Goals for the Week
Sampling Distributions
Confidence Intervals

STAT 140: Week 7

My hope is that after Week 7 (10/13-10/15), you will be able to:

After this week, you should be able to:

- Derive the sampling distribution for a sample proportion.
- Build and interpret a confidence interval for a proportion.
- Understand how changing different characteristics of a confidence interval affects the interval.

Sampling Distributions

- Remember, random variables are mostly theoretical! In these problems, we have been given parameters, like μ, σ, or p, and then we calculate probabilities based on what *should* be happening.
- In reality, what we actually observe gets compared to what we think should happen–i.e., we are using *statistics* to estimate *parameters* and then drawing conclusions about the population.
- We can think of statistics, like $\bar{x}$ and $\hat{p}$, as linear combinations of random variables.
- Then, a statistic has a distribution–this distribution is called the **sampling distribution**.

Sample Proportion, $\hat{p}$ 1

Remember,

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}x_1 + \frac{1}{n}x_2 + ... + \frac{1}{n}x_n$$

where $x_1, x_2, ..., x_n$ are either 0 (failure) or 1 (success).

- What variable can we use to describe a single trial that is either a success or failure?

Sample Proportion, $\hat{p}$ 2

- $x_1, x_2, ..., x_n$ are independent and identical *Bernoulli* random variables!
- Then, $x_1, x_2, ...x_n$ all have mean p, so...

$$\begin{aligned} E(\hat{p}) &= E\left(\frac{1}{n}x_1\right) + E\left(\frac{1}{n}x_2\right) + ... + E\left(\frac{1}{n}x_n\right) \\ &= \frac{1}{n}E(x_1) + \frac{1}{n}E(x_2) + ... + \frac{1}{n}E(x_n) \\ &= \frac{1}{n}p + \frac{1}{n}p + ... + \frac{1}{n}p = p \end{aligned}$$

Standard Error for Proportions

Because $x_1, ...x_n$ are Bernoulli random variables, they have variance $p(1-p)$, and

$$\begin{aligned} Var(\hat{p}) &= Var\left(\frac{1}{n}x_1\right) + Var\left(\frac{1}{n}x_2\right) + ... + Var\left(\frac{1}{n}x_n\right) \\ &= \frac{1}{n^2}Var(x_1) + \frac{1}{n^2}Var(x_2) + ... + \frac{1}{n^2}Var(x_n) \\ &= \frac{1}{n^2}p(1-p) + \frac{1}{n^2}p(1-p) + ... + \frac{1}{n^2}p(1-p)2 \\ &= \frac{p(1-p)}{n} \end{aligned}$$

The standard deviation would be $\sqrt{\frac{p(1-p)}{n}}$.

CLT for Proportions 1

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

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

and

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

CLT for Proportions 2

- The sample proportion $\hat{p}$ acts as a reasonable substitute for p when applying the Central Limit Theorem to a real-world setting.
- This **substitution approximation** of using $\hat{p}$ in place of p is also useful when computing the standard error of a sample proportion.
- For a particular sample size, the variability will be largest when $p = 0.5$.

Conditions

- Independence: The sample observations must be independent!
  - The most common way to satisfy this condition is to use a simple random sample from the population.
  - The data can also come from a random process (a.k.a., rolling a die).
- Normality:
  - In order for the sample size to be considered sufficiently large, we must have $np \geq 10$ and $n(1-p) \geq 10$.

Point Estimates

- **Parameter**: A parameter is a numerical characteristic describing an *entire population*.
- **Point Estimate**: A point estimate is a *statistic* estimating a parameter with a single value.
- Example: The true mean year, μ, in our population of pennies, is 1993.414, but the average year on a sample of 25 pennies, $\bar{x}$, is 1989.8.
- Example: The true proportion of pennies older than 1980, p, is 19.89899%, but the proportion from a sample of 25 pennies, $\hat{p}$ is 0.2.
- $\bar{x}$ and $\hat{p}$ are themselves point estimates!

Sampling Error

- **Error**: The difference between the parameter and the observed point estimate/statistic.
- **Sampling Error** describes how much an estimate will tend to vary from one sample to the next–a.k.a,
- Every time you take a sample, you can calculate a point estimate
  - e.g., 1989.8, 1995.16, 1994.44.
  - e.g., 20%, 0%, 28%.
- Each of these point estimates has an associated error.
  - e.g., -3.614, 1.746, 1.206.
  - e.g., 0.10101%, -19.89899%, 8.10101%
- Sampling error is uderstood and quantified using the standard error, which comes from the sampling distribution of a statistic.

Sampling Distributions

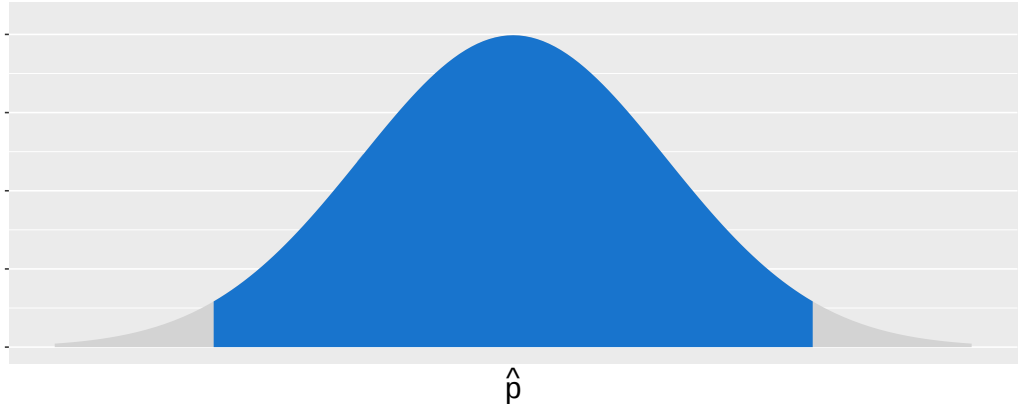
- Remember that probability distributions, including the sampling distribution, describe the theoretical behavior of a random variable.
- Now that we know that every statistic has a distribution, we can understand the theoretical behavior of the statistic.
- This means we can...
  - Understand how likely a statistic is to occur.
  - Understand what kinds of values are reasonable.

Confidence Intervals for a Proportion

- The *most* reasonable value for the parameter, like the population proportion p, is the point estimate (or statistic), but if we report a point estimate, like $\hat{p}$, we probably will not hit the exact value.
- However, if we report a range of plausible values, we have a good shot at capturing the parameter.

Capturing Values

- **Confidence Interval**: A **confidence interval** represents a range of plausible values where we are likely to find the population parameter.



95% CI for a Proportion 1

- The interval extends 1.96 standard errors from the sample proportion, so we are **95% confident** that the interval captures the population proportion.
- In a normal distribution, 95% of the data is within 1.96 standard deviations of the mean.

```
pnorm(1.96, mean = 0, sd = 1) - pnorm(-1.96, mean = 0, sd = 1)
```

```
## [1] 0.9500042
```

- 95% is known as the **confidence level**.
- 1.96 represents the 97.5^{th} percentile of the standard normal, and -1.96 represents the 2.5^{th} percentile of the standard normal.

```
qnorm(0.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

95% CI for a Proportion 2

$$\textit{Point Estimate} \pm 1.96 \times \textit{SE}$$

- The sample proportion $\hat{p}$ is the most reasonable value of the population proportion, so it makes sense to build a confidence interval around this point estimate.
- The standard error (SE) provides a guide for how large we should make the confidence interval.
- **95% Confident**: The interval extends 1.96 standard errors from the sample proportion, so we are **95% confident** that the interval captures the population proportion.

Changing the Confidence Level 1

- Suppose we want to consider confidence intervals where the confidence level is higher than 95%, such as 99%.

```
qnorm(0.975, mean = 0, sd = 1)
```

```
## [1] 1.959964
```

```
qnorm(0.995, mean = 0, sd = 1)
```

```
## [1] 2.575829
```

- What would happen to the confidence interval if we used a Z-score associated with a higher confidence level to build the confidence interval?

Changing the Confidence Level 2

- Increasing the confidence level makes the entire interval wider!!
- If a point estimate closely follows a normal model with standard error SE, then a confidence interval for the population parameter is

$$Point\ Estimate \pm z^* \times SE$$

where z^* corresponds to the confidence level selected.

- In a confidence interval, $z^* \times SE$ is called the **margin of error**.

Interpreting Confidence Intervals 1

- What does confidence mean?
- Confidence intervals built for the same population using the same methodology will contain the population parameter 90%, 95%, 99% of the time.
- Confidence intervals are always statements about parameters, not statistics.
- Confidence intervals are ONLY statements about parameters, not individual values.
- Confidence intervals are NOT statements about the probability that an interval captures a parameter.
- Confidence intervals only capture information about sampling error.

Interpreting Confidence Intervals 2

- **We are XX% confident what the population parameter falls between the lower bound and the upper bound.**
- Use the context!
- We are 90% confident that 87.1% to 90.4% of American adults support the expansion of solar power in 2018.
- We are 95% confident that the proportion of New York adults in October 2014 who supported a quarantine for anyone who had come into contact with an Ebola patient was between 0.796 and 0.844.
- We are 99% confident that the proportion of American adults that support expanding the use of wind turbines is between 81.9% and 87.7% in 2018.

Confidence Interval for a Single Proportion

Once you've determined a one-proportion confidence interval would be helpful for an application, there are four steps to constructing the interval.

1. **Prepare**. Identify $\hat{p}$ and n, and determine what confidence interval you wish to use.
2. **Check**. Verify the conditions to ensure $\hat{p}$ is nearly normal.
3. **Calculate**. If the conditions hold, compute the SE using $\hat{p}$, find z^*, and construct the interval.
4. **Conclude**. Interpret the confidence interval in the context of the problem.

Confidence Intervals: Example 1

In a sample of 300 Twitter users in the United States from 8/17/21 to 8/23/21, 72 said they were uncertain of or unwilling to get a SARS-COV-2 vaccine.

- What is the sample proportion, $\hat{p}$, of Twitter users in the United States who are uncertain of or unwilling to get the vaccine?
- What is the sampling distribution of $\hat{p}$? Can we say that it is nearly normal?
- What z^* should we use for a 99% confidence interval? What is the margin of error? What is the confidence interval?

Confidence Intervals: Example 2

$$\hat{p} = \frac{72}{300} = 0.24$$

$$0.24 \times 300 = 72 > 10, 0.76 \times 300 = 228 > 10$$

$$\hat{p} \sim N(0.24, \sqrt{\frac{0.24 \times 0.76}{300}}) = N(0.24, 0.025)$$

- The conditions for building a confidence interval hold since $n\hat{p}$ and $n(1 - \hat{p})$ are greater than 10.

Confidence Intervals: Example 3


```
SE <- sqrt((0.24*0.76)/300)
```

```
z_star <- qnorm(0.995, mean = 0, sd = 1)
z_star
```

```
## [1] 2.575829
```

```
MoE <- z_star*SE
MoE
```

```
## [1] 0.06351391
```

Confidence Intervals: Example 4


```
72/300 - MoE
```

```
## [1] 0.1764861
```

```
72/300 + MoE
```

```
## [1] 0.3035139
```

- We are 99% confident that the true number of Twitter users in the US who are uncertain of or unwilling to get the vaccine is between 17.6% and 30.4%.

Confidence Intervals: Example 5

`prop.test()` uses slightly different quantities/methods to calculate a CI, but we can use it for the purposes of this test.

```
prop.test(x = 72, n = 300, conf.level = 0.99)
```

```
##
##  1-sample proportions test with continuity correction
##
## data:  72 out of 300, null probability 0.5
## X-squared = 80.083, df = 1, p-value < 2.2e-16
## alternative hypothesis: true p is not equal to 0.5
## 99 percent confidence interval:
##  0.1810669 0.3104756
## sample estimates:
##    p
## 0.24
```