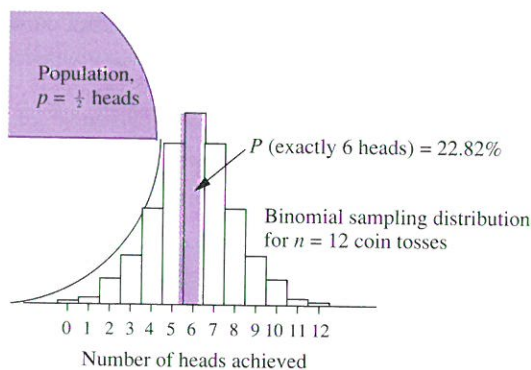
Resketch normal curve (for clarity) and shade area from 5.5 to 6.5. Using $\mu = 6$ and $\sigma = 1.73$, we solve as we would any normal curve problem by first calculating the z score at the cutoffs.

$$z = \frac{x - \mu}{\sigma} = \frac{5.5 - 6}{1.73} = \frac{-.5}{1.73} = -.29$$

The percentage of data from $z = 0$ to $z = -.29$ is 11.41%. Since there is an equal amount of data from $z = 0$ to $z = +.29$, we add 11.41% + 11.41% to get 22.82%.

Answer

Now we can say that the probability of achieving exactly 6 heads out of 12 tosses is 22.82%. Visually, this can be represented as follows:



Terminology

Continuity Correction Factor

This refers to the $\frac{1}{2}$ -unit shading adjustment(s) necessary to include the entire width of the histogram bar(s) in question.

This is necessary since it is the area occupied by the histogram bar that represents the probability that event will occur. So, remember, when using the normal curve to estimate binomial probabilities, we must shade the entire histogram bar(s) in question to get all the probability.

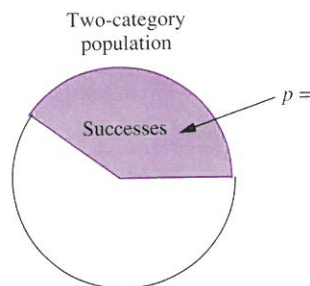
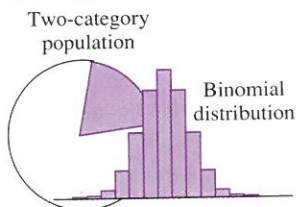
The *binomial sampling distribution* is sometimes referred to as a **discrete** data distribution, meaning the distribution contains only discrete values.

Discrete Values

Values that, when presented on a number line, occupy only distinct unconnected (or isolated) points.

Note in the histogram above, the data is classified only into values such as 0 heads, 1 head, 2 heads, 3 heads, etc. In other words, if 12 coins were dropped, you could never achieve $3\frac{1}{4}$ or $5\frac{1}{2}$ heads. When data can assume only isolated point values, such as in this case whole-number values, it is referred to as discrete.

4.5 Binomial Sampling Distribution: Applications



Of course, at this point, you may very well say, who cares about this; these are gambling experiments and I'm interested in business, psychology, medicine, education, or whatever.

Well, let's say, these binomial sampling distributions will form no matter what field of endeavor you apply them to, research in business, psychology, medicine, education, or whatever, provided you conform to the fundamental assumptions of binomial sampling, as follows.

Binomial sampling assumes selection from a two-category population (with members in *one* category labeled "success"), such that

1. there is a fixed number of selections, n (often referred to as " n trials"),
2. with each selection (or trial) independent* and each having the same probability, p , a success will be chosen.

***Independent** means: whether or not we achieve a success on one selection in no way affects the probability of achieving a success on any other selection.

Binomial sampling can be used in a wide variety of contemporary applications, provided we conform to these fundamental conditions. These conditions are necessary to conform to the basic fundamentals that are innate to coin, dice, and other gambling experiments, on which the theory and mathematics is based.

Although we must evaluate every contemporary experiment on the above formal terms, in actual practice these conditions can often be satisfied by simply

**Obtaining a Valid* Random Sample
from a Large Two-Category Population**

This will generally satisfy the above conditions for binomial sampling. A *large* population is defined as any population at least 20 times the size of the sample.

Let's see how all this applies to a contemporary experiment.

Example

From many thousands of users of a new experimental drug designed to cure a specific form of bladder inflammation, it was found that 60% were cured.

Suppose we randomly select $n = 15$ individuals from this large two-category population, what percentage of the time (or with what probability) would we find 12 or more cured?

Solution

Notice this is a binomial sampling experiment: there are 15 fixed selections from a two-category population, each independent and each having the same probability a cured individual (a success) will be chosen. *Random* selection from a

large two-category population generally satisfies these conditions. Furthermore, since expected value $(np) = (15)(.60) = 9$ and $n(1 - p) = (15)(.40) = 6$, and both are greater than 5, the resulting binomial sampling distribution will be approximately normally distributed with mean and standard deviation calculated as follows:

$$\begin{aligned}\mu &= np \\ &= 15(.60) \\ &= 9\end{aligned}\qquad\begin{aligned}\sigma &= \sqrt{np(1 - p)} \\ &= \sqrt{15(.60)(1 - .60)} \\ &= \sqrt{15(.60)(.40)} \\ &= 1.897 \\ &= 1.90\end{aligned}$$

Now a normal curve with these dimensions ($\mu = 9$, $\sigma = 1.90$) can be fitted over the histogram to estimate probabilities in any portion.

assumes both internal and external validity, as discussed in

Population: many thousands of
users of a new experimental drug

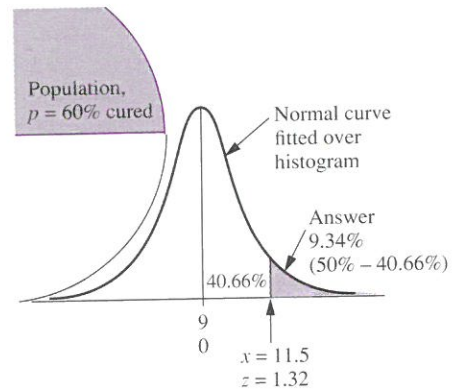
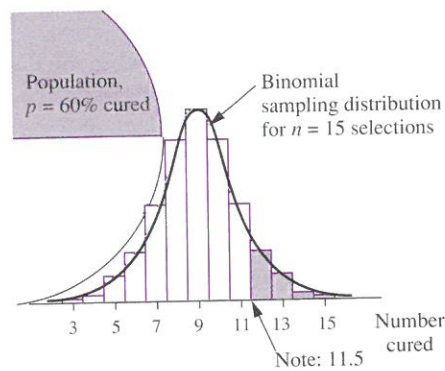
$p = 60\%$ cured

Binomial sampling distribution:
the result of
many thousands of
samples

$$Z = \frac{X - \mu}{\sigma}$$

$$\sigma = \sqrt{np(1-p)}$$

So, to answer our question, out of $n = 15$ randomly selected individuals what percentage of the time would we find 12 or more cured, we proceed as follows.



First, we shade the histogram bars representing 12 and above. Note the shading must extend to 11.5 to include the entire bar representing 12 cured. This half-unit adjustment is called your *continuity correction factor*. Now, we fit a normal curve over the histogram.

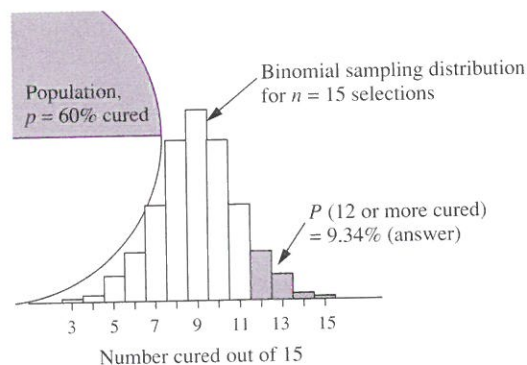
Second, we resketch the normal curve and shade the area 11.5 and above. Using $\mu = 9$ and $\sigma = 1.90$, we solve as we would solve any normal curve problem by first calculating the z score at the cutoff.

$$z = \frac{x - \mu}{\sigma} = \frac{11.5 - 9}{1.90} = \frac{2.5}{1.90} = 1.32$$

The percentage of data from $z = 0$ to $z = 1.32$ is 40.66%. Subtract this from 50% to get 9.34%.

Answer

Now we can say, 9.34% of the time we will achieve 12 or more cured when we randomly sample 15 from our population, and this can visually be represented as follows:



Let's summarize: since the population consists of 60% cured, any random sample will most likely consist of approximately 60% cured (60% of 15 = 9); thus, in this case, *approximately* 9 cured would be expected. Achieving 12 or more cured out of a sample of 15 is not very likely; in fact, this occurs only 9.34% of the time. ■

It is important when we conduct a binomial sampling experiment that we maintain the conditions of *independence* and a *constant probability of success* from selection to selection. *Random* selection from a *large* population allows for this.

Importance of random selection

Note if selection in the above medical experiment were *not* random: let's say we used only members of the same family for our sample of 15. Family members may very well have similar genetic reactions to a drug. In this case, it would not be unlikely to get 100% of the sample, or even 0% of the sample cured. Generally, *nonrandom* samples violate the prime conditions for binomial sampling, and will usually destroy our ability to predict probabilities. With *random* selection we can be assured of maintaining a constant probability of a success from selection to selection, and thus obtain a true representation of the population.

Importance of a large population

Second, if selection had been from a *small* population (under 20 times the size of your sample), this would violate our condition of independence. For instance, let's say our entire *population* in the medical experiment were not many thousands but instead merely 30 individuals, of which 18 were cured (60%). Now, if we were to randomly sample 15 from this very small population, how many cured individuals we selected, let's say, on the first few picks would greatly affect the probabilities associated with later picks.

Actually, sampling from small populations can be dealt with using other statistical tools, but not the binomial.

Remember, random selection from a large population allows us to maintain the conditions of the early coin and dice experiments, namely independence and a constant probability of success from selection to selection. Serious violation of these conditions can render your results valueless (and remember, for all sampling in this text, we assume internal and external validity has been assured, as discussed in chapter 1).

One more point before we continue. Keep in mind, the normal curve gives an approximation. The histogram bars are wide and the normal curve may fit well, but the fit is not perfect. For instance, the precise answer to the above problem is 9.05%. Our answer is 9.34%. Most would consider this quite close. Generally, for larger values of np and $n(1 - p)$ (for instance, when both np and $n(1 - p)$ exceed 14) the normal curve approximation for most purposes is almost

exact. Of course, this leaves somewhat of a gap for np and $n(1 - p)$ between 5 and 14 in which we must exercise some professional judgment in evaluating probabilities. As a general rule, for np or $n(1 - p)$ between 5 and 14, the probabilities in the broad central region of the normal curve are considered reasonably accurate, while probabilities in the “very extreme” tails might best be verified with other methods. Other methods are available to get more precise answers, however these methods can be quite tedious to implement (again, more is discussed on these special cases in chapter 11). Now let’s try another example.

This next example is presented not only for practice but to demonstrate that the approximating normal curve may peak at a value of μ that is not a whole number, even though the data in the binomial histogram is classified into discrete whole-number categories.

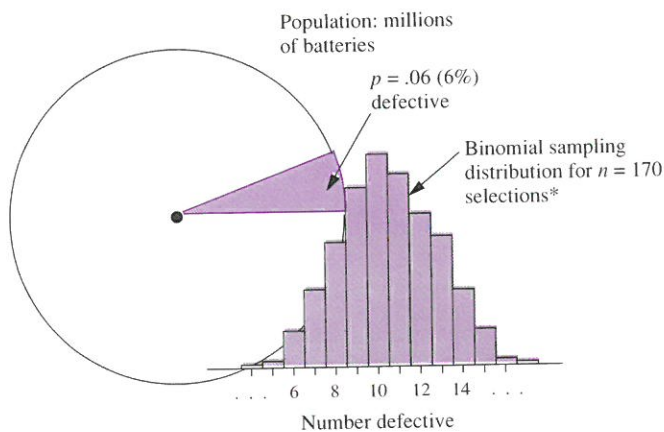
Example

Out of millions of assembly-line batteries produced last month by a large manufacturer, 6% were known to be defective.

Out of 170 randomly selected batteries from this population, find the probability that 14 or fewer of these will be defective.

Solution

This is binomial sampling since there are 170 fixed selections, each independent and each having the same 6% probability that a defective battery (a success) will be chosen. Random selection from a large two-category population generally satisfies the conditions for binomial sampling.



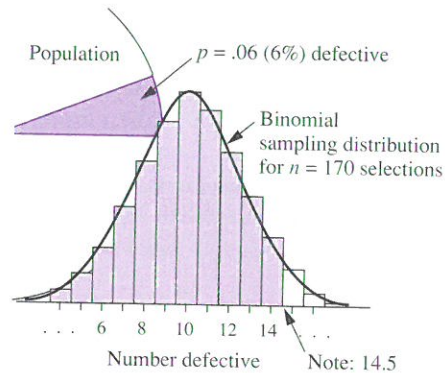
Since expected value $(np) = (170)(.06) = 10.2$ and $n(1 - p) = (170)(.94) = 159.8$, and both are greater than 5, the sampling distribution will be approximately normally distributed with mean and standard deviation calculated as follows:

$$\begin{aligned}\mu &= np \\ &= 170(.06) \\ &= 10.2\end{aligned}\qquad\begin{aligned}\sigma &= \sqrt{np(1 - p)} \\ &= \sqrt{170(.06)(.94)} \\ &= 3.096 \\ &= 3.1 \text{ (rounded)}\end{aligned}$$

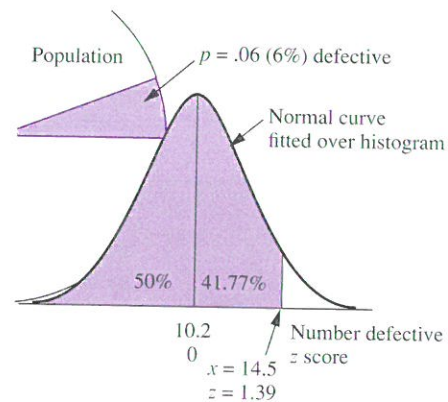
Note that the histogram shown is not quite symmetrical about any particular central value, thus the peak of the approximating normal curve will probably not be a whole number. In

this case, the approximating normal curve peaks at $\mu = 10.2$, which is not a whole number. So, to answer the question, what is the probability that out of our sample of 170 we will find 14 or fewer defective, we proceed as follows.

*Generated by computer simulation: Excel 5.0, Tools, Data Analysis, Random Number Generator, 1, 15000, Binomial Distribution, $p = .06$, 170, Histogram. The histogram represents the results of fifteen thousand random samples of size $n = 170$.



First, we shade the histogram bars representing 14 or less. Note the shading must extend to 14.5 to include the entire bar representing 14 defective. This half-unit adjustment is called your *continuity correction factor*. Now we fit a normal curve over the histogram.



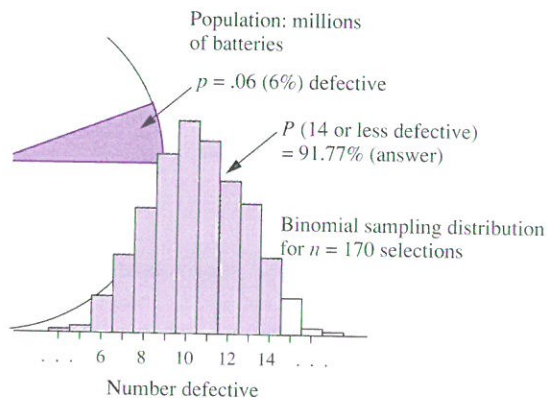
Second, we resketch the normal curve and shade the area 14.5 and below. Using $\mu = 10.2$ and $\sigma = 3.1$, we solve as we would any normal curve problem by first calculating the z score at the cutoff.

$$z = \frac{x - \mu}{\sigma} = \frac{14.5 - 10.2}{3.1} = \frac{4.3}{3.1} = 1.39$$

The % of data from $z = 0$ to $z = 1.39$ is 41.77%. Add this to 50% to get 91.77% (answer).

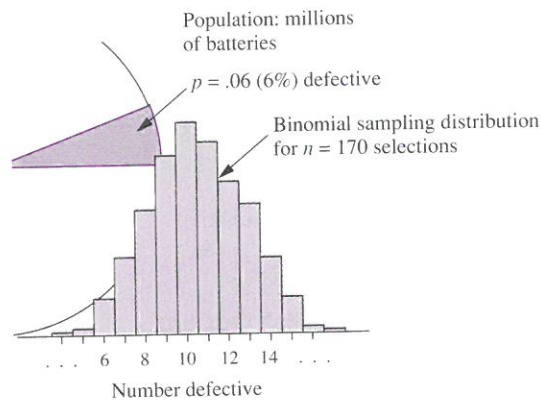
Answer

Now we can say, 91.77% of the time (or with probability .9177) we will achieve 14 or fewer defective batteries when we randomly sample 170 from our population, and this can be visually represented as follows:



Note that the normal curve is merely a tool, a device we use to lay over the histogram to help us determine the percentage of data in some portion of the histogram. In these binomial sampling experiments, there can never be, in reality, 10.2 defective batteries or 14.5 defective batteries. You can get 10 or 11 or 13 or 15 or any whole number of defective batteries but never 10.2 or 14.5. These numbers are location points on our estimating device, the normal curve.

One more point concerning this problem: Say we were employed as a Quality Control manager on the assembly line that produces these batteries and from a month's production we randomly sampled $n = 170$ and found 22 defective batteries, what would you conclude? Look at the histogram.



Certainly, if the population proportion were indeed $p = 6\%$ defective, then 22 defective batteries out of our sample of $n = 170$ would be an **extremely rare** event—in fact, nearly impossible. You can tell this just by looking at the histogram. The question is: did this *extremely rare* event occur or is the manufacturing process malfunctioning? In other words, is production out of control and no longer holding down the defective rate to $p = 6\%$? Certainly, a prudent quality control manager would investigate and would do so immediately before the process possibly degenerates further. ■

One last point concerning $np > 5$ and $n(1 - p) > 5$ in a binomial sampling experiment:

$$np = \text{Expected Number of Successes}$$

$$n(1 - p) = \text{Expected Number of Failures}$$

For example, in our battery experiment with $n = 170$ selections, we calculated $np = 10.2$ and $n(1 - p) = 159.8$. That is,

Expected Successes	+	Expected Failures	=	Total Selections
(10.2 defective batteries)	+	(159.8 okay batteries)	=	170 selections

So, instead of saying np and $n(1 - p)$ must each exceed 5 for the sampling distribution to be approximately normally distributed, we can say expected number of successes and expected number of failures must each exceed 5 for the sampling distribution to be approximately normally distributed, and the sum of these two numbers equals n , the total selections.

Summary

Perhaps the single most important distribution in all of statistics is the bell-shaped or normal distribution. The distribution was discovered seemingly under different circumstances by De Moivre (1733), Laplace (1781), and Gauss (1809) and encountered so frequently in experiments that sometime in the mid-to-late 1800s it adopted the name, *normal*.

Characteristics of the normal distribution:

Bell-shaped, fading at tails; symmetrical about μ , the mean, with 50% of the data in each half.
Approximately 68% of the data lies within ± 1

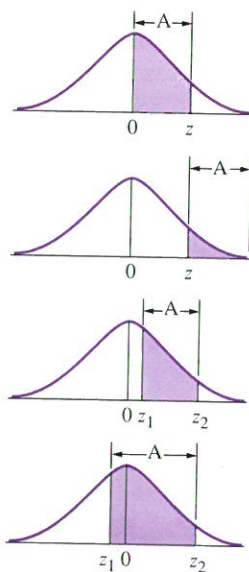
standard deviation of the mean, whereas approximately 95% of the data lies within ± 2 standard deviations of the mean.

Normal curve table: This table offers the percentage of data in the normal curve between $z = 0$ (the position of μ) and any z score you look up. Recall, a z score is the number of standard deviations a value is away from the mean. To precisely calculate the z score of a value, x , we use the formula

$$z = \frac{x - \mu}{\sigma}$$

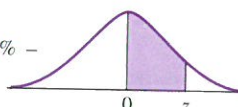
Normal Curve Table Usage

To Find Area, A

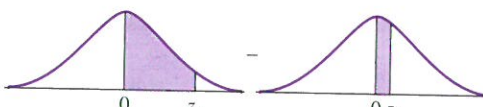


We Use the Following Procedure


A , the area between 0 and z can be found directly in the normal curve tables.

$$A = 50\% -$$


$$= 50\% - (\text{area between } 0 \text{ and } z)$$

$$A =$$


$$= (\text{area } 0 \text{ to } z_2) - (\text{area } 0 \text{ to } z_1)$$

$$A =$$


$$= (\text{area } 0 \text{ to } z_1) + (\text{area } 0 \text{ to } z_2)$$

Working backward: The normal curve table can also be used in reverse. If we know the percentage of data between 0 and z , we look up this area (in decimal form) in the table and determine the closest z value. If the percentage of data falls precisely midway between two values, we round to the higher z score.

Sampling from a Two-Category Population

Two-category population: A two-category population is a population where every member is classified into exactly one of two categories.

Sampling distribution: A sampling distribution shows us what we can expect when we randomly select n values (a fixed number) repeatedly from a particular population.

Binomial sampling distribution: The resulting sampling distribution when we randomly select n values repeatedly from a large two-category population, where each selection is independent and each has the same probability, p , a success will be chosen.

Large- n binomial sampling: For both np and $n(1 - p)$ greater than 5, the binomial sampling

distribution can be approximated with a normal curve with the following dimensions:

$$\begin{aligned}\mu &= np \\ \sigma &= \sqrt{np(1 - p)}\end{aligned}$$

Continuity correction: This refers to the $\frac{1}{2}$ -unit shading adjustment(s) necessary to include the entire width of the histogram bar(s) in question.

Discrete values: The binomial sampling distribution is sometimes referred to as a discrete data distribution, discrete, meaning values that when presented on a number line occupy only distinct unconnected (or isolated) points. However to assess probabilities we represent these discrete values with histogram bars, where the area of a histogram bar at some value represents the probability of achieving that value.

Small- n binomial sampling: For np or $n(1 - p)$ of 5 or less, the binomial sampling distribution is often skewed or sloping and the normal curve cannot be depended on to give proper estimates. For these cases, other techniques can be employed that are discussed in chapter 11, section 11.1.

Exercises

Note that full answers for exercises 1–5 and abbreviated answers for odd-numbered exercises thereafter are provided in the Answer Key.

4.1

- It was widely believed in the mid-1800s that given enough observations all natural phenomena, such as, heights, weights, reaction times, etc., from any common grouping will take on the shape of a normal distribution. Is this so? Explain.
- In the construction of the idealized normal curve, three primary assumptions were presented. List each and explain.
- The idealized normal curve has a number of characteristics. List four characteristics.

4.2 Use the normal curve table to determine the percentage of data in the normal curve

- between $z = 0$ and $z = .82$.
- above $z = 1.15$.
- between $z = -1.09$ and $z = .47$.
- between $z = 1.53$ and $z = 2.78$.

Work backward in the normal curve table to solve the following:

- 32% of the data in the normal curve can be found between $z = 0$ and $z = ?$
- Find the z score associated with the *lower* 5% of the data.
- Find the z scores associated with the *middle* 98% of the data.

4.3 Suppose the heights of all female students at University of Maryland in College Park are known to be normally distributed with $\mu = 5'5''$ and $\sigma = 2''$, find the percentage of female students

- under $5'2''$.
- between $5'2\frac{1}{2}''$ and $5'8''$.
- between $5'8\frac{1}{2}''$ and $5'9\frac{1}{2}''$.

4.4 Ebbinghaus in 1885, in a landmark experiment in Experimental Psychology, repeatedly measured the time necessary for an individual to memorize equal blocks of nonsense syllables (such as, zid, cuk, xot) and found the times to be normally distributed.

Using Ebbinghaus's data, suppose this individual takes an average of $\mu = 21.0$ minutes to complete the task of memorizing a block of nonsense syllables with standard deviation, $\sigma = 1.2$ minutes.

- Below what value would you expect to find the fastest 10% of the times? (Note: the *fastest* times would be *less than* 21.0 minutes, thus we shade the extreme left of the normal curve, estimating 10%.)
- Between what values would you expect to find the *middle* 50% of the times?

4.5 Selecting random samples of the same size *repeatedly* from a large two-category population creates a sampling distribution, known as a binomial sampling distribution, which is approximately normally distributed for expected value $(np) > 5$ and $n(1 - p) > 5$. Use this information to answer the following.

- In a population of many thousands of users of a new experimental drug designed to cure a specific form of bladder inflammation, it was found that 60% were cured. Suppose we randomly select 15 users from this population, what is the probability we will find 7 or less cured?
- Can we apply this population proportion ($p = 60\%$ cured) to future users, say in the case where the drug is to be distributed in another country? That is, can we expect about 60% cured? Discuss briefly.

4.6 Use the normal curve table to determine the percentage of data in the normal curve

- between ± 1 standard deviation.
- between ± 2 standard deviations.
- between ± 3 standard deviations.

4.7 Use the normal curve table to determine the percentage of data in the normal curve

- between $z = 0$ and $z = .38$.
- above $z = -1.45$.
- above $z = 1.45$.
- between $z = .77$ and $z = 1.92$.
- between $z = -.25$ and $z = 2.27$.
- between $z = -1.63$ and $z = -2.89$.

Work backward in the normal curve table to solve the following.

- 15% of the data in the normal curve can be found between $z = 0$ and $z = ?$
- Find the z score associated with the *upper* 73.57% of the data.
- Find the z scores associated with the *middle* 95% of the data.

4.8 Suppose standard IQ scores are known to be normally distributed with $\mu = 100$ and $\sigma = 15$. Find the percentage of individuals with IQ scores

- above 125.
- above 90.
- between 62 and 72.
- below 88.

Work backward in the normal curve table to solve the following.

- Above what value would you expect to find the *upper* 30% of IQ scores?
- Between what values would you expect to find the *middle* 75% of IQ scores?

4.9 Biological characteristics of a species are sometimes found to be near normally distributed. Suppose American anchovies, *Engraulis encrasicolus*, a species of herring commonly used on pizza, is known to have lengths that are normally

distributed with $\mu = 10.2$ centimeters (about 4") and $\sigma = .68$ centimeters (cm). Find the percentage of anchovies with lengths

- below 9.0 cm.
- below 10.7 cm.
- between 9.5 cm and 10.8 cm.
- between 11.0 cm and 11.4 cm.

Work backward in the normal curve table to solve the following.

- Above what length would you expect to find the *longest* 15% of anchovies?
- Between what lengths would you expect to find the *middle* 99% of anchovies?

4.10 Human characteristics are sometimes found to be near normally distributed. Quetelet in 1846 was probably the first to demonstrate this using the chest measurements of Scottish soldiers. He found the chest measurements to be normally distributed with $\mu = 39.5''$ and $\sigma = 2.5''$. Find the probability of randomly selecting a measurement

- between 36.5'' and 38.5''.
- above 38.2''.
- between 39.2'' and 40.6''.
- between 39.5'' and 44.7''.
- Below what value would you expect to find the *smallest* 40% of the chest measurements?
- Between what values would you expect to find the *middle* 96% of the chest measurements?

4.11 Galton demonstrated that large normal populations may, in fact, be comprised of several smaller normal populations. In 1875 he separated sweet pea seeds from the same parent by weight into several groups. Each group produced sweet peas with normally distributed weights but around different averages. When combined, these several smaller normal distributions formed into one large normal distribution centered around one common average.

Suppose the weights of a number of subspecies of Granny Smith apple combine to form one large normally distributed population of Granny Smith apple with $\mu = 6.9$ ounces (oz) and $\sigma = 1.1$ oz.

- What percentage of Granny Smith apples weigh more than 8.5 oz?
- What percentage of apples weigh between 7.2 oz and 8.0 oz?
- If you randomly select a Granny Smith apple, what is the probability the apple weighs less than 7.0 oz?
- Above what weight would you find the *heaviest* 65% of the apples?
- Between what weights would you find the *middle* 84% of the apples?

4.12 A binomial experiment is formally defined as a fixed number of trials (or selections), each independent and each having the same probability for success. Show how these conditions are met and solve the following.

- Out of 20 tosses of a coin, what is the probability of getting 13 to 15 heads?
- Out of $n = 50$ die tosses (one face of die is painted blue), what is the probability of turning up 10 or more blue faces?

4.13 The U.S. Military Academy at West Point is one of the nation's most selective colleges, accepting 11%* of applicants (according to the Insider's Guide to the Colleges). Out of $n = 60$ randomly selected applicants to the U.S. Military Academy,

- how many would you *expect* to be accepted?
- what is the probability 8 or less will be accepted?
- what is the probability between 5 and 7 will be accepted?
- what is the probability *exactly* 6 will be accepted?

4.14 67% of Americans feel secret files are being kept on them (based on data from *The Harper's Index*). Out of 25 randomly selected Americans, what is the probability 18 or more will feel secret files are being kept on them?

4.15 75% of those working in the visual, literary, or performing arts earn low wages from their art, under twelve thousand dollars per annum, based on data from Columbia's Research Center for Arts and

*Harvard accepts 15%.

Culture (*Columbia Magazine*, Summer 1990, p. 14). Out of 30 randomly selected artists,

- how many would you *expect* to earn low wages?
- what is the probability you will find at least 20 earning low wages?
- what is the probability you will find 24 to 27 earning low wages?

4.16 In a marketing population of phone calls, 3% produced a sale. If this population proportion ($p = 3\%$) can be applied to future phone calls, then out of 500 randomly monitored phone calls,

- how many would you expect to produce a sale?
- what is the probability of getting 11 to 14 sales?
- what is the probability of getting 12 or less sales?

4.17 In a study on aggression, 23% of mice exposed to severe conditions of overcrowding resorted to bizarre social behavior, such as cannibalism. If this is representative of all mice, out of a randomly selected group of $n = 100$ mice exposed to these severe conditions,

- find the probability you will get from 20 to 25 that exhibit bizarre social behavior.

- find the probability you will get 18 or less that exhibit bizarre social behavior.
- How valid is our assumption that $p = 23\%$ can be assigned to all mice? Discuss briefly.

4.18 88% of American high school students agree with their parents on the value of an education (according to studies from the University of Michigan, Institute for Social Research, "Monitoring the Future").* Out of $n = 45$ randomly selected American high school students,

- find the probability that 35 or more will agree with their parents on the value of an education.
- find the probability that 41 to 44 will agree with their parents on the value of an education.
- If we were to randomly sample $n = 45$ American high school students ten years from now, can we expect about 88% of the sample to agree with their parents on the value of an education? Discuss briefly.

*The same studies also revealed only 47% agreed with their parents on what's permitted on a date.

Endnotes

1. De Moivre, although born in France, was obliged to move to England as a young man under the Edict of Nantes (which restricted religious and civil liberties to Huguenots), and in England De Moivre worked as a mathematics tutor and consultant for wealthy patrons.

2. Actually De Moivre simulated the number of heads expected when n coins are dropped by using the expansion of $(1 + 1)^n$. One can also use the coefficients of the expansion $(a + b)^n$.

3. De Moivre did not use the term, standard deviation. In fact, technically the concept of standard deviation was not to be fully recognized for at least another seven decades, until after Legendre's work on least squares (1805) in which he demonstrated $\sum (x - \mu)^2$ was minimum about the mean (refer to chapter 9, section 9.0, under "Least-Squares Analysis" for

further reading on this). De Moivre's predictable distance was calculated to be $\frac{1}{2}\sqrt{n}$, which we now know as the standard deviation in a binomial experiment when $p = \frac{1}{2}$, that is, $\sigma = \sqrt{np(1-p)} = \sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}} = \frac{1}{2}\sqrt{n}$. In this case, where $n = 900$, $\sigma = \frac{1}{2}\sqrt{900} = \frac{1}{2} \cdot 30 = 15$. This is further discussed in section 4.4. De Moivre arrived at $\frac{1}{2}\sqrt{n}$ (actually, $\frac{1}{2}\sqrt{n-2}$, which is essentially equal to $\frac{1}{2}\sqrt{n}$ for large n) by determining the inflection points on the curve.

4. De Moivre's work at the time went relatively unnoticed and one can only speculate why. Perhaps the most probable reason is that mass statistical data was not as yet available, thus the practical application of De Moivre's discovery to social phenomenon could not be readily demonstrated—although De Moivre and a number of others felt it was only a matter of time until the laws of probability would be applied to a variety of social issues.

For an insightful discussion on this topic, refer to S. Stigler, *The History of Statistics* (Cambridge: Belknap Press, 1986), pp. 85–87.

5. It was unclear whether Laplace was familiar with De Moivre's work published 50 years earlier since he never mentioned De Moivre in his papers and his mathematical approach was quite different.

6. Laplace used the illustration of black and white tickets drawn from an urn.

7. At the time, Laplace (like De Moivre) was unaware of the concept of standard deviation. Laplace used a rather complex formulation to arrive at a suitable measure of spread. It was adequate for his purposes, but like much of Laplace's work exceedingly complex.

8. The actual figures were 251,527 males out of 493,472 births. All figures were scaled to 500,000 births for clarity.

9. The precise percentages were: Paris, 50.97% male births; London, 51.35%; Kingdom of Naples, 51.16% (Stigler, 1986).

10. According to *Newsweek* (April 16, 1990, p. 81), current averages worldwide are 50.6% male births, 49.4% female (102.5 males are born for every 100 females).

11. Gauss's reasoning essentially proceeded as follows: (1) it was generally acknowledged at the time that the arithmetic mean of several measurements was the best estimate of planetary position, (2) since the mean is most probable only if the errors are normally distributed, according to the method of least squares, then (3) errors must be normally distributed. Although one may find fault with Gauss's reasoning, the impact was monumental. Laplace seized on the argument giving it a solid base in logic based on his work with probability experiments.

Essentially, Laplace reasoned that a single observation must itself be an aggregate of more fundamental errors just like the outcome of 900 coins dropped on a table is the aggregate

of many head-tail outcomes. It is surprising Laplace himself had not made the discovery, considering his intense involvement in both astronomy and probability theory.

12. Planets such as Jupiter and Saturn were not used at sea to measure longitude because of their relatively slow movement and other difficulties of measurement while at sea. Moon craters were highly visible and the Moon's motion relatively fast, offering more precise measurements.

13. Gauss and others in the 1800s used a variety of standard distances from the mean, however many were multiples of the standard deviation, such as $.675 \sigma$, which was referred to as the *probable error*, since 50% of the errors were expected to fall within $\pm .675 \sigma$ of μ . In 1893, Pearson coined the term *standard deviation* and advocated its universal use.

14. Use of the normal distribution was confined mostly to astronomy for several decades and, thus, throughout much of the 1800s was

referred to as Gauss's law of error. Even to this day, the normal distribution is sometimes called the Gaussian distribution.

15. For further readings in this area, refer to H. Walker, *Studies in the History of Statistical Method* (Baltimore: Williams & Wilkins, 1929) and Stigler (1986).

16. In his newborn infant study (discussed in section 4.0), Laplace devised methods for calculating certain probabilities associated with the binomial distribution as $n \rightarrow \infty$, which Kramp in 1799 used to construct a full table of normal curve probabilities.

17. Kramp prepared the tables using $\sigma\sqrt{2}$ as the unit measure of dispersion, referred to as the *modulus*. Contemporary tables use σ , the standard deviation. Shepperd (1902) was the first to publish a table using σ , the standard deviation, as the unit measure.

18. E. W. Scripture, *The New Psychology* (1897), p. 443, as discussed and footnoted by Walker (1929), p. 24.

