

Practice Session 4

Introduction to Sampling Distributions

Recall that a “sample” is a subset of a population. We can use samples to calculate “statistics”, such as the sample mean, median, and standard deviation. If we were to **repeatedly** sample from the same population, and calculate the **mean** of each sample, we would obtain what is known as a Sampling Distribution of the **Sample Mean**. If we were to instead calculate a **Proportion** from each sample, we would obtain the Sampling Distribution of the **Sample Proportion**.

In general, Sampling Distributions are the distributions of a **statistic**. This is in contrast to the Population, which is the distribution of all individuals or sampling units.

Sampling Distribution Terminology

Match each term with its definition.

Terms:

- s
- σ
- Standard Error (SE)
- x
- \bar{x}

Definitions:

- Standard deviation of a population
- Individual unit of the population
- Standard deviation of a sample
- Sample mean
- Standard deviation of a sampling distribution

Generating Sampling Distributions

Let's generate a sampling distribution using polling data from the president's approval rating. Make sure to load the `library(SDS1000)` to have access to the functions for this exercise.

1.) Use the `get_approval_sample()` function to obtain a sample of $n = 1000$ individual approval ratings. Use the `get_proportions()` function to obtain the proportion of voters in this sample that approve of the president.

```
# Setting the seed so that we all get the same "random sample"
library(SDS1000)
set.seed(1000)
```

2.) Next, use the `do_it` function from class to generate 10000 samples, and repeat the above steps to obtain the proportion for each sample. Assign this vector of proportions to the variable `polling_distribution`, and create a histogram of it. Make sure to have an appropriate title.

Challenge: What do values on the x-axis of this histogram correspond to? How does this differ from histograms that we've previously encountered?

```
# Your code here
```

3.) Finally, calculate the mean and standard deviation of your collection of sample proportions. What is another name for the standard deviation?

```
# Your code here
```

Properties of Sampling Distribution

Answer the following True/False questions relating to sampling distributions.

- True/False: As the sample size n decreases, the standard error of a sampling distribution also decreases.
- True/False: The mean of a sampling distribution is always **greater than** the mean of population.
- True/False: The mean of a sampling distribution is always **less than** the mean of the population.
- True/False: Assuming that you have an unbiased sampling procedure with a sufficient sample size, the sampling distribution of the sample mean is approximately normally distributed.
- True/False: It is generally better to have a larger standard error than a smaller one.

Introduction to Confidence Intervals

Match each term with its definition:

Terms:

- Point Estimate
- Interval estimate
- Confidence Interval
- Confidence Level

Definitions:

- An interval that is generated with a specific method that guarantees that a certain percentage of these intervals will contain the population parameter.
- The percentage of confidence intervals that will contain the population parameter of interest. For example, a 95% confidence level means that 95% of confidence intervals generated will contain the population parameter of interest.
- A sample statistic that we use to estimate a population parameter. The sample mean is an example as we use it to estimate the population mean.
- An interval that contains a range of plausible values for a population parameter. The point estimate sits at the center of this interval.

Confidence Interval for Baseball Salaries

For this section, we will use the `BaseballSalaries2015` data set from the `Lock5Data` library.

1.) Using the `sample()` function, obtain a sample of 100 baseball player salaries. Make sure to sample **without replacement**. See the `help` page for `sample()` for details.

```
# Your code here
```

2.) Now, using the `do_it()` function, repeat this process $n = 10,000$ times, and calculate the mean each time.

```
# Your code here
```

3.) Using your original sample of 100 salaries from part 1.), calculate a point estimate of the mean salary.

```
# Your code here
```

4.) Calculate the standard deviation of your sampling distribution from part 2.) to estimate the standard error.

```
# Your code here
```

5.) Using your answers from part 3.) and 4.), generate a 95% interval estimate for the mean salary of baseball players. The interval estimate is of the form:

Point Estimate $\pm 1.96 \cdot$ Standard Error.

```
# Your code here
```

6.) Interpret your interval estimate.

7.) Compare your interval estimates to the actual mean salary from the entire data set. Comment on what you notice.

```
# Your code here
```