Lab 3: Sampling Distributions

Learning Objectives

Students will be able to:

- Increase skill and comfort with basic data management and syntax in R.
- Review and increase understanding of the basic methodology and syntax of the for loop.
- Simulating the sampling distributions for sample means.
- Review the concepts of the sampling distribution and the normal approximation.
- Practice constructing and interpreting confidence intervals.

Readings

To prepare for this lab assignment read the following tutorials

1. Normal Distribution
2. Sampling and Simulation
3. Sampling Distribution
4. Performing a t-test for μ in R

Required Data and Packages

Visit https://causeweb.org/tshs/hypoxia/ and download the hypoxia dataset, the format will be in `.rdata`

Once downloaded, you can access the dataset by using the `load()` command

```
load("path to hypoxia.rdata file")
```

Data Description

Obstructive sleep apnea (OSA) is a condition in which breathing repeatedly stops or becomes restricted during sleep. These interruptions can reduce oxygen levels in the blood, potentially affecting cardiovascular health.

The Hypoxia dataset contains information from 281 adult patients diagnosed with OSA who underwent bariatric surgery. Researchers investigated whether reduced oxygen levels during sleep were associated with decreased mean arterial pressure during surgery.

Mean arterial pressure (MAP) is a measure of average arterial blood pressure over a cardiac cycle. In this lab, we will examine the time-weighted average MAP during surgery, measured in millimeters of mercury (mmHg).

We will use the observed patients as a finite population for our sampling simulations. This is a learning exercise and should not be interpreted as a random sample of all patients with OSA.

For desriptions of the variables in the dataset, download the Data Dictionary

---

> **!** Important
>
> Submit a single PDF for your lab report using Quarto.
>
> Copy/paste each question from the lab, provide R code solution and explain solution *(if needed)*.
>
> Failure to follow the appropriate format will result in points reduction.

Section 1: Exploring the Data

1. How many observations and variables are contained in the dataset ? What does each row represent ?
2. Read over the data dictionary, What does `TWA` stand for ? What does `MAP` stand for ?

Extract `TWA MAP` variable and store the values in a vector called `

```
twa_map <- hypoxia$`TWA MAP`
```

3. How many observations are missing `twa_map` measurements ?
4. Create an informative histrogram, with appropriate titles and customization for the variable `twa_map`. Describe the distribution of `twa_map`
5. Are there any outliers in `twa_map` measurements ?
6. Report the mean and standard deviation of `twa_map`. Interpret the mean of these measurements.

Section 2: Simple Random Sampling

Although we have access to all available MAP measurements, suppose we could only collect data on 30 patients.

```
set.seed(10)
sample_0 <- mean(twa_map, size = 30)
```

1. Calculate the sample mean $\bar{x}$ and sample standard deviation (s) from
2. Repeat the sampling process using a different seed. Does the sample mean change ? Explain how this illustrates sampling variability.

Section 3: Simulating the Sampling Distribution

The sampling distribution of the sample mean describes how $\bar{X}$ varies over repeated random samples of a fixed size.

To investigate this distribution, we will draw $1,000$ random samples, each containing $n = 30$ patients, and calculate the sample mean for each sample.

We can accomplish this using a `for loop`.

```
n <- 30       # Sample size
M <- 1000     # Number of repetitions

# Create a vector to store sample means
xbars <- numeric(M)

set.seed(10)

for (i in 1:M) {
  sample_i <- sample(twa_map, size = n)
  xbars[i] <- mean(sample_i)
}
hist(xbars) # visualize sampling distribution
```

1. Describe the purpose of the following components of the code:

- `numeric(M)`
- `for (i in 1:M)`
- `sample(twa_map, size = n)`
- `xbars[i] <- mean(sample_i)`

Why should `set.seed()` be placed outside the loop?

2. Compute the mean and standard deviation of `xbars`. What do these values represent ?
3. Construct a histogram of the simulated sample means. Does it appear to be normally distributed ?

Section 4: Constructing a Confidence Interval

Suppose the random sample of 30 patients obtain in Section 2 is the only information available to us

```
set.seed(10)
sample_0 <- sample(twa_map, size = 30)
```

We want to estimate the population mean MAP using a 95% confidence interval. Since the population standard deviation is generally unknown in real applications, we use a one-sample t confidence interval

1. Using `sample_0`, compute the sample mean $\bar{x}$ and sample standard deviation (s).

2. Compute the critical value from a t-distribution with appropriate degrees of freedom, needed for a 90% confidence interval. Call this value `t_val`

3. Compute the standard error of $\bar{X}$, name it `std_error`. Obtain the margin of error using the critical value `t_val` from the previous problem.

```
margin_error <- t_val * std_error
```

4. Construct a 95% confidence interval for the mean MAP using the results from the previous questions.

```
sample_mean + c(-1,1)* margin_error
```

Here, `c(-1,1)` will account for the $\pm$

5. Interpret the constructed confidence interval in the context of the problem. Does our interval contain the population mean μ, in this case `mean(twa_map)`.