Chapter 9: Confidence Intervals for Population Proportions

1 Terminologies:

- A **population** is the entire group of people or things you wish to study.
  - Example : The population of all NBA players, the population of all high school aged students in America, the population of all pregnant women in America.
- A **parameter** is a numerical value that characterizes some aspect of the population.
  - Example : I want to know what percentage of all NBA players are over 6ft 5. I want to know what percentage of all high school students in America engaged in drug use, etc.
- A **sample** is a small collection of randomly selected people or things taken from the population of interest.
- Once we have a sample, a **statistic** is a numerical value that characterizes some aspect of the sample.

To help us understand how all four terms are applied, let's see an example: I am interested in the **population** of all senior citizen in the U.S. The proportion of all senior citizens that contracted COVID-19 is my **parameter**. My method for reporting will be to survey a **sample** of 100 randomly selected senior citizens. The proportion individuals from the sample that confirm they've had COVID-19 is my **statistic**.

We can make **statistical inferences** - meaning we can draw conclusions - about a population based on the trend we see from a small sample of the population.

Some things to note:

- A **statistic** is knowable because we can always find a statistic by studying a sample.
- A **parameter** is harder to know because it requires us to have data for the entire population which is difficult to collect. Therefore, a parameter is typically unknown.

Important Symbols

Now that we have introduced populations and samples, we should cover the symbols we use to describe them. Keep in mind, you have seen most of these before, but a table is useful to help you keep track of them all.

	Proportions	Means	Standard Deviation
Statistic	$\hat{p}$	$\bar{x}$	s
Parameter	p	μ	σ

Example 1: A researcher wants to estimate the average height of all women aged 20 years or older. From a simple random sample of 45 women, the researcher obtains a sample mean height of 63.9 inches. What is the population and what is the sample?

Population is all women aged 20 years or older while sample is the 45 women that are selected for the study.

Example 2: Researchers claim that 70% of all children, 10 years and younger, consume too much sodium. To study this phenomenon, a nutritionist randomly collects a sample of 75 children 10 and under. The nutritionist finds from their sample, 68% of children consumed too much sodium. In fact, from the sample the mean amount of sodium consumed by the children was 2993 milligrams with a standard deviation of 100 milligrams. What is the population and what is the sample? Identify the proper notation for all variables.

Population is all children, 10 years and younger whereas sample is 75 children who were selected who were 10 and under.
Population proportion (p) = 70%
Sample proportion ($\hat{p}$) = 68%
Sample mean ($\bar{x}$) = 2993 mg
Sample standard deviation (s) = 100 mg

Example 3: The university publishes a report that average GPA of all students at Montclair State University is 3.52 with standard deviation of 0.61. A student decides to do a research project to check the hypothesis that average GPA of all students at MSU is indeed what the university claims. She collects the random sample of 250 students and find that average GPA is 3.60 with a standard deviation of 0.55. What is the population and what is the sample? Identify the proper notation for all variables.

Population is all students at MSU, and sample is 250 students selected.
Population mean (μ) = 3.52
Population standard deviation (σ) = 0.61
Sample mean ($\bar{x}$) = 3.60
Sample standard deviation (s) = 0.55

2 Measuring the Bias in a Survey

No sample can be perfect, so we like to measure how much "bias" there is in our sample. To do this, we refer to a measurement called the **standard error**. The standard error tells us how much the population mean is likely to differ from the sample mean. In other words, standard error estimates the variability across multiple samples of the same population.

That being said, the **standard error** for the **population proportion, p** is :

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

 © 2024 Nishanta Baral. All rights reserved.

such that n = population size and $p = \frac{success}{total}$.

The reality is, we typically don't know the true value of the population proportion, p. This means we can't actually calculate the standard error. However, we can come pretty close by computing the standard error of the sample proportion which is a nearly identical formula!

The **standard error** for the **sample proportion,** $\hat{p}$ is :

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

such that n = sample size and $\hat{p} = \frac{success}{total}$.

3 Central Limit Theorem for sample proprotions

The **Central Limit Theorem** (CLT) that applies to estimating proportions in a population tells us that if the following conditions are met, then the sampling distribution of the sample proportion is close to the Normal distribution.

- 1. The sample is **random and independent**.
  - This means we want to randomly select our sample, and be sure that the selections are independent from each other.
- 2. We have a **large sample**.
  - This means we should compute $n \times p$ and $n \times (1-p)$ and make sure that both computations are greater than or equal to 10, to prove that we have enough successes and failures within our sample.
- 3. Our sample comes from a **big population**.
  - This means we should just be sure that the entire population is at least 10 times bigger than our sample. This is typically pretty trivial to prove and doesn't often require computation.

Example: According to the Pew Research Center, as of 2021 37.9% of people in America hold a bachelor's degree. I want to see how this this proportion represents my state of N.J. I take a random sample of 80 people from N.J. and 28 of them have a bachelor's degree. See if the conditions for central limit theorem are met and compute the standard error.

Given that the proportion of people in America with a bachelor's degree $p = 0.379$ and a sample size $n = 80$, where 28 out of the 80 individuals in New Jersey have a bachelor's degree, we first check the conditions for the Central Limit Theorem:

a. Random Sampling: Assuming the sample is selected randomly.
b. Independence: With a sample size of 80 , which is less than 10% of the total population of New Jersey, the condition for independence seems to be met.
c. Success-Failure Condition:

$$np = 80 \times 0.379 = 30.32$$
$$n(1-p) = 80 \times (1 - 0.379) = 49.68$$

Both np and $n(1-p)$ are greater than 10, thus satisfying the condition.

 © 2024 Nishanta Baral. All rights reserved.

Now, calculating the standard error (SE) for the sample proportion:
The formula for the standard error of a sample proportion is:

$$SE = \sqrt{\frac{p \times (1-p)}{n}}$$

Substituting the values:

$$SE = \sqrt{\frac{0.379 \times (1-0.379)}{80}}$$
$$SE = \sqrt{\frac{0.379 \times 0.621}{80}}$$
$$SE = \sqrt{\frac{0.235259}{80}}$$
$$SE \approx \sqrt{0.002941}$$
$$SE \approx 0.0542$$

Hence, the standard error for the sample proportion is approximately 0.0542 .

4 Confidence Intervals

As we discussed earlier, it is really difficult and pretty unrealistic to find the parameter values for an entire population. Instead, we look at samples within the population to help us represent the population. Now, keep in mind that samples have some variability because we can pick many different samples from the same population; they will differ slightly from each other.

We can improve upon our computation for standard error by considering a level of confidence. We can better capture the population parameter given this typical unreliability from sample to sample by using confidence intervals. A confidence interval refers to the probability that the population parameter will fall between two values. To provide some context, this might look something like this: You took your car to a dealership to trade it in. They value your car somewhere between \$40K and \$42K. It might look something like: I am confident that the value of the car is between \$40K and \$42K.

So far, we have a sample and a sample proportion. We check if it satisfies the Central Limit Theorem (CLT).

- Random and Independent
- Enough sucess and failures within a sample $np \geq 10$ and $n(1-p) \geq 10$
- Population size is at least 10 times greater than the sample.

Since we always don't have value of p, to estimate the standard error of our sample proportion in a normally distributed sampling distribution, we substitute p with $\hat{p}$ to get a close approximation:

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

 © 2024 Nishanta Baral. All rights reserved.

5 Estimating the Population Proportion with Confidence Intervals

To illustrate this concept, let's look at an example. Suppose a survey was conducted with 500 randomly selected residents of New York City. In this sample, 225 respondents i.e. $\frac{225}{500} = 45\%$ said they prefer using public transportation over driving a car. This percentage is based on our sample, but what if we want to know the proportion of all city residents who prefer public transportation? Could it be higher or lower than 45%?

We don't know the exact population proportion (p), but we have our sample proportion ($\hat{p}$) of 0.45. Here are some key points:

1. Our sample proportion is an unbiased estimator, meaning it should closely match the true population proportion, even if not exactly.

2. We can calculate the standard error (SE) to gauge the variability of our estimator. The formula for SE is:

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{0.45 \times 0.55}{500}} \approx 0.022, \text{ or about } 2.2\%$$

   Since our estimator is unbiased, the SE helps us understand how much our sample proportion might deviate from the true population proportion.

3. With a large sample size, the distribution of our sample proportion is approximately normal. This means there's a 68% chance that our sample proportion is within one standard error of the population proportion, a 95% chance it's within two standard errors, and nearly a 100% chance it's within three standard errors. For our example, three standard errors are $3 \times 2.2\% = 6.6\%$, so we can confidently say that the true proportion is within 6.6 percentage points of 45%.

Therefore, we can be quite certain that the true population proportion falls between:

$$45\% - 6.6\% \text{ and } 45\% + 6.6\%, \text{ or } 38.4\% \text{ to } 51.6\%.$$

In summary, what we've done is estimating a confidence interval around our sample proportion, indicating where we believe the true population proportion lies.

5.1 Margin of Error

We select a margin of error that will produce the desired confidence level. Margin of Error is calculated as:

$$\text{Margin of error} = z^* \times SE$$

In this formula, z^* is a multiplier that determines the number of standard errors to include in the margin of error. For example, if $z^* = 1$, the confidence level is 68%. If $z^* = 2$, the confidence level is 95%. Table 7.6 below summarizes the margin of error for common confidence levels.

$$MOE = z^* \cdot SE_{EST}$$

Where z^* is the critical value.

For example, let's see for a 95% confidence interval. CLT ensures that my sampling distribution is approximately normal with mean of p and standard error of SE.

 © 2024 Nishanta Baral. All rights reserved.

Confidence Level	Margin of Error Is...
99.7%	3.0 standard errors
99%	2.58 standard errors
95%	1.96 standard errors
90%	1.645 standard errors
80%	1.28 standard errors

6 Finding confidence intervals when p is not known

When we are looking for a confidence interval for a population proportion, it has the following structure:

$$\hat{p} \pm m$$

where m is the margin of error. By substituting the margin of error, we can write:

$$\hat{p} \pm z^* \times SE$$

To find the standard error, we need to know the value of p:

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

However, since p is usually unknown, we use the sample proportion $\hat{p}$ to estimate the standard error:

$$SE_{\text{est}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

This gives us a confidence interval that is close to the correct confidence level. Therefore, in practice, we use the following formula to find approximate confidence intervals for a population proportion:

$$\hat{p} \pm z^* \times SE_{\text{est}}$$

where

$$SE_{\text{est}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Confidence Level	Margin of Error Is...
99.7%	3.0 standard errors
99%	2.58 standard errors
95%	1.96 standard errors
90%	1.645 standard errors
80%	1.28 standard errors

where: m is the margin of error, $\hat{p}$ is the sample proportion of successes, n is the sample size, z^* is a multiplier based on the desired confidence level, and SE_{est} is the estimated standard error.

 © 2024 Nishanta Baral. All rights reserved.

Example 1: The article "How Do People View Climate Change?" (Global Times, April 5, 2022) describes a survey of 850 adults where 510 of them indicated that they believe climate change is a serious issue. The survey was conducted by the International Center for Climate Studies and the sample was selected in a way that makes it reasonable to regard the sample as representative of the adult population. Calculate a 95% confidence interval for the proportion of adults who believe climate change is a serious issue.

Check the conditions of Central Limit Theorem:

The three necessary conditions of Central Limit Theorem are met as:

- The sample is random and independent.
- $\hat{p} = \frac{510}{850} \approx 0.60$

  $n\hat{p} = 850(0.60) = 510$ and $n(1-\hat{p}) = 850(1-0.60) = 850(0.40) = 340$ are both greater than or equal to 10.
- The population size (the number of adults) is at least 10 times the sample size.

Calculate the estimated standard error (SE_{est}):

$$SE_{\text{est}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.60)(1-0.60)}{850}} \approx 0.017$$

Calculate the margin of error ($z^* = 1.96$):

$$\text{Margin of error } (m) = z^* SE_{\text{est}} = 1.96 \times 0.017 \approx 0.033$$

Calculate the confidence interval and write down the conclusion:

$$\text{Confidence Interval} = \hat{p} \pm m = 0.60 \pm 0.033 = (0.567, 0.633)$$

We can be 95% confident that the proportion of adults who believe climate change is a serious issue is between 0.567 and 0.633.

Example 2: The article "How Many People Believe in Daily Exercise?" (Health News, March 15, 2023) reported the results of a survey of 950 random and independent adults. Of those surveyed, 720 indicated that they believe daily exercise is essential for good health. Based on these sample data, calculate a 90% confidence interval for the proportion of all adults who believe daily exercise is essential for good health.

Check the conditions of Central Limit Theorem:

The three necessary conditions of Central Limit Theorem are met as:

- The sample is random and independent.
- $\hat{p} = \frac{720}{950} \approx 0.758$

  $n\hat{p} = 950(0.758) \approx 719$ and $n(1-\hat{p}) = 950(1-0.758) \approx 231$ are both greater than or equal to 10.

 © 2024 Nishanta Baral. All rights reserved.

- The population size is at least 10 times the sample size.

Calculate the estimated standard error (SE_{est}):

$$SE_{\text{est}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.758)(1-0.758)}{950}} \approx 0.014$$

Calculate the margin of error ($z^* = 1.645$):

$$\text{Margin of error } (m) = z^* SE_{\text{est}} = 1.645 \times 0.014 \approx 0.023$$

Calculate the confidence interval and write down the conclusion:

$$\text{Confidence Interval} = \hat{p} \pm m = 0.758 \pm 0.023 = (0.735, 0.781)$$

We can be 90% confident that the proportion of all adults who believe daily exercise is essential for good health is between 0.735 and 0.781.

For the same problem, calculate a 95% confidence interval for the proportion of all adults who believe daily exercise is essential for good health.

We need to calculate the new margin of error.

$$\text{Margin of error } (m) = z^* SE_{\text{est}} = 1.96 \times 0.014 \approx 0.027$$

$$\text{Confidence Interval} = \hat{p} \pm m = 0.758 \pm 0.027 = (0.731, 0.785)$$

With the increase in confidence level, does the width of your confidence interval increase or decrease? Compare the width of your 90% and 95% confidence intervals.

As the confidence level increases, the margin of error also increases. Hence, the width of the confidence interval also increases. Among the standard confidence intervals (C.I.), 99% C.I. has the biggest width, followed by 95%, then 90%, and finally 80% being the smallest.

7 Confidence Intervals for Difference in Population Proportions in Independent Sample

To estimate the difference in proportions from two separate populations, for instance, control and treatment groups, Republicans and Democrats, or residents in 2015 with residents in 2013, we use confidence intervals for difference in population proportions. Let p_1 and p_2 be the true population proportions. The two sample statistics will be $\hat{p}_1$ and $\hat{p}_2$. We will use $\hat{p}_1 - \hat{p}_2$ to estimate the true difference in proportions $p_1 - p_2$.

Sampling Distribution for Difference in Proportion:

1. We ensure the sampling distribution is normally distributed for which we need to check the following conditions for Central Limit Theorem:

   (a) Random and Independent. Both samples are *randomly* drawn from their populations, and observations are independent of each other.

 © 2024 Nishanta Baral. All rights reserved.

(b) Large Samples. Both sample sizes are large enough that at least 10 successes and 10 failures can be expected in both samples. Verify yourself!

(c) Big Populations. If the samples are collected without replacement, then both population sizes must be at least 10 times bigger than their samples.

(d) Independent Samples. The samples must be independent of each other. This condition is new!

2. We will have $\hat{p}_1, \hat{p}_2, n_1$, and n_2.

3. The mean of the sampling distribution of the difference in sample proportions $(\hat{p}_1 - \hat{p}_2)$ is given by the difference in the population proportions $(p_1 - p_2)$. When estimating this from sample data, we use $(\hat{p}_1 - \hat{p}_2)$. The standard error is then calculated as:

$$S.E._{EST} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

4. Now we have to calculate the margin of error as $z^* \times SE_{EST}$ and confidence intervals as $(\hat{p}_1 - \hat{p}_2) \pm m$ where m is the margin of error.

Example: We want to know whether there was a difference between the proportion of callbacks for the applications the researchers identified as being perceived as female and the proportion of callbacks for the applications the researchers identified as being perceived as male.

- Of the 2746 applications with female names, 309 received callbacks.
- Of the 1121 applications with male names, 83 received callbacks.

What is the 95% confidence interval for the difference in proportions of applicants who received callbacks between those perceived as female and those perceived as male? Interpret the meaning of confidence interval.

We need to check if our sampling distribution is approximately normal by ensuring the conditions of the Central Limit Theorem are met.

1. **Random and Independent:** Both samples are randomly drawn from their populations, and the observations are independent of each other.

2. **Large Sample Size:** Both sample sizes are large enough that at least 10 successes and 10 failures can be expected in both samples. $n_1\hat{p}_1 \geq 10, n_1(1-\hat{p}_1) \geq 10$ and $n_2\hat{p}_2 \geq 10, n_2(1-\hat{p}_2) \geq 10$.

3. **Population Size:** The population size is at least 10 times larger than the sample size for both samples.

4. **Independent Samples.** The samples are independent of each other.

To find the 95% confidence interval for the difference in proportions of applicants who received callbacks between those perceived as female and those perceived as male, we will use the following steps:

1. Calculate the sample proportions:

$$\hat{p}_f = \frac{309}{2746} \approx 0.1126$$

 © 2024 Nishanta Baral. All rights reserved.

$$\hat{p}_m = \frac{83}{1121} \approx 0.0741$$

2. Calculate the standard error of the difference in proportions:

$$SE = \sqrt{\frac{\hat{p}_f(1-\hat{p}_f)}{n_f} + \frac{\hat{p}_m(1-\hat{p}_m)}{n_m}} = \sqrt{\frac{0.1126 \times (1-0.1126)}{2746} + \frac{0.0741 \times (1-0.0741)}{1121}} \approx 0.0096$$

3. Calculate the margin of error for a 95% confidence interval:

$$ME = z^* \times SE = 1.96 \times 0.0096 \approx 0.0188$$

4. Calculate the confidence interval:

$$(\hat{p}_f - \hat{p}_m) \pm ME = (0.1126 - 0.0741) \pm 0.0188 = 0.0385 \pm 0.0188$$

$$\text{Confidence Interval} = (0.0197, 0.0573)$$

Interpretation: We are 95% confident that the true difference in proportions of applicants who received callbacks between those perceived as female and those perceived as male lies between 0.0197 and 0.0573. Since 0 is not included in this interval, it suggests that there is a statistically significant difference in callback rates between the two groups. Specifically, this suggests that applications perceived as female have a higher callback rate compared to those perceived as male.

Example 2: Example: We want to know whether there was a difference between the proportion of women with bachelor's degrees in the millennial generation and the Gen X generation. In a 2017 survey of 400 randomly selected women of the millennial generation, 36% graduated with a bachelor's degree. Similarly, in a 2017 survey of 400 randomly selected women of the Gen X generation, 40% graduated with a bachelor's degree. We wish to find a 95% confidence interval for the difference in the proportion of women with bachelor's degrees between the millennial and the Gen X generations.

Solution: We need to check if our sampling distribution is approximately normal by ensuring the conditions of the Central Limit Theorem are met. They do meet and you can verify it yourself.

To find the 95% confidence interval for the difference in proportions of women with bachelor's degrees between the millennial and the Gen X generations, we will use the following steps:
1. Calculate the sample proportions:

$$\hat{p}_M = 0.36$$

$$\hat{p}_X = 0.40$$

2. Calculate the standard error of the difference in proportions:

$$SE = \sqrt{\frac{\hat{p}_M(1-\hat{p}_M)}{n_M} + \frac{\hat{p}_X(1-\hat{p}_X)}{n_X}} = \sqrt{\frac{0.36 \times (1-0.36)}{400} + \frac{0.40 \times (1-0.40)}{400}} \approx 0.0346$$

3. Calculate the margin of error for a 95% confidence interval:

$$ME = z^* \times SE = 1.96 \times 0.0346 \approx 0.0678$$

4. Calculate the confidence interval:

$$(\hat{p}_M - \hat{p}_X) \pm ME = (0.36 - 0.40) \pm 0.0678 = -0.04 \pm 0.0678$$

 © 2024 Nishanta Baral. All rights reserved.

$$\text{Confidence Interval} = (-0.1078, 0.0278)$$

Interpretation: We are 95% confident that the true difference in proportions of women with bachelor's degrees between the millennial generation and the Gen X generation lies between -0.1078 and 0.0278. Since 0 is included in this interval, it suggests that there is no statistically significant difference in the proportions of women with bachelor's degrees between the two generations.

 © 2024 Nishanta Baral. All rights reserved.