Suppose there are $n$ experiments, and the probability that someone gets the right answer on any given experiment is $p$. So in the first example above, $n = 5$ and $p = 0.2$. Let $X$ be the number of correct results — this is a random variable (discrete). The formula (page 270 of the text) says

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}.$$ 

Here $n!$ is a special symbol known as factorial $n$. Factorial means you multiply together all the numbers from 1 to $n$. So

$$5! = 1 \times 2 \times 3 \times 4 \times 5 = 120.$$
\[ P(x) = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}. \]

The second part of the formula gives the probability of a specific sequence of S’s and F’s — so if there are exactly \( x \) S’s and \( n - x \) F’s, we calculate the probability as \( p \times p \times p \times \ldots \ (x \text{ times}) \) multiplied by \( (1 - p) \times (1 - p) \times (1 - p) \times \ldots \ (n - x \text{ times}) \) for an answer of \( p^x (1 - p)^{n-x} \).
\[ P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}. \]

The first part of the formula is the number of possible ways of arranging the S’s and F’s. For example, with \( n = 5 \) and \( x = 2 \) we find

\[
\frac{n!}{x!(n-x)!} = \frac{5!}{2!3!} = \frac{5 \times 4 \times 3 \times 2 \times 1}{2 \times 1 \times 3 \times 2 \times 1} = \frac{120}{12} = 10
\]

exactly as we got before by counting. However if \( n \) and \( x \) were large it would be very tedious to write out all the possible combinations of S’s and F’s, whereas this formula is relatively easy to apply.
So in the case $n = 5, \ x = 2, \ p = 0.2$ we have

$$P(x) = \frac{5!}{2!3!} \times 0.2^2 \times 0.8^3$$
$$= 10 \times 0.04 \times 0.512$$
$$= 0.2048.$$ 

Or with $n = 5, \ x = 4, \ p = 0.2$ we have

$$P(x) = \frac{5!}{4!1!} \times 0.2^4 \times 0.8^1$$
$$= \frac{5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1 \times 1} \times 0.0016 \times 0.8$$
$$= 5 \times 0.0016 \times 0.8$$
$$= .0064.$$
Assumptions

This formula is known as the *binomial distribution*. The essential conditions for a binomial distribution are:

1. Each outcome of the experiment has exactly two possible outcomes, which we label “success” and “failure” though they could have many different interpretations (alive or dead, rain or no rain, flight arrives on time or not, etc.). Data having this form are called *binary data*.

2. The experiment is repeated a number of times \((n)\) and the different experiments are *independent*.

3. The probability of “success” is some number \(p\) (between 0 and 1) and is the same for all the experiments.
Here is another example.

A tennis player serves her first serve in 70% of the time. Assume each serve is independent of all the others. She serves the ball six times. What is the probability that she gets

(a) All 6 serves in?

(b) Exactly 4 serves in?

(c) At least 4 serves in?

(d) No more than 4 serves in?
Solution:

(a) \( (0.7)^6 = 0.118 \).

(b) \( \frac{6!}{4!2!} \times (0.7)^4 \times (0.3)^2 = \frac{6 \times 5 \times 4 \times 3 \times 2 \times 1}{4 \times 3 \times 2 \times 1 \times 2 \times 1} \times (0.7)^4 \times (0.3)^2 = 15 \times 0.2401 \times 0.09 = 0.324 \).

(c) The probability of exactly 5 is \( \frac{6!}{5!1!} \times (0.7)^5 \times 0.3 = 6 \times 0.16807 \times 0.3 = 0.303 \). So the probability of at least 4 is 0.324 + 0.303 + 0.118 = 0.745.

(d) The probability of at least 5 is 0.303 + 0.118 = 0.421 so the probability of not more than 4 is 1 − 0.421 = 0.579.
The Normal approximation

Although the binomial formula is faster to calculate than trying to count all possibilities, it would still be hard to use for large samples, say \( n = 100 \). In this case, we use an alternative approach based on approximating the binomial distribution by a normal distribution.

Suppose we have a binomial distribution with \( n \) trials and probability of success \( p \), and \( n \) is some large number (say, 100). In this situation, we are not usually interested in the exact number of successes, but in the probability that the number will be more or less than some given number.
Example 1 (from text): In a certain week in 1997, the police at a certain location in Philadelphia made 262 car stops. Of these, 207 drivers were African American. Among the whole population of Philadelphia, 42.2% are African American. Does this prove the police were guilty of “racial profiling”, i.e. deliberately stopping drivers because they were African Americans?

Assuming the traffic stops are independent and the proportion of African Americans driving at this particular location is the same as the proportion in the whole city, this corresponds to the random variable $X$ (number of African Americans among those stopped) having a binomial distribution with $n = 262$, $p = 0.422$. The question is, what is the probability that $X \geq 207$ if the binomial distribution is correct?
Note: In this case it wouldn’t make sense to try to calculate the probability that $X$ is exactly 207. What we’re really concerned about is that the number is so large, so a natural question is “what is the probability that the number would have been as large as this by chance?” That leads us to consider $X \geq 207$ rather than $X = 207$. 
A Key Formula (page 274)

The binomial distribution for $n$ trials with probability $p$ of success on each trial has mean $\mu$ and standard deviation $\sigma$ given by

$$
\mu = np, \\
\sigma = \sqrt{np(1 - p)}.
$$
The solution proceeds by several steps:

**Step 1:** Calculate the mean of $X$. This is given by the formula

$$
\mu = np = 262 \times 0.422 = 110.6.
$$

**Step 2:** Calculate the standard deviation of $X$. This is given by the formula

$$
\sigma = \sqrt{np(1-p)} = \sqrt{262 \times 0.422 \times 0.518} = 7.99.
$$

**Step 3:** Convert the given $x$ value (207) to $z$. So

$$
z = \frac{x - \mu}{\sigma} = \frac{207 - 110.6}{7.99} = 12.07.
$$

**Step 4:** Calculate the probability associated with this $z$ value.
The only problem with step 4 is: the number’s off the chart! The regular table only goes up to 3.49. In fact, if you look at the little table in the bottom corner of page A2, you can see some further numbers:

<table>
<thead>
<tr>
<th>( z )</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>.999767</td>
</tr>
<tr>
<td>4.0</td>
<td>.9999683</td>
</tr>
<tr>
<td>4.5</td>
<td>.9999966</td>
</tr>
<tr>
<td>5.0</td>
<td>.999999713</td>
</tr>
</tbody>
</table>

Even at \( z = 5 \), the probability to the left of \( z \) (i.e. less than 5) is more than .999999, which means that the probability to the right of \( z \) is less than .0000001. Replace \( z = 5 \) by \( z = 12 \), and the probability of that is much smaller again.
Conclusion. The probability that we could have got this result (207 African Americans out of 262) by chance is so small that it is effectively 0. This seems to be completely convincing evidence that the police were engaging in the practice of racial profiling. However, there are other possible explanations — for example, perhaps the proportion of African Americans driving past this particular checkpoint was much greater than 42.2%.

Further Discussion. It is possible to compute the exact probability that $X \geq 207$ in this example: the answer is $4.9 \times 10^{-34}$. To give an idea of how small a probability that is, it is roughly equivalent to the probability that your favorite baseball team win the World Series 23 times in succession! [On the assumption that there are 30 Major League Baseball teams, that any one of them is equally likely to win in a given year, and that results from year to year are independent. $\left(\frac{1}{30}\right)^{23} = 1.1 \times 10^{-34}$.]
Example 2:

Consider our earlier example about the tennis player who gets in 70% of her serves. In a whole match she serves 80 times. What is the probability she makes at least 65 of these?
Solution:

First calculate $\mu$ and $\sigma$:

\[
\mu = 80 \times 0.7 = 56, \\
\sigma = \sqrt{80 \times 0.7 \times 0.3} = 4.1.
\]

Then for $x = 65$, we have $z = \frac{65 - 56}{4.1} = 2.20$. Look up 2.20 in the normal table: the corresponding left-hand probability is .9861. So the answer is $1 - 0.9861 = 0.0139$. In other words, it would be very unusual if she actually achieved this in a game, though it would be nothing like as “surprising” as our racial profiling example!

[Again it is possible to use a computer to calculate the exact probability. In this case it comes to .0161, compared with the above approximate answer of .0139. This gives an idea how accurate the normal approximation is. It’s not perfect, but it’s good enough for most practical calculations.]
Guidelines for the normal approximation to the binomial distribution (see sidebar p. 276): the binomial distribution can be approximated well by a normal distribution when the expected number of successes, $np$, expected number of failures, $n(1 - p)$, are both at least 15.

In the racial profiling example, $n = 262$, $p = 0.422$ so $np = 110.6$, $n(1 - p) = 151.4$. In the tennis example, $n = 70$, $p = 0.7$ so $np = 56$, $n(1 - p) = 24$. In both cases, the number is greater than 15 so the condition is satisfied.
Sampling Distributions

Example: An ABC News/Washington Post opinion poll published February 23 stated that President Obama currently has an approval rating of 68% (among all voters — the ratings are sharply different among Democrats, Independents and Republicans).

This is based on a sample of 1001 voters. The margin of error is described as plus or minus 3%.

What exactly does this mean?
Let’s focus on the proportion of people who support the President. In this case 68% is a statistic — the number calculated from the sample. The true proportion in the population is an unknown value $p$.

Collecting a sample is essentially a binomial distribution, with $n = 1001$. However most opinion polls are reported as the percentage or proportion of people who vote a certain way, rather than the total number. Therefore, our interest is in the sample proportion. If $X$ is the number of people who support Obama in the poll, then the sample proportion is $X/n$ (so in this example, $X$ was about 681, which would lead to $X/n = 681/1001 = 0.68$ to two decimal places).
For a sample proportion we have (see sidebar, page 280):

\[
\text{Mean} = p,
\]

\[
\text{Standard Deviation} = \sqrt{\frac{p(1 - p)}{n}}.
\]

So assuming \( p = 0.68 \), in this case we get a mean of 0.68 and a standard deviation of \( \sqrt{\frac{0.68 \times 0.32}{1001}} = 0.0147 \).

Also the normal distribution applies (because again \( np > 15 \), \( n(1 - p) > 15 \)), so we can assume the sampling distribution is approximately normal.
**Conclusion:** If the true value of $p = 0.68$, then in repeated samples of size 1008, the sampling distribution will be approximately normal with a mean of 0.68 and a standard deviation of 0.0147.

In particular, approximately 95% of all polls will result in a sample proportion within 0.0294 (2 standard deviations, or 3 percentage points) of the mean, and approximately 99.7% of all polls will result in a sample proportion within 0.0441 (3 standard deviations, or 4.4 percentage points) of the mean.

Because the quantity $\sqrt{\frac{p(1-p)}{n}}$ plays a special role in this calculation, it is given a special name — the *standard error*. 
Side comment: In this example, we don’t actually know that $p = 0.68$ is the correct value for the whole population — that’s only an estimate based on the sample. But the value of the standard error is not all that sensitive to $p$ — for example, it would be 0.0136 at $p = 0.75$ and 0.0157 at $p = 0.5$ (compared with 0.0147 at $p = 0.68$) — these numbers are not very much different. Strictly speaking, the standard error is only an estimate of the standard deviation of the proportion, but this point is usually ignored in practice.
Example (Problems 6.47 and 6.48, p. 287). In the exit poll of 3160 voters in the California recall election for Governor, the actual proportion of voters who voted for the recall was 55%.

(a) Define a binary random variable that represents the vote for one voter (1=vote for recall). State probability distribution.

(b) Identify \( n \) and \( p \) for the binomial distribution that is the sampling distribution for the number in the sample who voted for the recall out of the sample of size 3160.

(c) Find mean and SD of the binomial random variable in (b).

(d) Find mean and SD of the sampling distribution of the proportion of the 3160 people in the sample who voted for the recall.

(e) In (d), if the mean was 0.55 and the standard error was 0.0089, based on the approximate normality of the sampling distribution, give the interval of values within which the sample proportion will almost certainly fall.

(f) Based on the result in (e), if you take an exit poll and observe a sample proportion of 0.60, would you be safe in concluding that the population proportion probably exceeds 0.55? Why?
Brief answers:
(a) $P(X = 1) = 0.55$, $P(X = 0) = 0.45$.
(b) $n = 3160$, $p = 0.55$.
(c) $\mu = np = 3160 \times 0.55 = 1738$. $\sigma = \sqrt{np(1 - p)} = \sqrt{3160 \times 0.55 \times 0.45} = 27.97$.
(d) $\mu = p = 0.55$, $\sigma = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.55 \times 0.45}{3160}} = 0.0089$.
(e) By the “empirical rule” from Chapter 2, which applies because the binomial distribution is approximately normal for such a large sample size, the sample proportion is almost certain to lie within three standard deviations of the population proportion, which in this case would translate to between $0.55 - 3 \times 0.0089 = 0.523$ and $0.55 + 3 \times 0.0089 = 0.577$.
(f) Yes, because if $p = 0.55$ the sample proportion is almost certain to be less than 0.60, and if $p$ were any smaller than 0.55, it would be even more certain that the sample proportion is less than 0.60. Therefore, the only reasonable conclusion is that $p > 0.55$. 
Additional Comment

We could make the calculation in (f) a little more precise by computing the normal approximation that the sample proportion is greater than 0.6 based on \( \mu = 0.55, \sigma = 0.0089 \). In this case \( z = \frac{0.60 - 0.55}{0.0089} = 5.62 \), so the left-hand tail probability is greater than .9999997 by the thumbnail in the bottom left corner of page A2, hence the right-hand tail probability is less than .0000003. This gives some quantitative interpretation to the phrase “almost certain” — in this case, between 4 and 5 consecutive World Series wins for the Red Sox!
How close are sample means to population means?

Let’s go back to the “women’s heights” example from Chapter 2.

In our class, I took a survey in which I asked students their heights. Of the 61 female students who responded, the mean was 65.5 and the standard deviation 3.23. The question now is: how much variability is there in these numbers? In particular, how close is that 65.5 likely to be to the true mean of all female students at UNC?
We do a simulation experiment. Individual samples have means like 66.1, 65.8, 65.3, 65.5...

Now let’s do 1000 samples. The mean of the sample means is 65.495 and the standard deviation 0.406.

Try increasing the size of an individual sample, say to 1001 (a typical size for an opinion survey)

Now the individual samples have means like 65.44, 65.41, 65.55, 65.49, 65.56

In a repetition of 1000 samples of size 1001, the mean is 65.496 and the standard deviation .099.
Conclusions:

1. The standard deviation of the sample mean (resp. 0.406, 0.099 for the two sample sizes) is much smaller than the standard deviation for the observations themselves.

2. The distribution of the sample mean is also much closer to normal as indicated by the shape of the histogram.

Property 2 is called the **Central Limit Theorem**.
Formula: The standard deviation of $\bar{x}$ (the sample mean of a sample of size $n$ is

$$\frac{\sigma}{\sqrt{n}}.$$ where $\sigma$ is the standard deviation of a single observation.

The quantity is known as the standard error (the same word as we used with sample proportions, but this is more general, because we’re not necessarily estimating the mean of binary random variables — they can be anything).

With $\sigma = 3.23$, $n = 61$ we have $\frac{\sigma}{\sqrt{n}} = 0.414$, very close to the 0.406 we saw in our simulation.

With $\sigma = 3.23$, $n = 1001$ we have $\frac{\sigma}{\sqrt{n}} = 0.102$, very close to the 0.099 we saw in our simulation.

Moreover, it’s very close to a normal distribution so we can use that to calculate probabilities.
Problem 6.55.

A roulette wheel in Las Vegas has 38 slots. If you bet $1 on a particular number, you win $35 if the ball ends up in that slot and $0 otherwise. All outcomes are equally likely.

(a) Let $X$ be your winnings when you play once. What is the probability distribution of $X$? (It has mean 0.921 and standard deviation 5.603.)

(b) You play 5040 times. Show that the sampling distribution of your sample mean winnings has mean 0.921 and standard deviation 0.079.

(c) Using the central limit theorem, find the probability that your mean winnings is at least $1, so that you don’t lose money overall.
Solution

(a) The table is

<table>
<thead>
<tr>
<th>$x$</th>
<th>$P(x)$</th>
<th>$xP(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$35$</td>
<td>$\frac{1}{38}$</td>
<td>$\frac{35}{38}$</td>
</tr>
<tr>
<td>$0$</td>
<td>$\frac{37}{38}$</td>
<td>$0$</td>
</tr>
</tbody>
</table>

Mean is $\frac{35}{38} = 0.921$; SD is $5.603$ (given)

(b) $\frac{5.603}{\sqrt{5040}} = 0.079$.

(c) With $x = 1, \mu = 0.921, \sigma = 0.079$, we have $z = \frac{1-0.921}{0.079} = 1$; left tail probability is $0.84$; right tail probability is $1 - 0.84 = 0.16$ (the answer).
Problem 6.57.

Weekly income of farm workers has \( \mu = 500, \sigma = 160 \) in N.Z. dollars (right-skewed distribution). A researcher plans to sample \( n = 100 \) and use sample mean \( \bar{x} \) to estimate \( \mu \).

(a) Show that the standard error of \( \bar{x} \) is 16.0

(b) Explain why it is almost certain that the sample mean will fall within $48 of $500

(c) Find the probability that \( \bar{x} \) lies within $20 of \( \mu \) (regardless of the true value of \( \mu \)).
Solution

(a) \( \frac{\sigma}{\sqrt{n}} = \frac{160}{\sqrt{100}} = 16 \).

(b) $48$ is three times the standard error, so it’s true by “empirical rule”

(c) The \( z \) scores are \( \frac{20}{16} = 1.25 \) and \( -\frac{20}{16} = -1.25 \).

The respective left tail probabilities are 0.894 and 0.106.

The difference is 0.894–0.106=0.788.