Sampling for Proportions

Suppose we know that a certain percentage of the individuals in a population have a certain characteristic.

Sampling for Proportions

Suppose we know that a certain percentage of the individuals in a population have a certain characteristic.

Possible Examples:

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

Sampling for Proportions

Suppose we know that a certain percentage of the individuals in a population have 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

Suppose we know that a certain percentage of the individuals in a population have 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

Suppose we know that a certain percentage of the individuals in a population have 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

Suppose we know that a certain percentage of the individuals in a population have 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

Suppose we know that a certain percentage of the individuals in a population have 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 are interested in the composition of *samples* from such a population.

Sampling for Proportions

We are interested 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 are interested 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: $1 - p$

Sampling for Proportions

If our sample is drawn *with* replacement, the number of individuals in the population that have the specified characteristic remains constant.

Sampling for Proportions

If our sample is drawn *with* replacement, the number of individuals in the population that have the specified characteristic remains constant.

Consequently, a random sample of size n drawn with replacement is exactly like n Bernoulli trials.

Sampling for Proportions

If our sample is drawn *with* replacement, the number of individuals in the population that have the specified characteristic remains constant.

Consequently, a random sample of size n drawn with replacement is exactly like n Bernoulli trials.

In each trial, the probability of an outcome of 1 is equal to the proportion p of the population that has the characteristic.

Sampling for Proportions

If our sample is drawn *with* replacement, the number of individuals in the population that have the specified characteristic remains constant.

Consequently, a random sample of size n drawn with replacement is exactly like n Bernoulli trials.

In each trial, the probability of an outcome of 1 is equal to the proportion p of the population that has the characteristic.

The probability of an outcome of 0 is equal to $1 - p$, one minus the proportion of the population that has the characteristic.

Sampling for Proportions

Usually we assume the sample is drawn *without* replacement.

Sampling for Proportions

Usually we assume the sample is drawn *without* replacement.

Strictly speaking, this does not result in a sequence of n Bernoulli trials having outcome 1 with probability p.

Sampling for Proportions

Usually we assume the sample is drawn *without* replacement.

Strictly speaking, this does not result in a sequence of n Bernoulli trials having outcome 1 with probability p.

The reason for this is that every time we draw an individual for the sample, they are no longer eligible to be chosen on subsequent draws.

Sampling for Proportions

Usually we assume the sample is drawn *without* replacement.

Strictly speaking, this does not result in a sequence of n Bernoulli trials having outcome 1 with probability p.

The reason for this is that every time we draw an individual for the sample, they are no longer eligible to be chosen on subsequent draws.

This means that they are effectively removed from the population, and this changes the proportion p of individuals that have the characteristic.

Sampling for Proportions

For a very large population, the effect of removing a single individual, or even a sample of n, is considered to be negligible.

Sampling for Proportions

For a very large population, the effect of removing a single individual, or even a sample of n, is considered to be negligible.

The experiment is treated as a sequence of n Bernoulli trials with probability of outcome 1 equal to p.

Sampling for Proportions

For a very large population, the effect of removing a single individual, or even a sample of n, is considered to be negligible.

The experiment is treated as a sequence of n Bernoulli trials with probability of outcome 1 equal to p.

So even though the sampling is usually done without replacement, the results are analyzed as if the sampling was done with replacement.

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 Normal Approximation

Example: A sample of size 100 is drawn from a population in which 30% of the individuals have Type O blood.

What is the expected value and standard deviation of the proportion of this sample that has Type O blood?

The Normal Approximation

Example: A sample of size 100 is drawn from a population in which 30% of the individuals have Type O blood.

What is the expected value and standard deviation of the proportion of this sample that has Type O blood?

Solution:

$$\text{mean} = p = 0.30$$

and

$$\text{standard deviation} = \sqrt{\frac{0.3(1-0.3)}{100}} = 0.04583$$

The Normal Approximation

Example: A sample of size 50 is drawn from a population in which 10% of the individuals are left handed.

What is the expected value and standard deviation of the proportion of this sample that are left handed?

The Normal Approximation

Example: A sample of size 50 is drawn from a population in which 10% of the individuals are left handed.

What is the expected value and standard deviation of the proportion of this sample that are left handed?

Solution:

$$\text{mean} = p = 0.10$$

and

$$\text{standard deviation} = \sqrt{\frac{0.1(1-0.1)}{50}} = 0.04243$$

The Normal Approximation

Example: The standard deviation formula is $\sqrt{p(1-p)}$. If you graph this function with a calculator or computer, you find that for values of p between zero and one, the maximum value occurs at $p = 1/2$.

If I draw a population where the proportion of individuals having a characteristic is unknown, what is the *largest possible* standard deviation for the sample proportion if the sample size is 80?

The Normal Approximation

Example: The standard deviation formula is $\sqrt{p(1-p)}$. If you graph this function with a calculator or computer, you find that for values of p between zero and one, the maximum value occurs at $p = 1/2$.

If I draw a population where the proportion of individuals having a characteristic is unknown, what is the *largest possible* standard deviation for the sample proportion if the sample size is 80?

Solution:

$$\text{when mean} = p = 0.50$$

and

$$\text{standard deviation} = \sqrt{\frac{0.5(1-0.5)}{80}} = 0.0559$$

The Normal Approximation

The formulas

$$\text{when mean} = p$$

and

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

are exactly true for a sample that represents n independent Bernoulli trials.

The Normal Approximation

The formulas

$$\text{when mean} = p$$

and

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

are exactly true for a sample that represents n independent Bernoulli trials.

The assertion that the sample proportion has a normal distribution is an approximation.

The Normal Approximation

The formulas

$$\text{when mean} = p$$

and

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

are exactly true for a sample that represents n independent Bernoulli trials.

The assertion that the sample proportion has a normal distribution is an approximation.

One rule of thumb statisticians use is that the approximation will be reasonably accurate if

$$n \cdot p(1-p) \geq 10$$

The Normal Approximation

Example: In a population 15% of the individuals have a certain characteristic. What is the smallest sample size for which the rule of thumb

$$n \cdot p(1-p) \geq 10$$

holds?

The Normal Approximation

Example: In a population 15% of the individuals have a certain characteristic. What is the smallest sample size for which the rule of thumb

$$n \cdot p(1 - p) \geq 10$$

holds? Solution:

$$n \geq \frac{10}{.15 \cdot .85} = 79$$