Chapter 8: Sampling Distributions

Consider the following three news items.

*The **average price of unleaded regular fell by 1.6 cents to $3.667 a gallon** on Saturday, from $3.683 a gallon, according to survey results from the motorist group AAA. (Source: CNNMoney.com)*

*The Census Bureau on Tuesday released the 2007 American Community Survey, **the government's annual estimates** of social, economic and housing characteristics for the nation. Among the highlights: **25.3 - In minutes, the average commute to work in 2007**, an increase from 25.0 minutes in 2006. (Source: Chicago Tribune)*

*Barack Obama leads John McCain, **49% to 44%**, when registered voters are asked who they would vote for if the election were held today, according to the latest Gallup Poll Daily tracking update. (Source: Gallup)*

All three of these are *estimates based on samples.* In fact, they're probably not correct, due to sampling error. Our goal in this chapter and the next is to characterize how close we think we are. In Section 8.1, we'll talk about sample means (the first two examples above), and in Section 8.2, we'll talk about proportions (the third example).

If you're ready to begin, just click on the "start" link below, or one of the section links on the left.

Section 8.1: Distribution of the Sample Mean

8.1 Distribution of the Sample Mean
8.2 Distribution of the Sample Proportion

Objectives

By the end of this lesson, you will be able to...

1. describe the distribution of the sample mean for samples obtained from normal populations
2. describe the distribution of the sample mean for samples obtained from a population that is not normal

---

Sampling Distributions

Consider the following three news items.

> The **average price of unleaded regular fell by 1.6 cents to $3.667 a gallon** on Saturday, from $3.683 a gallon, according to survey results from the motorist group AAA. (Source: CNNMoney.com)

> The Census Bureau on Tuesday released the 2007 American Community Survey, **the government's annual estimates** of social, economic and housing characteristics for the nation. Among the highlights: **25.3 - In minutes, the average commute to work in 2007**, an increase from 25.0 minutes in 2006. (Source: Chicago Tribune)

> Barack Obama leads John McCain, **49% to 44%**, when registered voters are asked who they would vote for if the election were held today, according to the latest Gallup Poll Daily tracking update. (Source: Gallup)

All three of these are *estimates based on samples* In fact, they're probably not correct, due to sampling error. From Section 1.4,

Sampling error is the error that results from using a sample to estimate information regarding a population.

The idea is this - unless we sample every single individual in the sample, there will be some error in our results. Our goal in this section will be to characterize the distribution of the sample mean.

The Distribution of the Sample Mean

Let's look again at the definition of a **random variable**, from Section 6.1.

A **random variable** is a numerical measure of the outcome of a probability experiment whose value is determined by chance.

Think about the sample mean, $\bar{x}$. Isn't it's value determined by chance as well? Since we the individuals in a sample are *randomly* selected, the sample mean will depend on those individuals selected, so it, too, is a random variable. The big question, then, is the distribution of $\bar{x}$ - in other words, what are its mean (the mean of the sample mean, $\mu_{\bar{x}}$) and its standard deviation (the standard deviation of the sample mean, $\sigma_{\bar{x}}$)?

To investigate these, let's look at a particular population.

Consider the heights of the players from the starting line-up from the 2008 Men's Olympic Basketball gold medal game - Jason Kidd (76"), LeBron James (80"), Kobe Bryant (78"), Carmelo Anthony (78"), and Dwight Howard (83"). (Source: NBC Sports) The mean of the population is 79", with a standard deviation of 2.37"

First, let's consider the different samples of size 2. There are 10 such samples ($_5C_2 = 10$), shown below, along with their corresponding sample means.

sample	heights	$\overline{x}$
Kidd, James	76, 80	78
Kidd, Bryant	76, 78	77
Kidd, Anthony	76, 78	77
Kidd, Howard	76, 83	79.5
James, Bryant	80, 78	79
James, Anthony	80, 78	79
James, Howard	80, 83	81.5
Bryant, Anthony	78, 78	78
Bryant, Howard	78, 83	80.5
Anthony, Howard	78, 83	80.5

Interestingly, the mean of the sample means of size 2 is 79" as well. This is actually reasonable, though, because we know that the mean of a random variable is also its *expected value*, and it makes perfect sense that the value we should *expect* from the sample mean is the same as the population mean!

The standard deviation, though, is very different. It helps to look at things visually. The image below represents all possible sample means for samples of size 1 (individuals), 2, 3, 4, and 5 (the population). Pay particular attention to the standard deviation.
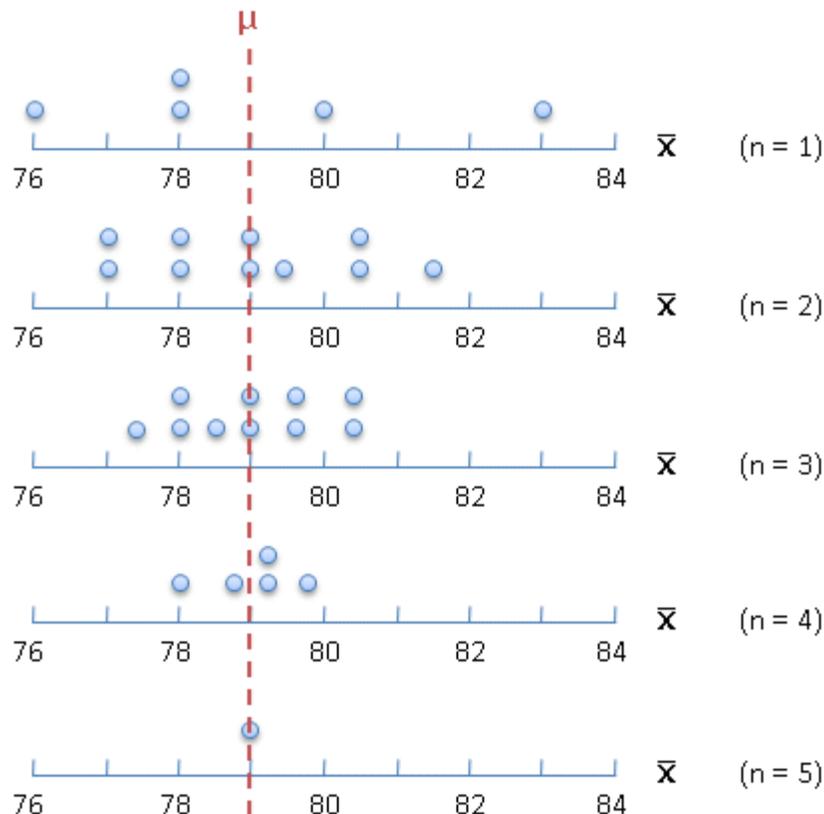
The interesting things to note here are that $\mu_{\overline{x}} = 79$, regardless of the sample size, but the standard deviation decreases as n increases. If we think about this a bit, this too, is reasonable. The more individuals we have in our sample, the more likely we are to be closer to the true mean. Things brings us to our first major point.

> **The Law of Large Numbers**
>
> As n increases, the difference between $\bar{x}$ and μ approaches zero.

We're now ready to investigate the standard deviation of $\bar{x}$ a bit more in-depth.

The Central Limit Theorem

> **The Central Limit Theorem**
>
> Regardless of the distribution shape of the population, the sampling distribution of $\bar{x}$ becomes approximately normal as the sample size n increases (conservatively n≥30).

This is very interesting! So it doesn't matter if the distribution shape was left-skewed, right-skewed, uniform, binomial, anything - the distribution of the sample mean will always become normal as the sample size increases. What an amazing result!

Exploring the Distribution of the Sample Mean

To do some exploring yourself, go to the Demonstrations Project from Wolfram Research, and download the Central Limit Theorem demonstration. If you haven't already, download and install the player by clicking on the image to the right.

Once you have the player installed and the Central Limit Theorem demonstration downloaded, move the slider for the *sample size* to get a sense of its affect on the distribution shape. You can also move the *new sample* slider to get a different sample.
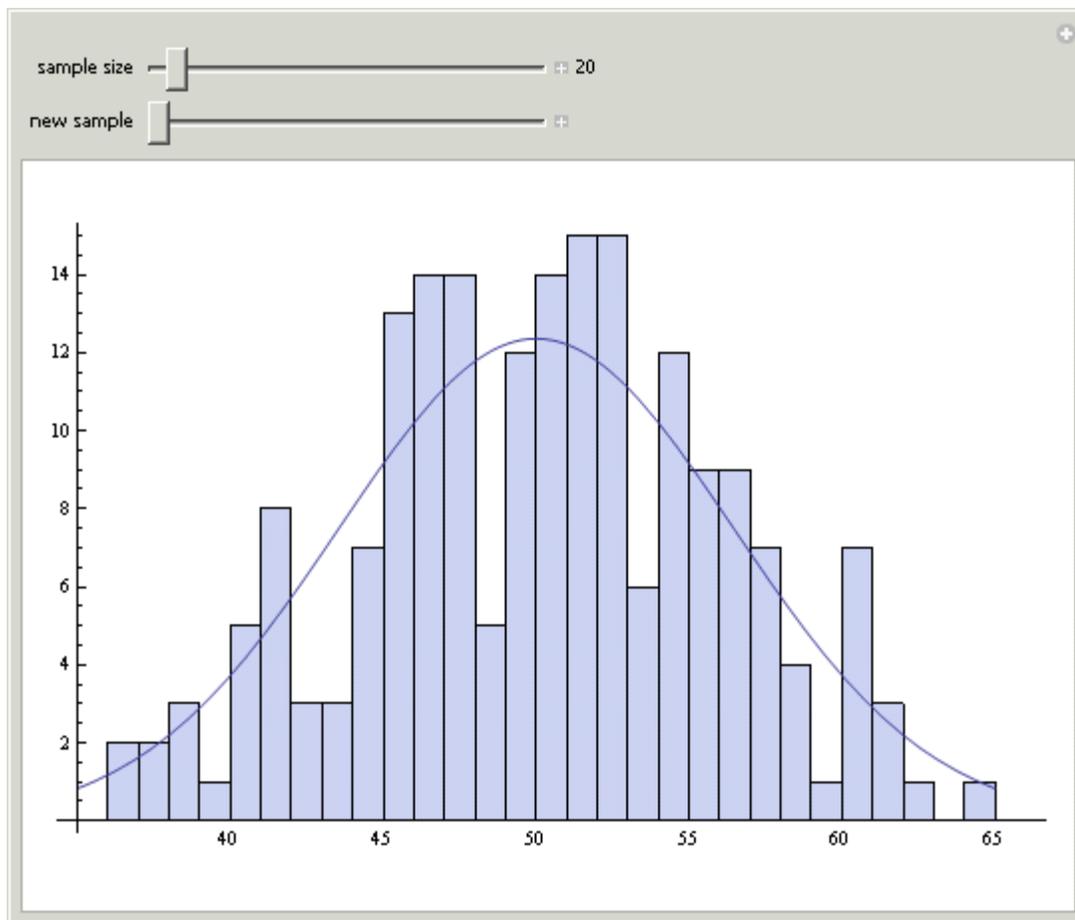
The Distribution of the Sample Mean

We can even be more specific about the distribution of $\bar{x}$:

The Sampling Distribution of $\bar{x}$

If a simple random sample of size n is drawn from a large population with mean μ and standard deviation σ, the sampling distribution of $\bar{x}$ will have mean and standard deviation:

$$\mu_{\bar{x}} = \mu \text{ and } \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where $\sigma_{\bar{x}}$ is the **standard error of the mean**.

Now let's apply this distribution to various problems.

Using the Central Limit Theorem

In order to find probabilities about a normal random variable, we need to first know its mean and standard deviation. With the results of the Central Limit Theorem, we now know the distribution of the sample mean, so let's try using that in some examples.

Let's see a couple examples.

Example 1

Let's consider again the distribution of IQs that we looked at in Example 1 in Section 7.1.

We saw in that example that tests for an individual's intelligence quotient (IQ) are designed to be normally distributed, with a mean of 100 and a standard deviation of 15.

What is the probability that a randomly selected sample of 20 individuals would have a mean IQ of more than 105?

Solution:

To answer this question, we need to find $P(\bar{x} > 105)$, if n = 20. Before we can do that, we need to first find the distribution of $\bar{x}$. From the distribution of the sample mean, we know $\mu_{\bar{x}} = \mu = 100$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 15/\sqrt{20} \approx 3.35$.

Here's what the distribution of $\bar{x}$ looks like in relation to the distribution of X.

$\bar{X}: \mu_{\bar{x}} = 100, \sigma_{\bar{x}} = 3.35$

$X: \mu = 100, \sigma = 15$
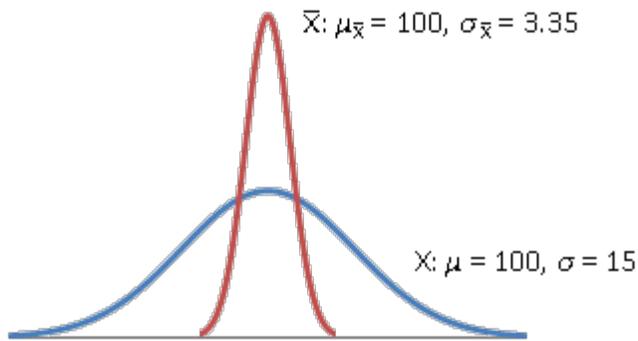
Now that we have the standard deviation, we can find the probability. Using StatCrunch...

| Mean: 100 | Std. Dev.: 3.35 |
| Prob(X => ▼ 105 |) = 0.06777917 |

Example 2

In Example 2 in Section 7.1, we were told that weights of 1-year-old boys are approximately normally distributed, with a mean of 22.8 lbs and a standard deviation of about 2.15. (Source: About.com)

Suppose the sample mean of the 10 1-year-old boys at the Kiddie Care day care center is 22.3 lbs. Is that unusual?

Source: stock.xchng

Solution:

In order to determine if an event is *unusual*, we need to find its probability. If the probability of the event is less than 5%, we can classify it as an unusual event.

In this case, we want to find the probability of observing a sample mean of 22.3 or less. Using the distribution of the sample mean, $\mu_{\bar{x}} = \mu = 22.8$ and $\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2.15/\sqrt{10} \approx 0.68$. Using StatCrunch...

| Mean: 22.8 | Std. Dev.: 0.68 |
| Prob(X <= ▼ 22.3 |) = 0.23108019 |

So we'd observe a sample mean of 22.3 lbs or less from a sample of 10 1-year-old boys about 23% of the time, which is not very unusual at all.

Here's one for you to try:

Example 3

Suppose that a particular professor's Statistics exam on probability traditionally has a mean score of 74, with a standard deviation of 11.

The professor suspects that his current crop of students is very strong. To

compare, he gives them the same exam he has in the past. The sample mean of the 28 students in his current class was a 78.

Was the professor correct? Is his current class of students unusual compared to those from the past?

[reveal answer]

Section 8.2: Distribution of the Sample Proportion

8.1 Distribution of the Sample Mean
8.2 Distribution of the Sample Proportion

Objectives

By the end of this lesson, you will be able to...

1. describe the sampling distribution of a sample proportion
2. compute probabilities of a sample proportion

The Sample Proportion

Consider these recent headlines:

> ***Hispanics See Their Situation in U.S. Deteriorating***
> *Half (50%) of all Latinos say that the situation of Latinos in this country is worse now than it was a year ago, according to a new nationwide survey of 2,015 Hispanic adults conducted by the Pew Hispanic Center. (Source: Pew Research)*

> ***Automatic enrollment in 401(k) doesn't take care of everything***
> *Never got around to signing up for the company retirement plan? The boss may have done it for you. Forty-two percent of employers with 401(k) plans automatically enroll new or existing employees in the plans, nearly double the 23 percent from 2006, according to estimates from the 2008 401(k) Benchmarking Survey by the International Foundation of Employee Benefit Plans and Deloitte Consulting. The survey polled 436 employers with workforces of all sizes. (Source: Chicago Tribune)*

> ***Stem cell, marijuana proposals lead in Mich. poll***
> *A recent poll shows voter support leading opposition for ballot proposals to loosen Michigan's restrictions on embryonic stem cell research and allow medical use of marijuana. The EPIC-MRA poll conducted for The Detroit News and television stations WXYZ, WILX, WOOD and WJRT found 50 percent of likely Michigan voters support the stem cell proposal, 32 percent against and 18 percent undecided. (Source: Associated Press)*

These three articles all have something in common - they're referring to *sample proportions* - 50% of all Latinos, 42% of employers, and 50% of likely Michigan voters, respectively, in the three articles above. Proportions are the number with that certain characteristics divided by the sample size.

> In general, if we let x = the number with the specific characteristic, then the **sample proportion**, $\hat{p}$, (read "*p-hat*") is given by:
>
> $$\hat{p} = \frac{x}{n}$$
>
> Where $\hat{p}$ is an estimate for the population proportion, p.

Let's focus for a bit on x, the number with that characteristic. If we rephrase that a bit, and consider an individual having that characteristic as a "success", we can see that x follows the binomial distribution.

From Section 6.2, we know that the distribution of a binomial random variable becomes bell-shaped as n increases. The three histograms below demonstrate the effect of the sample size on the distribution shape.

n=10, p=0.8

$\mu_x = 8$

n=20, p=0.8

n=50, p=0.8
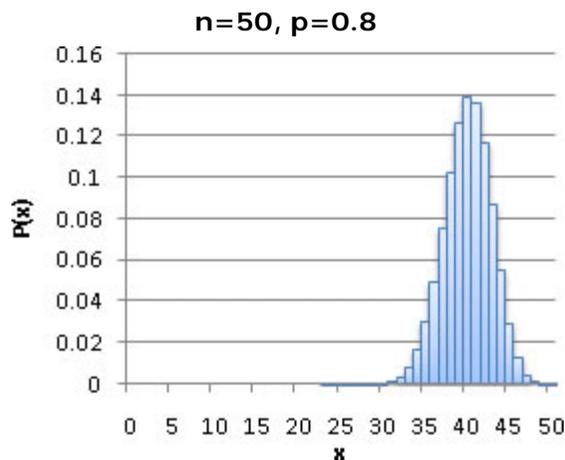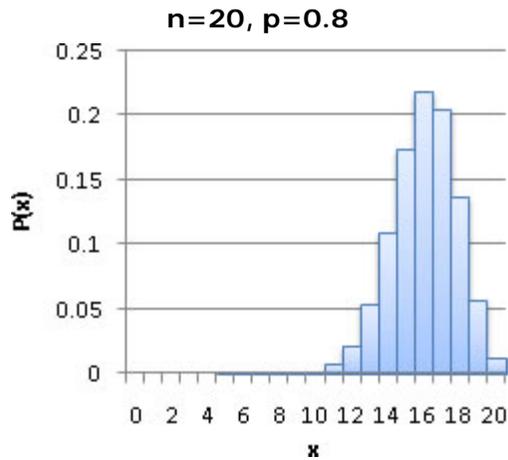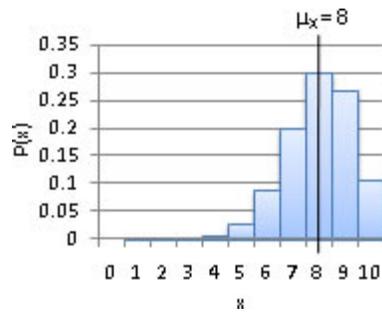
As the number of trials in a binomial experiment increases, the probability distribution becomes bell-shaped. As a rule of thumb, **if np(1-p)≥10, the distribution will be approximately bell-shaped**.

With our new knowledge of the normal distribution, it appears that if np(1-p)≥10, x is normally distributed, which implies that $\hat{p}$ is as well. All we need to know, then, is its mean and standard deviation.

Sampling Distribution of $\hat{p}$

For a simple random sample of size n such that n≤0.05N (in other words, the sample is less than 5% of the population),

- The shape of the sampling distribution of $\hat{p}$ is approximately normal provided np(1-p)≥10
- The mean of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$.

The standard deviation of the sampling distribution of $\hat{p}$ is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

Now that we know how $\hat{p}$ is distributed, we can use that information to find some probabilities.

Let's see a couple examples.

Example 1

In a typical class, about 70% of students receive a C or better. Out of a random sample of 100 students, what is the probability that less than 60 receive a C or better?

Solution:

Since there are millions of students, 100 is definitely less than 5% of the population.

Because $np(1-p) = 100(0.7)(1-0.7) = 21 \geq 10$, we can say that the distribution of $\hat{p}$ is normal.

We then need to find the mean and standard deviation. From the distribution of the sample proportion, we know

$$\mu_{\hat{p}} = p = 0.7$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.7(1-0.7)}{100}} \approx 0.046$$

Using StatCrunch, the probability of observing a sample proportion of less than 60/100 = 0.6 is

Mean: 0.7		Std. Dev.: 0.046
Prob(X <= ▼ 0.6	) =	0.014855833
	Close Compute	

So the probability of observing less than 60 out of 100 students is about 0.015 - pretty unusual.

Example 2

A basketball player traditionally makes 85% of her free throws. Suppose she shoots 100 free-throws during practice. Would it be unusual for her to make less than 75?

Source: stock.xchng

Solution:

Since our player can shoot an unlimited number of free throws, we can assume that 100 is less than 5% of the population. Checking the distribution:
$$np(1-p) = 100(0.85)(1-0.85) = 12.75 \geq 10$$

So the distribution should be normally distributed, with mean and

standard deviation

$$\mu_{\hat{p}} = p = 0.85$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.85(1-0.85)}{100}} \approx 0.036$$

Using StatCrunch, with p = 75/100 = 0.75:

Mean: 0.85	Std. Dev.: 0.036
Prob(X <= ▼ 0.75) =	0.0027366017
Close Compute	

So the probability that she makes less than 75 out of 100 is about 0.0027. That means she'll make less than 75 out of 100 about 0.27% of the time, which is very unusual.

Here's one for you to try:

Example 3

According to data from the American Cancer Society, about 3.86% of women develop breast cancer between the ages of 40-59.

What is the probability that in a random sample of 500 39-year-old women without breast cancer, more than 20 will develop breast cancer by the age of 60?

[reveal answer]

A note on rounding: We've already seen that rounding is an important topic in this course. In this particular section, rounding can make a significant difference in your calculations. It's important that you take a minute and watch this video regarding the effect of rounding when doing the calculations from this section: The Dangers of Rounding (Quicktime or iPod)