Point Estimates and Sampling Variability

August 19, 2019

Final Exam Options

Option 1: The final exam is NOT comprehensive.

Option 2: The final exam IS comprehensive, but if you do better on the final than on the midterm, your final exam score will replace your midterm score. The score comparison will be based on raw scores, NOT scores with extra credit included.

Foundations for Inference

- Statistical inference is where we get to take all of the concepts we've learned and use them on our data.
- We want to understand and quantify uncertainty related to parameter estimates.
- The details will vary, but the foundations will carry you far beyond this class.

Foundations for Inference

In this chapter, we will

1. Think about using a sample proportion to estimate a population proportion.
2. Build confidence intervals, or ranges of plausible values for the population parameter.
3. Introduce hypothesis testing, which allows us to formally test some of those research questions we talked about in Chapters 1 and 2.

Point Estimates

- A recent poll suggests Trump's approval rating among US adults is 41%.
- We consider 41% to be a **point estimate** for the true approval rating.
  - The true rating is what we would see if we could get responses from every single adult in the US.
- The response from the entire population is the **parameter** of interest.

Point Estimates

- When the parameter is a proportion, it is often denoted by p.
- The sample proportion is denoted $\hat{p}$ (p-hat).
- Unless we collect responses from every individual in the population, p is unknown.
- We use $\hat{p}$ as our estimate of p.

Error

- The difference between the sample proportion and the parameter is called the **error** in the estimate.
- Error consists of two aspects:
    1. sampling error
    2. bias.

Sampling error

- **Sampling error** is how much an estimate tends to vary between samples.
- This is also referred to as *sampling uncertainty*.
- E.g., in one sample, the estimate might be 1% above the true population value.
- In another sample, the estimate might be 2% below the truth.
- Our goal is often to quantify this error.

Bias

- **Bias** is a *systematic* tendency to over- or under-estimate the population true value.
- E.g., Suppose we were taking a student poll asking about support for a UCR football team.
- Depending on how we phrased the question, we might end up with very different estimates for the proportion of support.
- We try to minimize bias through thoughtful data collection procedures.

Variability of a Point Estimate

- Suppose the true proportion of American adults who support the expansion of solar energy is $p = 0.88$
  - This is our parameter of interest.
- If we took a poll of 1000 American adults, we wouldn't get a perfect estimate.
- Assume the poll is well-written (unbiased) and we have a random sample.

Variability of a Point Estimate

- How close might the sample proportion ($\hat{p}$) be to the true value?
- We can think about this using simulations.
- This is possible because we know the true proportion to be $p = 0.88$.

Variability of a Point Estimate

Here's how we might go about constructing such a simulation:

1. There were about 250 million American adults in 2018. On 250 million pieces of paper, write "support" on 88% of them and "not" on the other 12%.
2. Mix up the pieces of paper and randomly select 1000 pieces to represent our sample of 1000 American adults.
3. Compute the fraction of the sample that say "support".

Variability of a Point Estimate

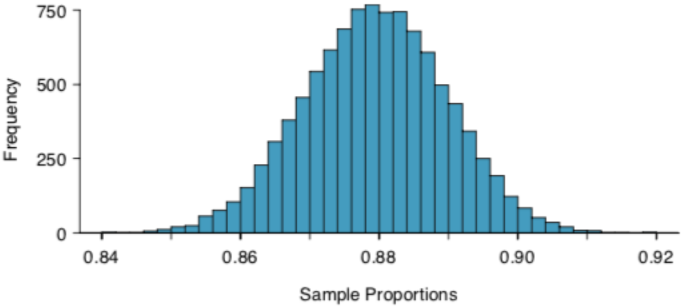
- Obviously we don't want to do this with paper, so we will use a computer.
- Using `R`, we got a point estimate of $\hat{p}_1 = .894$.
- This means that we had an error of $0.894 - 0.88 = +0.014$

Note: the `R` code for this simulation may be found on page 171 of the textbook.

Variability of a Point Estimate

- This code will give a different estimate each time it's run (so the error will change each time).
- Therefore, we need to run multiple simulations.
- In more simulations, we get
  1. $\hat{p}_2 = 0.885$, which has an error of $+0.005$
  2. $\hat{p}_3 = 0.878$ with an error of -0.002
  3. $\hat{p}_4 = 0.859$ with an error of -0.021

Variability of a Point Estimate



The histogram shows the estimates across 10,000 simulations. This distribution of sample proportions is called a **sampling distribution**.

Sampling Distribution

We can characterize this sampling distribution as follows:

Center:

- The center is $\bar{x}_{\hat{p}} = 0.880$, the same as our parameter.
- This means that our estimate is unbiased.
- The simulations mimicked a simple random sample, an approach that helps avoid bias.

We can characterize this sampling distribution as follows:

Spread.

- The standard deviation of the sampling distribution is $s_{\hat{p}} = 0.010$.
- When we're talking about a sampling distribution or the variability of a point estimate, we use the term **standard error** instead of standard deviation.
- Standard error for the sample proportion is denoted $SE_{\hat{p}}$.

Sampling Distribution

We can characterize this sampling distribution as follows:

Shape.

- The distribution is symmetric and bell-shaped - it resembles a normal distribution.

These are all good! When the population proportion is $p = 0.88$ and the sample size is $n = 1000$, the sample proportion $\hat{p}$ is a good estimate *on average*.

Sampling Distribution

Note that the sampling distribution is never observed!

However,

- It is useful to think of a point estimate as coming from a distribution.
- The sampling distribution will help us make sense of the point estimates that we do observe.

Example

What do you think would happen if we had a sample size of 50 instead of 1000?

- Intuitively, more data is better.
- This is true!
- If we have less data, we expect our sampling distribution to have higher variability.
- In fact, the standard error will increase if we decrease sample size.

Central Limit Theorem

- The sampling distribution histogram we saw looked a lot like a normal distribution.
- This is no coincidence!
- This is the result of a principle called the **Central Limit Theorem**.

Central Limit Theorem

When observations are independent and the sample size is sufficiently large, the sample proportion $\hat{p}$ will tend to follow a normal distribution with mean

$$\mu_{\hat{p}} = p$$

and standard error

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

The Success-Failure Condition

In order for the Central Limit Theorem to hold, the sample size is typically considered sufficiently large when

$$np \geq 10$$

and

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

This is called the **success-failure condition**.

Standard Error

Using the standard error, we can see that the variability of a sampling distribution *decreases* as sample size *increases*.

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

Example

Confirm that the Central Limit Theorem applies for our example with $p = 0.88$ and $n = 1000$. Confirm that the Central Limit Theorem applies.

Example

Independence. There are $n = 1000$ observations for each sample proportion $\hat{p}$, and each of those observations are independent draws.

- The most common way for observations to be considered independent is if they are from a simple random sample.
- If a sample is from a seemingly random process, checking independence is more difficult. Use your best judgement.

Example

Success-failure condition.

$$np = 1000 \times 0.88 = 880 \geq 10$$

and

$$n(1-p) = 1000 \times (1 - 0.88) = 120 \geq 10$$

The independence and success-failure conditions are both satisfied, so the Central Limit Theorem applies and it's reasonable to model $\hat{p}$ using a normal distribution.

Example

Compute the theoretical mean and standard error of $\hat{p}$ when $p = 0.88$ and $n = 1000$, according to the Central Limit Theorem.

$$\mu_{\hat{p}} = p = 0.88$$

and

$$SE_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.88 \times (1 - 0.88)}{1000}} = 0.010$$

So $\hat{p}$ is distributed approximately $N(0.88, 0.010)$.

Example

Estimate how frequently the sample proportion $\hat{p}$ should be within 0.02 of the population value, $p = 0.88$.

Example

Within 0.02 of 0.88 is between 0.86 and x0.90. As before, we will find the Z-scores.

$$z_{0.86} = \frac{0.86 - 0.88}{0.010} = -2$$

and

$$z_{0.90} = \frac{0.86 - 0.88}{0.010} = 2$$

Example

Using software,

$$\begin{aligned}P(-2 < Z < 2) &= 1 - P(Z < 2) - P(Z > -2)\\&= P(Z < 2) - P(Z < -2)\\&= 0.977 - 0.023\\&= 0.954\end{aligned}$$

So 95.4% of the proportions should fall within 0.02 of the true population value.

Central Limit Theorem in the Real World

- In a real-world setting, we almost never know the true population proportion.
- However, we use the population proportion to determine whether the Central Limit Theorem is appropriate.
- How do we verify use of the Central Limit Theorem?

Central Limit Theorem in the Real World

Independence. The poll is a simple random sample of American adults, which means that the observations are independent.

Success-failure condition. Without the sample proportion, we use $\hat{p}$ as our next best way to check the success-failure condition.

$$n\hat{p} \geq 10$$

and

$$n(1 - \hat{p}) \geq 10$$

Central Limit Theorem in the Real World

We call this a **substitution approximation** or the *plug-in principle*.

This can also be used to estimate the standard error:

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

This estimate of the standard error tends to be a good approximation of the true standard error.

More About the Central Limit Theorem

What is our conditions don't hold and either

$$np < 10$$

or

$$n(1 - p) < 10?$$

More About the Central Limit Theorem

Let's do another simulation. Suppose $p = 0.25$. Here's a sample of size $n = 10$:

$$\text{no, no, yes, yes, no, no, no, no, no, no}$$

Here, $\hat{p} = 0.2$ for yeses.

More About the CLT

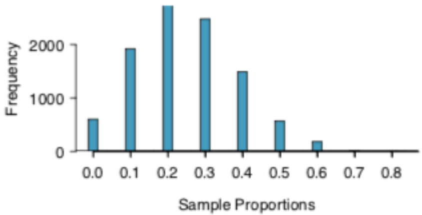
Notice that

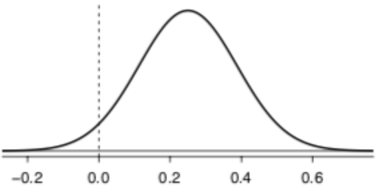
$$np = 10 \times 0.25 = 2.5 < 10$$

The mean and standard deviation for this binomial distribution are 2.5 and 0.137, respectively.

If we simulate many samples with $n = 10$ and $p = 0.25$, what happens to the sampling distribution?

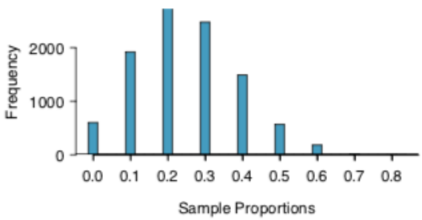
More About the CLT

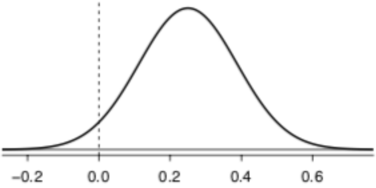




- The histogram shows simulations of $\hat{p}$ for $n = 10$ and $p = 0.25$
- The normal distribution has the same mean (0.25) and standard deviation (0.137).

More About the CLT





- The normal distribution is unimodal, smooth, and symmetric.
- The sampling distribution is unimodal, but it is neither smooth nor symmetric.

More About the CLT

In general, when np or $n(1-p)$ are less than 10,

- The distribution is not continuous.
- The skew is more noteworthy.

When np and $n(1-p)$ are greater than 10,

- The larger both np and $n(1-p)$, the more normal the distribution.

More About the CLT

- The sampling distribution is always centered at the true population proportion p (i.e., $\mu = p$).
- This means that the sample proportion $\hat{p}$ is an *unbiased* estimate of p.
  - This is true as long as the data are independent.

More About the CLT

- The variability decreases as the sample size n increases.
- Remember our formula for standard error!
- Estimates based on a larger sample are intuitively more likely to be accurate.

More About the CLT

- For a particular sample size, the standard error is largest when $p = 0.5$
- This is also reflected in the standard error formula.

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

$p(1-p)$ is maximized at $p = 0.5$.

More About the CLT

- The distribution of $\hat{p}$ will always be discrete.
- However, the normal distribution is still a good *approximation* when the success-failure condition holds.

There are about 25 examples of sampling distributions with different values of n and p on pages 176 and 177 of the textbook.

Extending the Framework

- Using a sample statistic to estimate a parameter is quite common.
- We can apply this to many other statistics (other than proportions).
- The mean is also a very common statistic and parameter.
- In this case, we use $\bar{x}$ to estimate μ.

We will talk more about estimation strategies for the mean in another chapter.