Class 12: Sampling Distributions: Proportions (Text: Sections 3.3 and 5.2)

TWO TYPES OF VARIABLE

There are two types of **random variable—categorical and quantitative**. What is the difference between them?

- **Categorical**: Individuals we are studying are in categories. For example, gender, race, has BA. We get a count of the number in each category.
- **Quantitative**: Each individual has their own number. For example, income, weight, height, test score.

Why do we care? Formulas for the sampling distributions of the two variables will be different.

How do we tell which kind of variable we have*?*
Either
—-Look at an individual in the population. Ask yourself if the individual has this/her own number? Or is the individual in a category?
Or
—-Look at a sample. Ask yourself if the variable has a mean or a proportion in the sample

Ex: Is the *underlying* variable quantitative or categorical?
Proportion of men in a sample: Each individual is a man or a women; Categorical
Average income in a sample from a community: Each individual has an income: Quantitative
Fraction of a sample that has finished 12 years of education: Each person has either finished 12 years or not: Categorical
Mean weight of people in a sample from a hospital: Each person has a weight: Quantitative
Notice that even categorical variables give rise to numbers at the sample level: Each sample has a count and a proportion that satisfy the condition.

Standard Notation: When we take a sample and look at values from the sample, we usually write:

	Population Parameter	Sample Statistic
Size	N	n
Mean	μ	$\bar{x}$
Standard Deviation	σ	s
Proportion	p	$\hat{p}$

NOW WE LOOK AT VALUES FROM A SAMPLE:

Goal of This Classes

- To understand the distribution of the proportions we get from samples
- Why? Because when we take a sample, we want to know how to interpret what we get. To do that, we need to know what we expected.

This means we want to know the distribution of $\bar{x}$ and $\hat{p}$. Today we will figure out the distribution of $\hat{p}$; last time, we figured out the distribution of $\bar{x}$.

THE SAMPLING DISTRIBUTION

The distribution of proportions from *all possible samples* of a fixed size is called the ***sampling distribution of proportions***. We will approximate the sampling distribution of proportions by ***simulation***. The histogram we draw will be an approximation because we will not take all samples.

We can also find out about the sampling distribution from **theory**. This information is given by the Central Limit Theorem (CLT). There are two CLTs, one for categorical variables and one for quantitative variables. (We saw the quantitative CLT last time.)

SAMPLING SIMULATION with a FIXED SAMPLE SIZE of 30

Activity:

Using the emailed file *SimulationProportionsCLT.xlsx*,

- Take 20 samples of size 20 each (Fill down the white area)
- Count the proportion of women in each. (Fill down the tan area)
- Count the number of samples in each bin. (Yellow area.)
- We will combine the number of samples for the class to make an approximate distribution.

Conclusion:

Want to observe

- Shape of sampling distribution
- Mean of sampling distribution
- Standard deviation of sampling distribution

THE CENTRAL LIMIT THEOREM FOR PROPORTIONS (CATEGORICAL VARIABLES)

When we take random samples of a fixed size *n* from a population with a population proportion *p*, the sample statistic, $\hat{p}$, has distribution given by the **Central Limit Theorem:**

- Distribution of $\hat{p}$ is approximately normal
- Mean is population proportion, p
- Standard deviation is $\sqrt{\frac{p(1-p)}{n}}$, also called the **standard error**

The sampling distribution is approximately normal, and the approximation gets better as the sample size increases. We assume the sampling distribution is normal for samples of size 30 or greater.

The name "standard error" for the standard deviation of a sampling distribution is used to emphasize the idea that the sample proportions are estimates for the population proportion. The standard deviation of the sampling distribution gives a measure of the error in the estimate.

When we take a real sample to draw conclusions about a population—this is *statistical inference*— we use theory rather that simulation to figure out what to expect, as we will only have one sample and so can't figure out the standard error except from theory.

SAMPLING SIMULATION with a FIXED SAMPLE SIZE of 30

Results from theory to compare *SimulationProportionsCLT.xlsx*

Ex: What is the population parameter, *p*? What are the sample statistics, $\hat{p}$?

Population parameter is proportion F in population $p = 0.399 = 39.9\%$;
Sample statistics are $\hat{p}$, the proportions of females in the samples.

Sampling Distribution of Proportions:

- Shape should be approximately normal
- Mean should be 39.9%
- Standard error is

$$SE = \sqrt{\frac{0.399(1-0.399)}{30}} = 0.0894 = 8.94\%.$$

SAMPLING FROM A POPULATION with 62.9% WOMEN with a VARIETY of SAMPLE SIZES

Ex: We take samples of a fixed size n. What is the population parameter, p? What are the sample statistics, $\hat{p}$?

Population parameter is $p = 62.9\%$. This is constant; it does not vary from sample to sample.
Sample statistics are $\hat{p}$, the values of the proportion of women from each sample. Varies from sample to sample.

How Does the Sampling Distribution Depend on Samples Size?

Intuition about values of $\hat{p}$. What is the effect of a larger sample size?

- Mean of should remain 0.629.
- For larger samples, values close to 62.9% are more likely than those that are far away.
- In other words, with a larger sample, there is less variation.

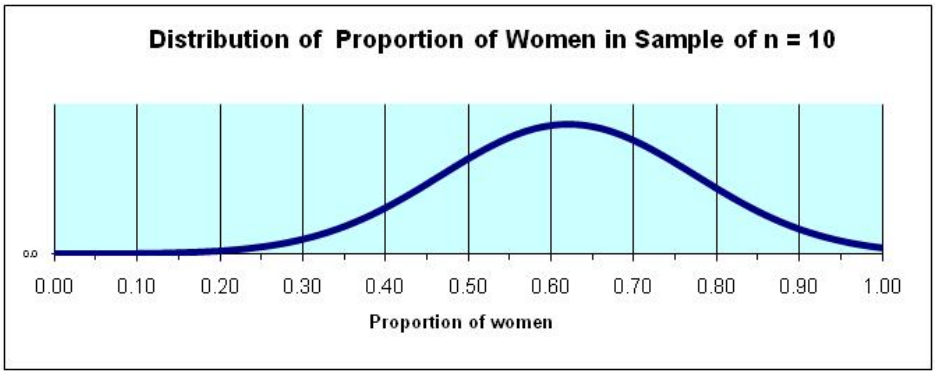
Theory about values of $\hat{p}$. What is the effect of a larger sample size?

CLT says

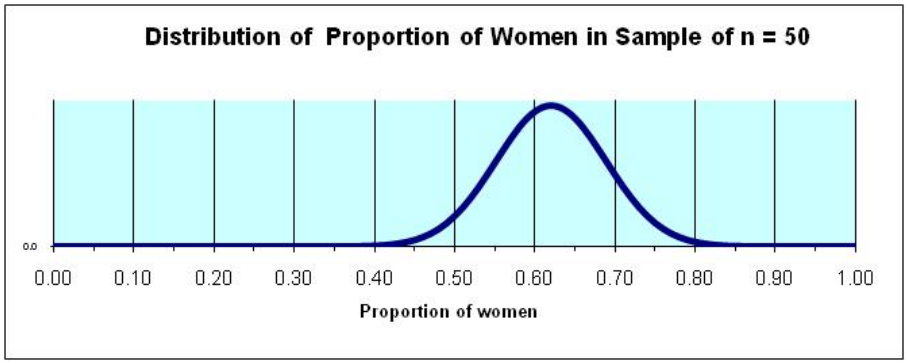
- Mean $= 0.629 = 62.9\%$, does not depend on sample size.
- Standard deviation $= \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.629(1-0.629)}{n}}$, decreases with increasing sample size.

Sampling Distribution for Various Sample Sizes (for $p = 0.629$)

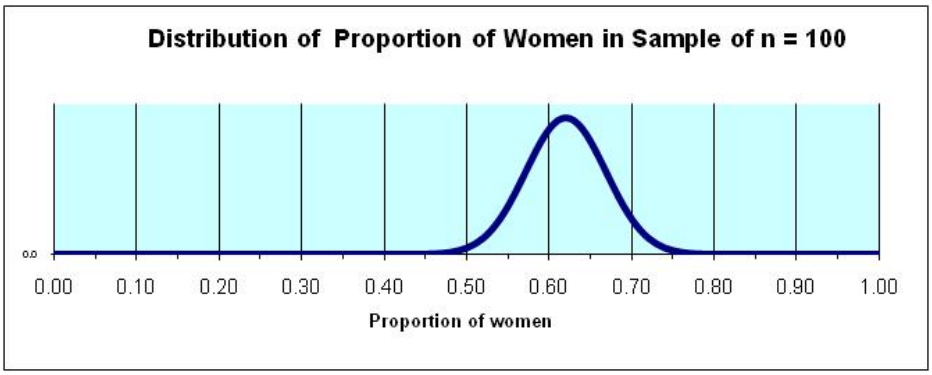
For a sample of size $n = 10$:



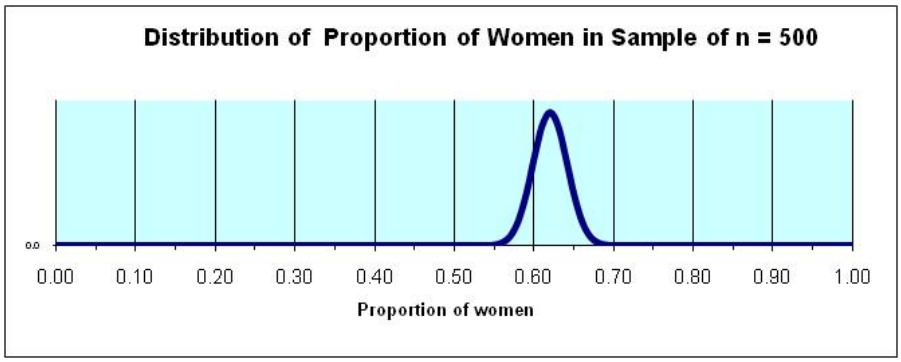
For a sample of size $n = 50$:



For a sample of size $n = 100$:



For a sample of size $n = 500$:



Ex: What is along horizontal axis? $\hat{p}$

Ex: Could all the scales on the vertical axes be the same? Are all the distributions the same height?
Because the total area under each one has to be 1, as the distribution becomes narrower, it becomes taller. As the sample size gets larger, more sample statistics are close to the mean, so pdf graph is higher. Thus the vertical scales on the four graphs cannot be the same.

Notice that:

- The sampling distributions look approximately normal.
  Values far above or far below 62.9% are unlikely; those close to 62.9% are more likely.
- Mean is always 62.9%.
  On average we expect to have 62.9% women.
- The sampling distribution gets narrower as the sample size gets bigger.
  The larger the sample, the less likely there are to be big fluctuations in $\hat{p}$.

From Central Limit Theorem

Ex: For samples from a population containing 62.9% women, what does CLT tell us about the sampling distribution for the proportion of women in a sample of size n = 10? Size n = 100? Size n = 1000? What are the means and standard errors of these distributions?

For n = 10: Normal with Mean = 62.9% and

Standard error $= \sqrt{\frac{0.629(1-0.629)}{10}} = 0.153 = 15.3\%.$

For n = 50: Normal with Mean = 62.9% and

Standard error $= \sqrt{\frac{0.629(1-0.629)}{50}} = 0.069 = 6.9\%.$

For n = 100: Normal with Mean = 62% and

Standard error $= \sqrt{\frac{0.62(91-0.629)}{100}} = 0.049 = 4.9\%.$

For n = 500: Normal with Mean = 62.9% and

Standard error $= \sqrt{\frac{0.629(1-0.629)}{500}} = 0.022 = 2.2\%.$

For n = 1000: Normal with Mean = 62.9% and

Standard error $= \sqrt{\frac{0.629(1-0.629)}{1000}} = 0.015 = 1.5\%$---smaller than before!

Notice that mean is constant and standard deviation shrinks—-Distributions is always centered on 62.9% but gets "more scrunched" and the sample size increases.

Ex: Compare the standard errors for samples of $n = 100$ and of $n = 400$. (Population proportion is 62.9%)

For $n = 100$: Keeping more decimal places: Standard error $= \sqrt{\frac{0.629(1-0.629)}{100}} = 0.048 = 4.8\%.$

For $n = 400$: Standard error $= \sqrt{\frac{0.629(1-0.629)}{400}} = 0.024 = 2.4\%.$

Notice that the standard error for $n = 100$ is twice that for $n = 400$.

- To **halve** the standard error, you have to **quadruple** the sample size.

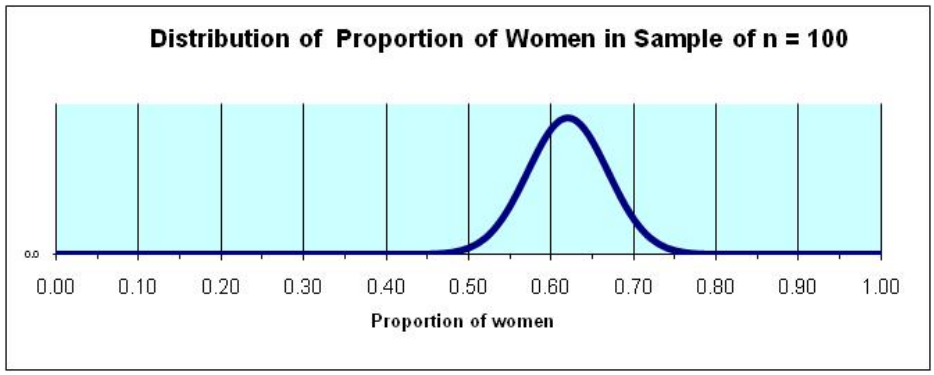
Ex: Show that quadrupling the sample size halves the standard error for any sample size.
If the original sample size is n, then quadrupling it gives $4n$. Then

$$\text{New } SE = \sqrt{\frac{0.629(1-0.629)}{4n}} = \sqrt{\frac{1}{4n}} \cdot \sqrt{\frac{0.629(1-0.629)}{4n}} = \frac{1}{2}\sqrt{\frac{0.629(1-0.629)}{n}} = \frac{1}{2} \cdot \text{Old } SE$$

Calculating Probabilities from a Sampling Distribution

Intuition for $n = 100$

The sampling distribution of $\hat{p}$ shows what values of $\hat{p}$ are possible, and which are likely and which are unlikely. Smoothed out, the sampling distribution looks roughly like this:



Ex: How likely is it that we see 60 or less in a sample of 100? Then $\hat{p} = 0.6$; very likely.

Ex: How likely is it that we see 20 or less in a sample of 100? Then $\hat{p} = 0.2$; very unlikely.

Ex: How likely is it that we see 75 or more in a sample of 100? Then $\hat{p} = 0.75$. Not likely; but not impossible.

Calculation for $n = 100$

The standard deviation (standard error) is 0.049.

Ex: How likely is it that we see 60 or less in a sample of 100?
Then $\hat{p} = 0.6$.so

$$z = \frac{0.6 - 0.629}{0.049} = -0.59$$

The probability is $P(Z < -0.59) = 0.2776 = 27.76\%$.

Ex: How likely is it that we see 20 or less in a sample of 100?
Then $\hat{p} = 0.2$.so

$$z = \frac{0.2 - 0.629}{0.049} = -8.8$$

Since -8.8 is off the table, the probability is $P(Z < -8.8) \approx 0$. Very unlikely to happen.

Ex: How likely is it that we see 75 or more in a sample of 100?
Then $\hat{p} = 0.75$.so

$$z = \frac{0.75 - 0.629}{0.049} = 2.47$$

The probability is $P(Z > 2.47) = 1 - 0.9932 = 0.068 = 0.68\%$. Not likely, but not impossible.

Optional Background: MEAN and STANDARD DEVIATION of the SAMPLING DISTRIBUTION

Mean and Standard Deviation from Theory

Mean comes from $\mu_{a+bx} = E(a + bX) = a + bE(X)$
Standard deviation comes from the variance $\sigma^2_{a+by} = V(a + bY) = b^2V(Y)$

The proportion, $\hat{p}$, is obtained from the count by the formula

$$\hat{p} = \frac{X}{n}$$

The distribution of the count is called binomial. It has mean $E(X) = np$ and the variance $V(X) = np(1 - p)$. From this we can obtain the mean, variance and standard deviation of $\hat{p}$.

The proportion, $\hat{p}$, is ***not*** binomially distributed, but its mean, variance, and standard deviation are given by

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{E(X)}{n} = \frac{np}{n} = p$$

$$V(\hat{p}) = V\left(\frac{X}{n}\right) = \frac{E(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

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