Sampling for Proportions

Previously, we assumed that we knew that a certain percentage of the individuals in a population had a certain characteristic.

Sampling for Proportions

Previously, we assumed that we knew that a certain percentage of the individuals in a population had a certain characteristic.

Possible Examples:

We know that 7% of the population is left handed.

Sampling for Proportions

Previously, we assumed that we knew that a certain percentage of the individuals in a population had a certain characteristic.

Possible Examples:

We know that 7% of the population is left handed.

We know that 35% of the population has Type O blood.

Sampling for Proportions

Previously, we assumed that we knew that a certain percentage of the individuals in a population had a certain characteristic.

Possible Examples:

We know that 7% of the population is left handed.

We know that 35% of the population has Type O blood.

We know that 4% of the population is unemployed.

Sampling for Proportions

Previously, we assumed that we knew that a certain percentage of the individuals in a population had a certain characteristic.

Possible Examples:

We know that 7% of the population is left handed.

We know that 35% of the population has Type O blood.

We know that 4% of the population is unemployed.

We know that 23% of the population is over 50.

Sampling for Proportions

Previously, we assumed that we knew that a certain percentage of the individuals in a population had a certain characteristic.

Possible Examples:

We know that 7% of the population is left handed.

We know that 35% of the population has Type O blood.

We know that 4% of the population is unemployed.

We know that 23% of the population is over 50.

We know that 94% of households own a car.

Sampling for Proportions

Previously, we assumed that we knew that a certain percentage of the individuals in a population had a certain characteristic.

Possible Examples:

We know that 7% of the population is left handed.

We know that 35% of the population has Type O blood.

We know that 4% of the population is unemployed.

We know that 23% of the population is over 50.

We know that 94% of households own a car.

We know that 2% of the population is homeless.

Sampling for Proportions

We examined in the composition of *samples* from such a population.

Sampling for Proportions

We examined in the composition of *samples* from such a population.

The word *samples* can mean different things, but in this case we mean a **simple random sample**

Sampling for Proportions

We examined in the composition of *samples* from such a population.

The word *samples* can mean different things, but in this case we mean a **simple random sample**

By a *simple random sample*, we mean a sample chosen in a way that:

- Every member of the population an equal chance of being chosen
- The sample is drawn *without* replacement

Sampling for Proportions

The act of selecting a single individual from a population at random can be though of as an experiment with two possible outcomes:

- The individual has the specified characteristic
- The individual does *not* have the specified characteristic

Sampling for Proportions

The act of selecting a single individual from a population at random can be though of as an experiment with two possible outcomes:

- The individual has the specified characteristic
- The individual does *not* have the specified characteristic

This idealized experiment is called a *Bernoulli trial*

Sampling for Proportions

Recall that the Bernoulli distribution associates one of two values, zero or one, with the outcome of the experiment.

Sampling for Proportions

Recall that the Bernoulli distribution associates one of two values, zero or one, with the outcome of the experiment.

Assume p is the proportion of time we expect an outcome of one.

Then the experiment produces:

- An outcome of 1 with probability: p
- An outcome of 0 with probability: $1 - p$

Computing the Standard Error

For a single Bernoulli trial, consider the outcome X to be a random variable that:

- assumes the value 1 with probability p
- assumes the value 0 with probability $1 - p$

Computing the Standard Error

For a single Bernoulli trial, consider the outcome X to be a random variable that:

- assumes the value 1 with probability p
- assumes the value 0 with probability $1 - p$

The *standard deviation* of the random variable X is given by the formula:

$$SD_x = \sqrt{p(1-p)}$$

Computing the Standard Error

If we conduct a series of n Bernoulli trials, each with probability p of outcome 1, then the proportion of outcomes in the sample having outcome 1 is just the average of the X values over the n replications of the experiment:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Computing the Standard Error

If we conduct a series of n Bernoulli trials, each with probability p of outcome 1, then the proportion of outcomes in the sample having outcome 1 is just the average of the X values over the n replications of the experiment:

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Recall that we obtain the standard error of the mean of a sample of size n by dividing the population standard deviation by the square root of n:

$$SD_{\overline{x}} = \frac{SD_x}{\sqrt{n}}$$

Computing the Standard Error

The population standard deviation in this case is:

$$SD_x = \sqrt{p(1-p)}$$

Computing the Standard Error

The population standard deviation in this case is:

$$SD_x = \sqrt{p(1-p)}$$

So the standard error of the sample mean (which is just the sample proportion) is:

$$SD_{\overline{x}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

Computing the Standard Error

The population standard deviation in this case is:

$$SD_x = \sqrt{p(1-p)}$$

So the standard error of the sample mean (which is just the sample proportion) is:

$$SD_{\overline{x}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

It is important to note that p is the *population* proportion, that is, the proportion of individuals in the population that have the specified characteristic.

Computing the Standard Error

The population standard deviation in this case is:

$$SD_x = \sqrt{p(1-p)}$$

So the standard error of the sample mean (which is just the sample proportion) is:

$$SD_{\overline{x}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$$

It is important to note that p is the *population* proportion, that is, the proportion of individuals in the population that have the specified characteristic.

The standard error of the sample mean *does not* depend on the sample proportion.

The Normal Approximation

The expected value of the sample mean is the population proportion, p.

The Normal Approximation

The expected value of the sample mean is the population proportion, p.

The distribution of the sample mean is *approximately* a bell curve with:

$$\text{mean} = p$$

and

$$\text{standard deviation} = \sqrt{\frac{p(1-p)}{n}}$$

The case where p is unknown

In practice, more often than not we *do not* know the actual proportion p of individuals that have the characteristic.

The case where p is unknown

In practice, more often than not we *do not* know the actual proportion p of individuals that have the characteristic.

Also, in practice it is usually not possible to examine every member of the population to determine the actual proportion.

The case where p is unknown

In practice, more often than not we *do not* know the actual proportion p of individuals that have the characteristic.

Also, in practice it is usually not possible to examine every member of the population to determine the actual proportion.

Often it is difficult or impossible to *identify* every member of the population.

The case where p is unknown

In practice, more often than not we *do not* know the actual proportion p of individuals that have the characteristic.

Also, in practice it is usually not possible to examine every member of the population to determine the actual proportion.

Often it is difficult or impossible to *identify* every member of the population.

Even if every member of the population can be identified, it is usually impractical to contact all of them and determine whether they have the characteristic.

Confidence Intervals

For these reasons, usually the only practical solution is sampling.

Confidence Intervals

For these reasons, usually the only practical solution is sampling.

As before, we collect a sample of, say 100 individuals, and determine the proportion p that has the characteristic we are interested in.

Confidence Intervals

For these reasons, usually the only practical solution is sampling.

As before, we collect a sample of, say 100 individuals, and determine the proportion p that has the characteristic we are interested in.

Usually, the sample proportion p will differ from the actual population proportion, so there will be uncertainty as to how accurate p is as an estimate.

Confidence Intervals

For these reasons, usually the only practical solution is sampling.

As before, we collect a sample of, say 100 individuals, and determine the proportion p that has the characteristic we are interested in.

Usually, the sample proportion p will differ from the actual population proportion, so there will be uncertainty as to how accurate p is as an estimate.

Depending on the sample size and the actual proportion in the population, p may be a very precise estimate, or very imprecise.

Confidence Intervals

A **confidence interval** is a way of expressing the precision of an estimate.

Confidence Intervals

A **confidence interval** is a way of expressing the precision of an estimate.

The idea is to construct the interval so that a statement like the following can be made:

If we repeated the experiment of drawing this sample over and over, X percent of the time the confidence interval will contain the true population proportion.

Confidence Intervals

A **confidence interval** is a way of expressing the precision of an estimate.

The idea is to construct the interval so that a statement like the following can be made:

If we repeated the experiment of drawing this sample over and over, X percent of the time the confidence interval will contain the true population proportion.

It is important to realize what we are *not* saying here

Confidence Intervals

A **confidence interval** is a way of expressing the precision of an estimate.

The idea is to construct the interval so that a statement like the following can be made:

If we repeated the experiment of drawing this sample over and over, X percent of the time the confidence interval will contain the true population proportion.

It is important to realize what we are *not* saying here

We are *not* saying "There is an X percent chance that the population parameter is in the interval"

Confidence Intervals

This is a rather subtle, but important idea.

Confidence Intervals

This is a rather subtle, but important idea.

The philosophy is that the true population proportion is a fixed quantity, not a random variable.

Confidence Intervals

This is a rather subtle, but important idea.

The philosophy is that the true population proportion is a fixed quantity, not a random variable.

As such, it is what it is, and the idea of the probability that it falls in a certain range is meaningless.

Confidence Intervals

This is a rather subtle, but important idea.

The philosophy is that the true population proportion is a fixed quantity, not a random variable.

As such, it is what it is, and the idea of the probability that it falls in a certain range is meaningless.

The *sample* proportions *are* random quantities, and so is the confidence interval, and it does make sense to talk about the probability that the confidence interval contains the true proportion.

Review: The case where p is Known

Recall that

Review: The case where p is Known

Recall that

The distribution of the sample mean is *approximately* a bell curve with:

$$\text{mean} = p$$

and

$$\text{standard deviation} = \sqrt{\frac{p(1-p)}{n}}$$

The case where p is Unknown

The computations are the same, but this time p is the *sample* proportion.

The case where p is Unknown

The computations are the same, but this time p is the *sample* proportion.

We *estimate* the standard deviation of the sample proportion as

$$\text{standard deviation} = \sqrt{\frac{p(1-p)}{n}}$$

The Confidence Interval for p

We compute the *sample* proportion p.

The Confidence Interval for p

We compute the *sample* proportion p.

Next compute the *sample* standard deviation of the sample proportion as

$$\text{sample standard deviation} = \sqrt{\frac{p(1-p)}{n}} = s_p$$

The Confidence Interval for p

We compute the *sample* proportion p.

Next compute the *sample* standard deviation of the sample proportion as

$$\text{sample standard deviation} = \sqrt{\frac{p(1-p)}{n}} = s_p$$

For a 95% confidence interval, using the normal approximation, the upper and lower limits are:

$$p \pm 1.96 s_p$$