Chapter 7: Sampling Distributions

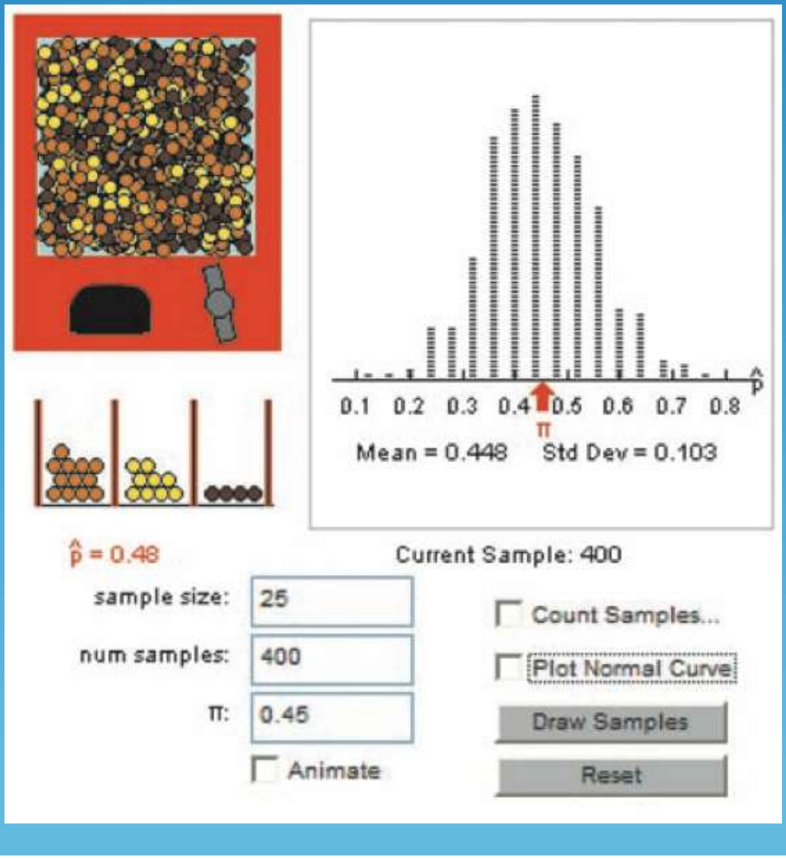
Section 7.2
Sample Proportions

The Sampling Distribution of $\hat{p}$

How good is the statistic $\hat{p}$ as an estimate of the parameter p?

To find out, we ask, “What would happen if we took many samples?” The **sampling distribution of** $\hat{p}$ answers this question.

The figure below shows one set of possible results from Step 4 of “The Candy Machine” Activity. Let’s describe what we see.



Mean = 0.448 Std Dev = 0.103

$\hat{p}$ = 0.48

Current Sample: 400

sample size: 25

num samples: 400

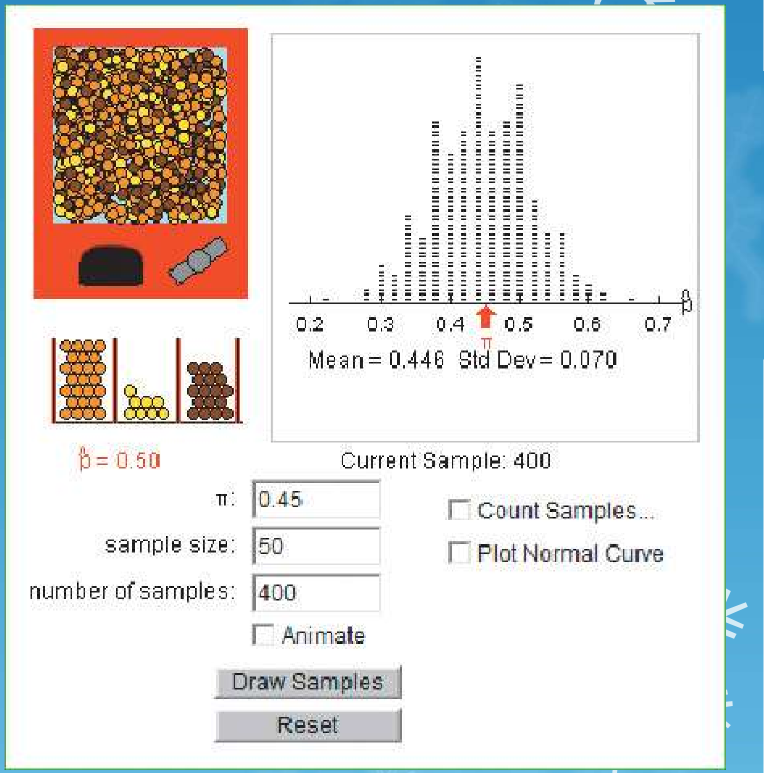
π: 0.45

□ Animate

□ Count Samples...

□ Plot Normal Curve

Example 1: In a similar way, we can explore the sampling distribution of $\hat{p}$ when n = 50 (Step 5 of the Activity). As the figure below shows, the dotplot is once again roughly symmetric and somewhat bell-shaped. This graph is also centered at about 0.45. With samples of size 50, however, there is less spread in the values of $\hat{p}$. The standard deviation in the figure below is 0.070. For the samples of size 25 in the figure above, it is 0.105. To repeat what we said earlier, larger samples give the sampling distribution a smaller spread.



Mean = 0.446 Std Dev = 0.070

$\hat{p}$ = 0.50

Current Sample: 400

π: 0.45

sample size: 50

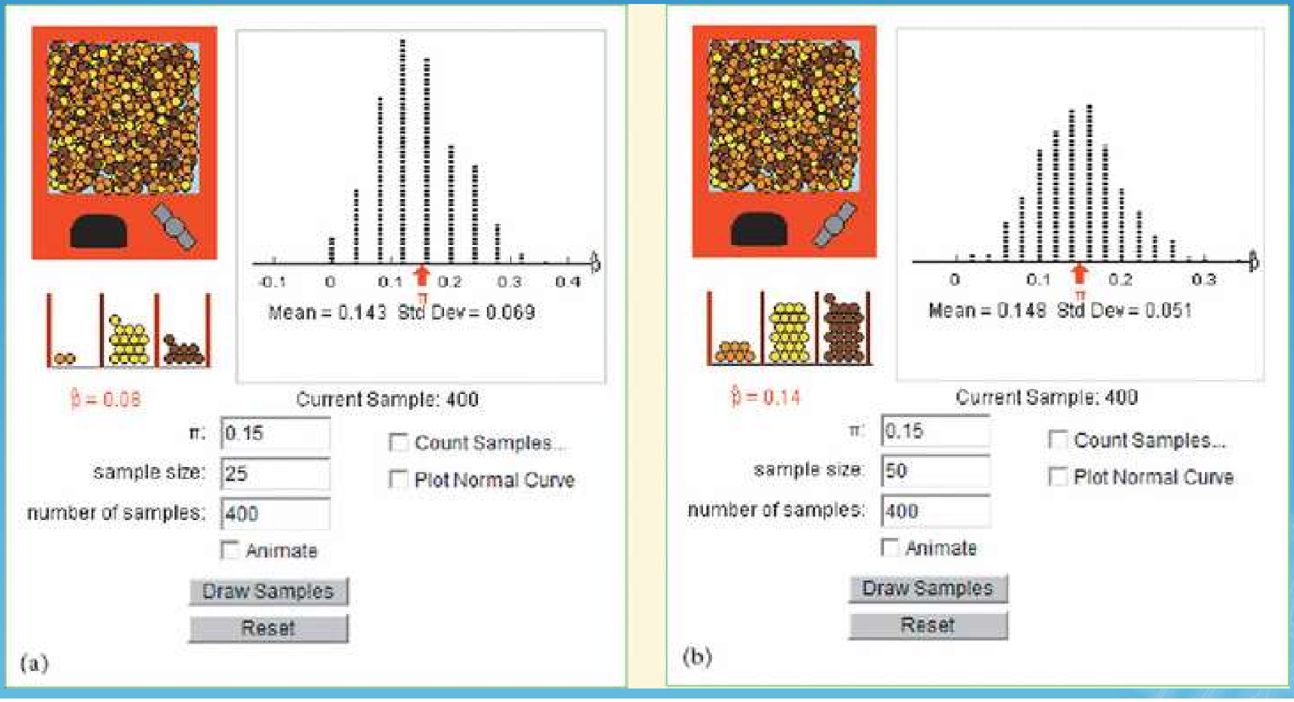
number of samples: 400

□ Count Samples...

□ Plot Normal Curve

□ Animate

What if the actual proportion of orange candies in the machine were p = 0.15? The figure below on the left shows the approximate sampling distribution of $\hat{p}$ when n = 25. Notice that the dotplot is slightly right-skewed. The graph is centered close to the population parameter, p = 0.15. As for the spread, it's similar to the standard deviation on the previous page, where n = 50 and p = 0.45. If we increase the sample size to n = 50, the sampling distribution of $\hat{p}$ should show less variability. The standard deviation below on the right confirms this. Note that we can't just visually compare the graphs because the horizontal scales are different. The dotplot is more symmetrical than the graph in the figure below on the left and is once again centered at a value that is close to p = 0.15.



(a)

Mean = 0.143 Std Dev = 0.069

$\hat{p}$ = 0.08

Current Sample: 400

π: 0.15

sample size: 25

number of samples: 400

□ Count Samples...

□ Plot Normal Curve

□ Animate

(b)

Mean = 0.148 Std Dev = 0.051

$\hat{p}$ = 0.14

Current Sample: 400

π: 0.15

sample size: 50

number of samples: 400

□ Count Samples...

□ Plot Normal Curve

□ Animate

What have we learned so far about the sampling distribution of $\hat{p}$

Shape: In some cases, the sampling distribution of $\hat{p}$ can be approximated by a Normal curve. This seems to depend on both the sample size n and the population proportion p.

Center: The mean of the sampling distribution is $\mu_{\hat{p}} = p$. This makes sense because the sample proportion $\hat{p}$ is an *unbiased estimator* of p.

Spread: For a specific value of p, the standard deviation $\sigma_{\hat{p}}$ gets smaller as n gets larger. The value of $\sigma_{\hat{p}}$ depends on both n and p. To sort out the details of shape and spread, we need to make an important connection between the sample proportion $\hat{p}$ and the number of "successes" X in the sample.

There is an important connection between the sample proportion $\hat{p}$ and the number of "successes" X in the sample.

$$\hat{p} = \frac{\text{count of successes in sample}}{\text{size of sample}} = \frac{X}{n}$$

Here's a summary of the important facts about the sampling distribution of $\hat{p}$.

Sampling Distribution of a Sample Proportion
Choose an SRS of size n from a population of size N with proportion p of successes. Let $\hat{p}$ be the sample proportion of successes. Then:

The **mean** of the sampling distribution of $\hat{p}$ is $\mu_{\hat{p}} = p$.
$\hat{p}$ is an unbiased estimator of p.

The standard deviation of the sampling distribution of $\hat{p}$ is $\sigma_{\hat{p}} = \sqrt{\dfrac{p(1-p)}{n}}$ as long as the *10% condition* is satisfied: $n \leq 0.10N$. As sample size increases, the spread decreases.

As n increases, the sampling distribution of $\hat{p}$ becomes approximately Normal. Before you perform Normal calculations, check that the *Large Counts condition* is satisfied: $np \geq 10$ and $n(1 - p) \geq 10$.

Using the Normal Approximation for $\hat{p}$

Inference about a population proportion p is based on the sampling distribution of $\hat{p}$. When the sample size is large enough for np and $n(1-p)$ to both be at least 10 (the Large Counts condition), the sampling distribution of $\hat{p}$ is approximately Normal.

In that case, we can use a Normal distribution to calculate the probability of obtaining an SRS in which lies in a specified interval of values.

Example 2: A polling organization asks an SRS of 1500 first-year college students how far away their home is. Suppose that 35% of all first-year students attend college within 50 miles of home. Find the probability that the random sample of 1500 students will give a result within 2 percentage points of this true value. Show your work.

Step 1: State: We want to find $P(0.33 \leq \hat{p} \leq 0.37)$.

We have an SRS of size n = 1500

p = 0.35 is the true proportion of first year college students who attend college within 50 miles of home.

mean is $\mu_{\hat{p}} = p = 0.35.$

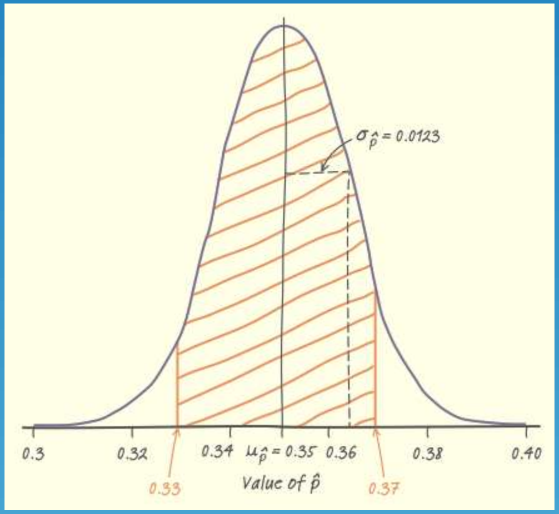
We need to check the 10% condition, to use the standard deviation formula. The population must contain at least 10(1500) = 15,000 people, it is safe to assume that this is true, so

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.35)(0.65)}{1500}} = 0.0123$$

Check the Large Counts condition:

np **= 1500(0.35) = 525 and *n*(1 – *p*) = 1500(0.65) = 975.**

Both are much larger than 10, so the Normal approximation will be quite accurate.



Step 2: Perform calculations—show your work!

$$P(0.33 \le \hat{p} \le 0.37) = \text{ normalcdf}\left(\text{lower: } 0.33, \text{upper: } 0.37, \mu : 0.35, \sigma : 0.0123\right)$$

$$= 0.8961$$

Step 3: Answer the question: About 90% of all SRSs of size 1500 will give a result within 2 percentage points of the truth about the population.