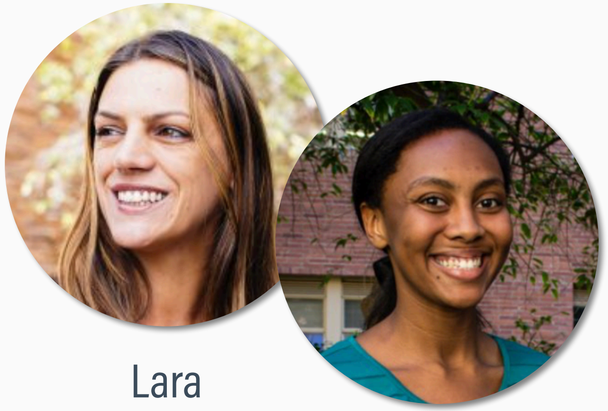
LAB WEEK 3

SAMPLING ERROR AND SAMPLING DISTRIBUTIONS

SMOKING AND DRINKING CESSATION TRIAL

Pharmacological treatments that can concomitantly address cigarette smoking and heavy drinking stand to improve health care delivery for these highly prevalent co-occurring conditions. This superiority trial compared the combination of varenicline and naltrexone against varenicline alone for smoking cessation and drinking reduction among heavy-drinking smokers.



Lara Ray

ReJoyce Green

Ray, L.A., Green, R., Enders, C., et al. (2021). Efficacy of combining varenicline and naltrexone for smoking cessation and drinking reduction: A randomized clinical trial. *American Journal of Psychiatry*, *178*, 818–828.

KEY VARIABLES

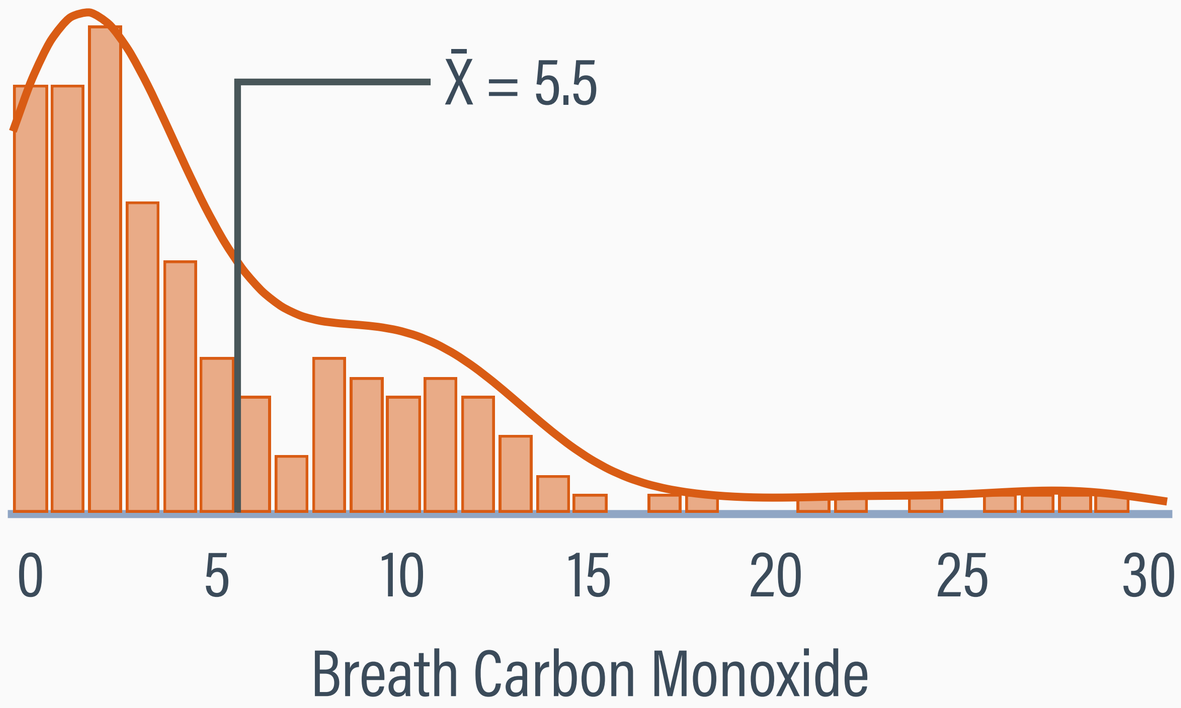
Breath (alveolar) carbon monoxide

A measure of carbon monoxide in the lungs. Breath carbon monoxide is a biomarker of smoking behavior common in clinical trials.

Medication arm

Participants were randomly assigned to receive one of two meds: varenicline plus naltrexone or varenicline plus placebo pills

SAMPLE STATISTICS REVISITED



Mean ($\bar{X}$)	5.53
Std. Dev. (s)	5.96
N	165

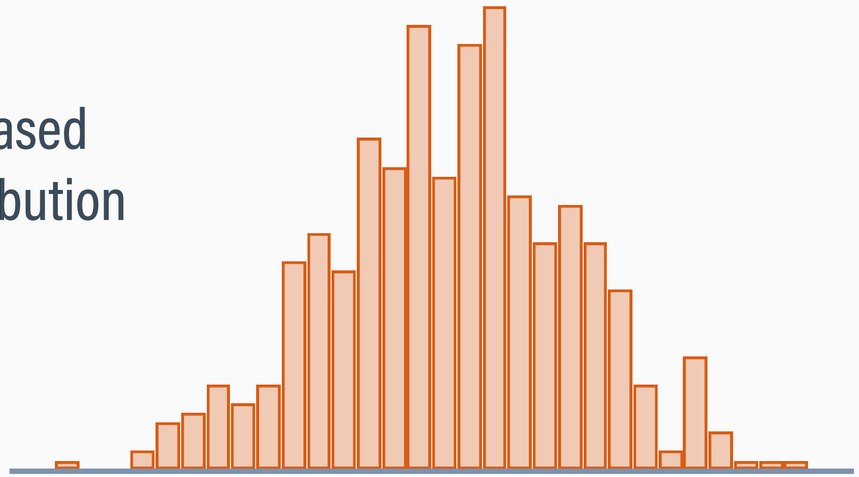
SAMPLING ERROR

- The sample mean is not the same as the true mean in the population
- How close is the estimate to the true mean?
- We can use computer simulation and statistical theory to answer a related question: Across many different samples, how close would the sample means be, on average?

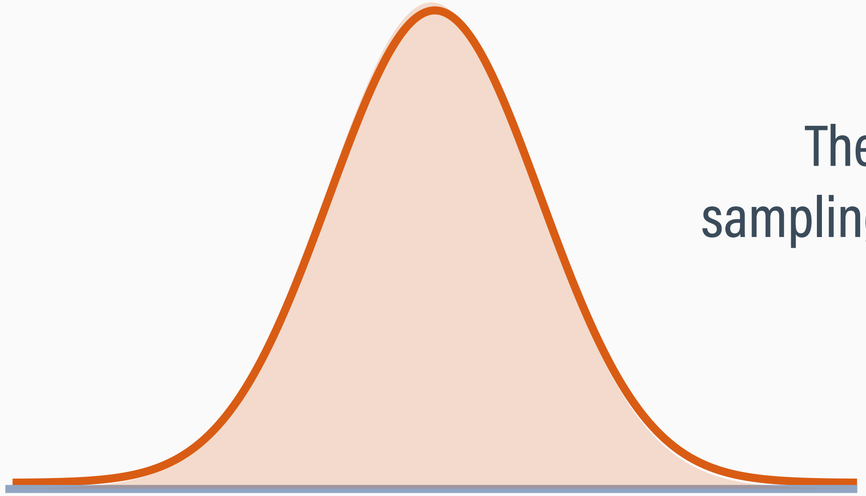
CENTRAL LIMIT THEOREM PART 1

- With a large enough sample size, the means from many random samples follow a normal distribution

Simulation-based sampling distribution



Sample Means (Estimates)



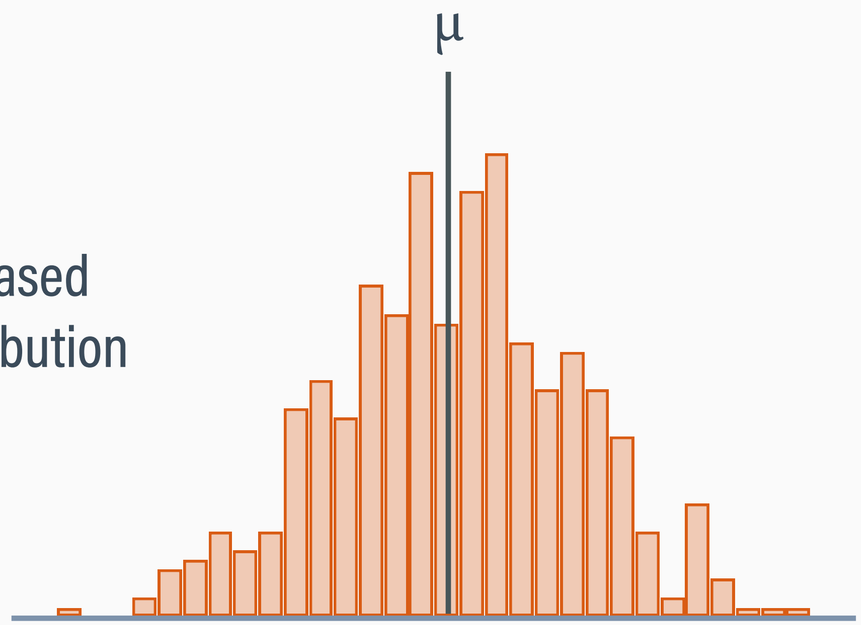
Theoretical sampling distribution

Sample Means (Estimates)

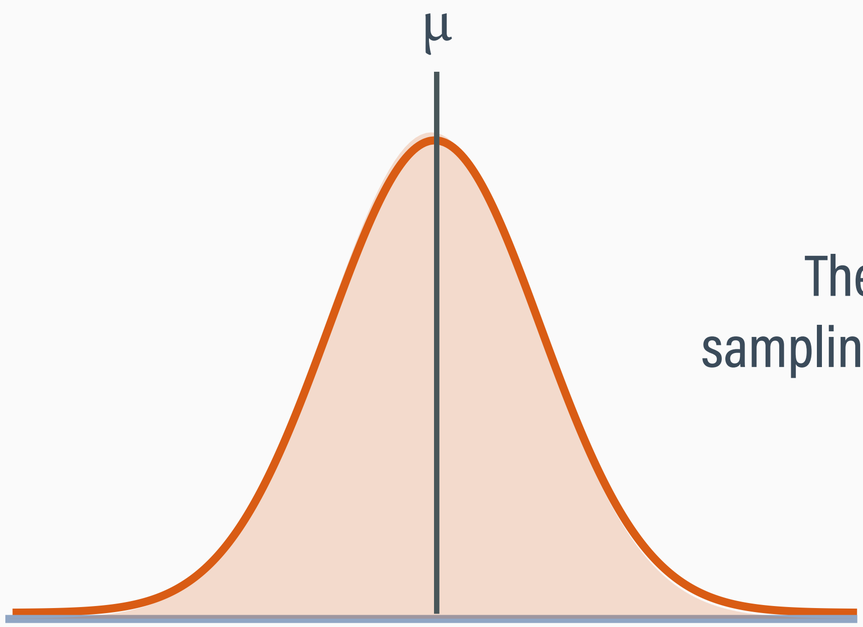
CENTRAL LIMIT THEOREM PART 2

- Estimates vary symmetrically around the true mean (across many hypothetical samples, estimates average out to the true mean)

Simulation-based sampling distribution



Sample Means (Estimates)



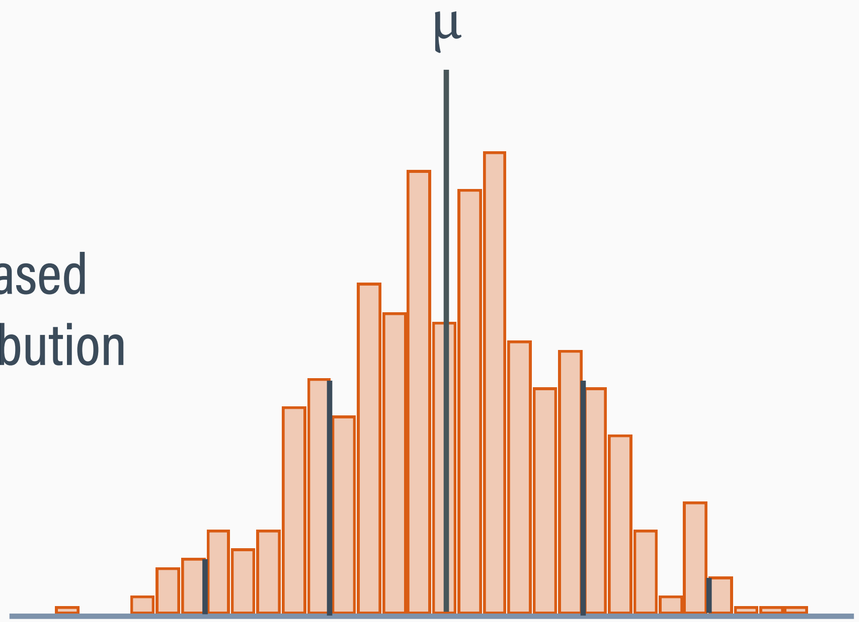
Theoretical sampling distribution

Sample Means (Estimates)

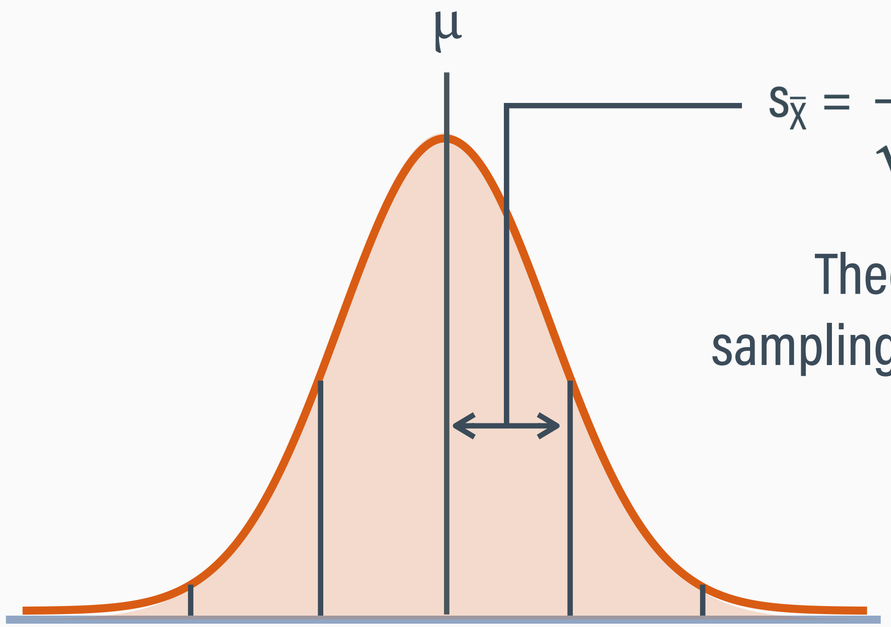
CENTRAL LIMIT THEOREM PART 3

- The standard error of the sample means can be estimated by dividing the standard deviation of the scores by $\sqrt{N}$

Simulation-based sampling distribution



Sample Means (Estimates)



Theoretical sampling distribution

Sample Means (Estimates)

LOAD PACKAGES AND IMPORT DATA

= data frame name
= variable name
= raw data file name

```
# LOAD R PACKAGES ----

# load R packages
library(psych)

# READ DATA ----

# github url for raw data
filepath <-
  'https://raw.githubusercontent.com/craigenders/psych250a/main/data/ClinicalTrialData.csv'

# create data frame called ClinicalTrial from github data
ClinicalTrial <- read.csv(filepath, stringsAsFactors = T)
```

SUMMARIZING DATA

= data frame name

= variable name

```
# INSPECT DATA ----

# summarize entire data frame (summarytools package)
dfSummary(ClinicalTrial)

# DESCRIPTIVE STATISTICS ----

# descriptive statistics (psych package)
describe(ClinicalTrial$COWeek8)
```

R OUTPUT

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
COWeek8	6	165	5.53	5.96	3	4.55	2.97	0	29	29	1.75	3.35	0.46

THEORY-BASED STANDARD ERROR

- The formula from the central limit theorem predicts that the average amount of sampling error is 0.46 breath CO points
- The average distance from many hypothetical sample means (N = 165) and the true mean is ±0.46

Mean ($\bar{X}$)	5.53
Std. Dev. (s)	5.96
N	165

$$s_{\bar{X}} = \frac{s}{\sqrt{N}} = \frac{5.96}{\sqrt{165}} = 0.46$$

COMPUTER SIMULATION RECIPE

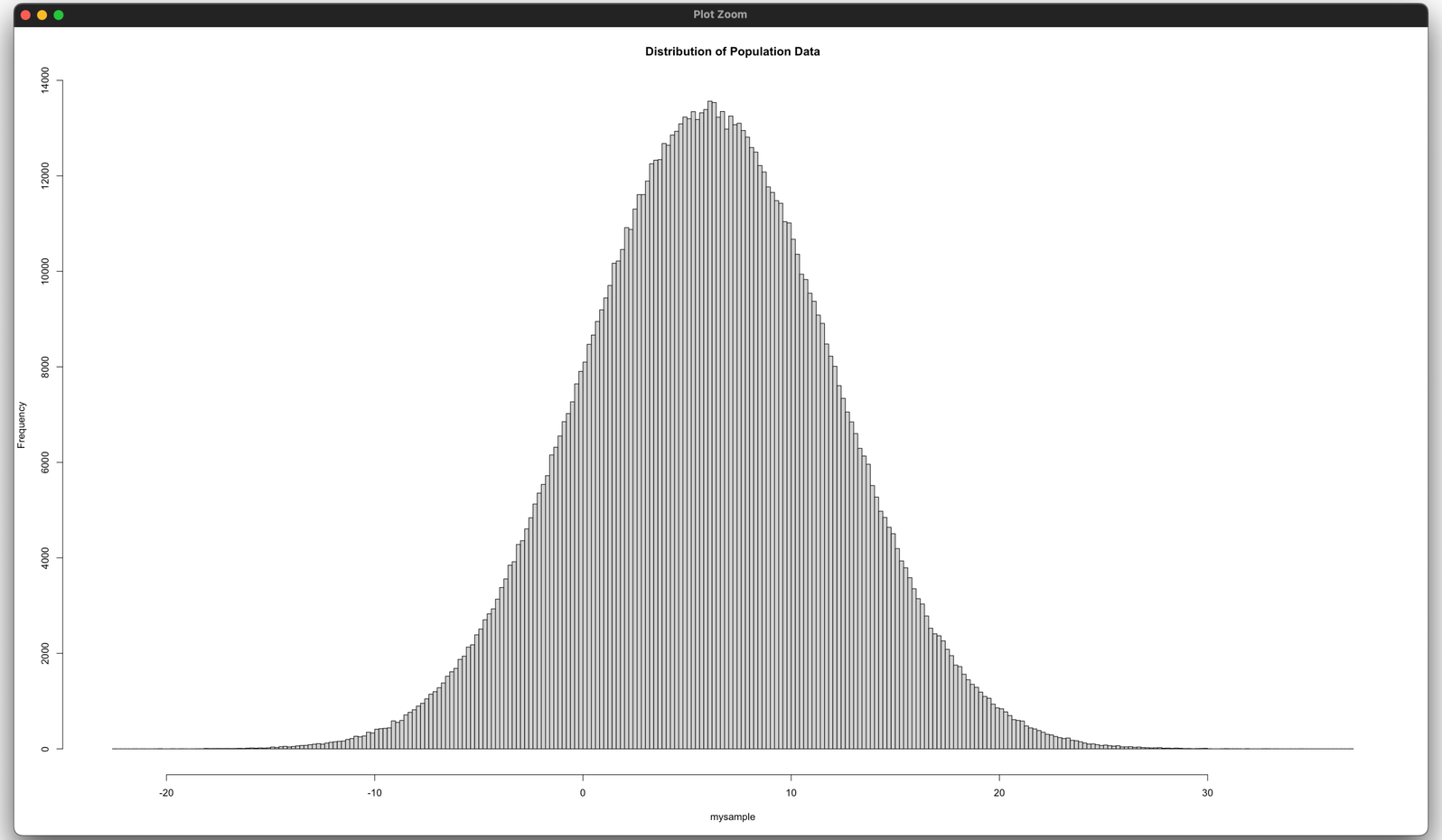
1. Provide the true mean and standard deviation in a hypothetical population (a real data set can provide good guesses)

2. Use RStudio to create many random samples of artificial data from the hypothetical population

3. Compute the mean from each artificial data set

4. Compute the average distance from the simulated sample means to the true population mean (the average error)

SIMULATION 1: NORMAL POPULATION



SIMULATION INPUTS

```
# SPECIFY POPULATION PARAMETERS ----

# simulation inputs
pop_mean <- 6           # population mean
pop_sd <- 5.96          # population standard deviation
sample_N <- 165         # sample size
normal_data <- TRUE     # TRUE = normal data, FALSE = skewed data
```

GENERATE ONE DATA SET

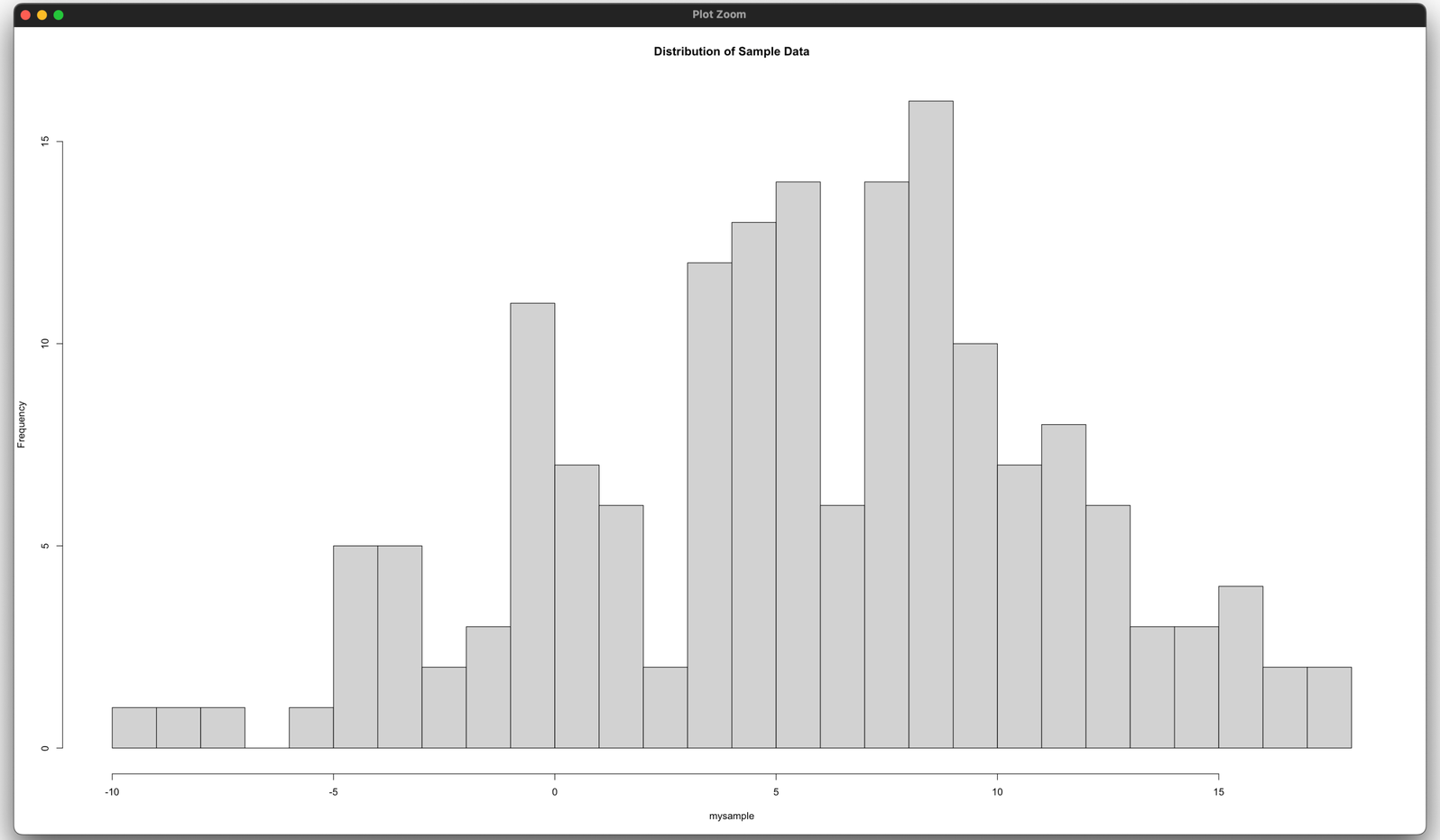
- Running this code section (do not modify) generates a single sample of data to illustrate the process of simulating a random sample from a population

```
# GENERATE AND INSPECT ONE SAMPLE OF DATA ----

# simulate artificial data set from the population
if(normal_data){ # simulate normal data
  mysample <- rnorm(n = sample_N, mean = pop_mean, sd = pop_sd)
} else { # simulate skewed data
  mysample <- pop_mean + ((rchisq(sample_N, pop_mean) - pop_mean)/sqrt(2*pop_mean)) * pop_sd
}

# graph the simulated sample data
hist(mysample, breaks = 50, main = 'Distribution of Simulated Data')
```

SCORE DISTRIBUTION FROM ONE SAMPLE



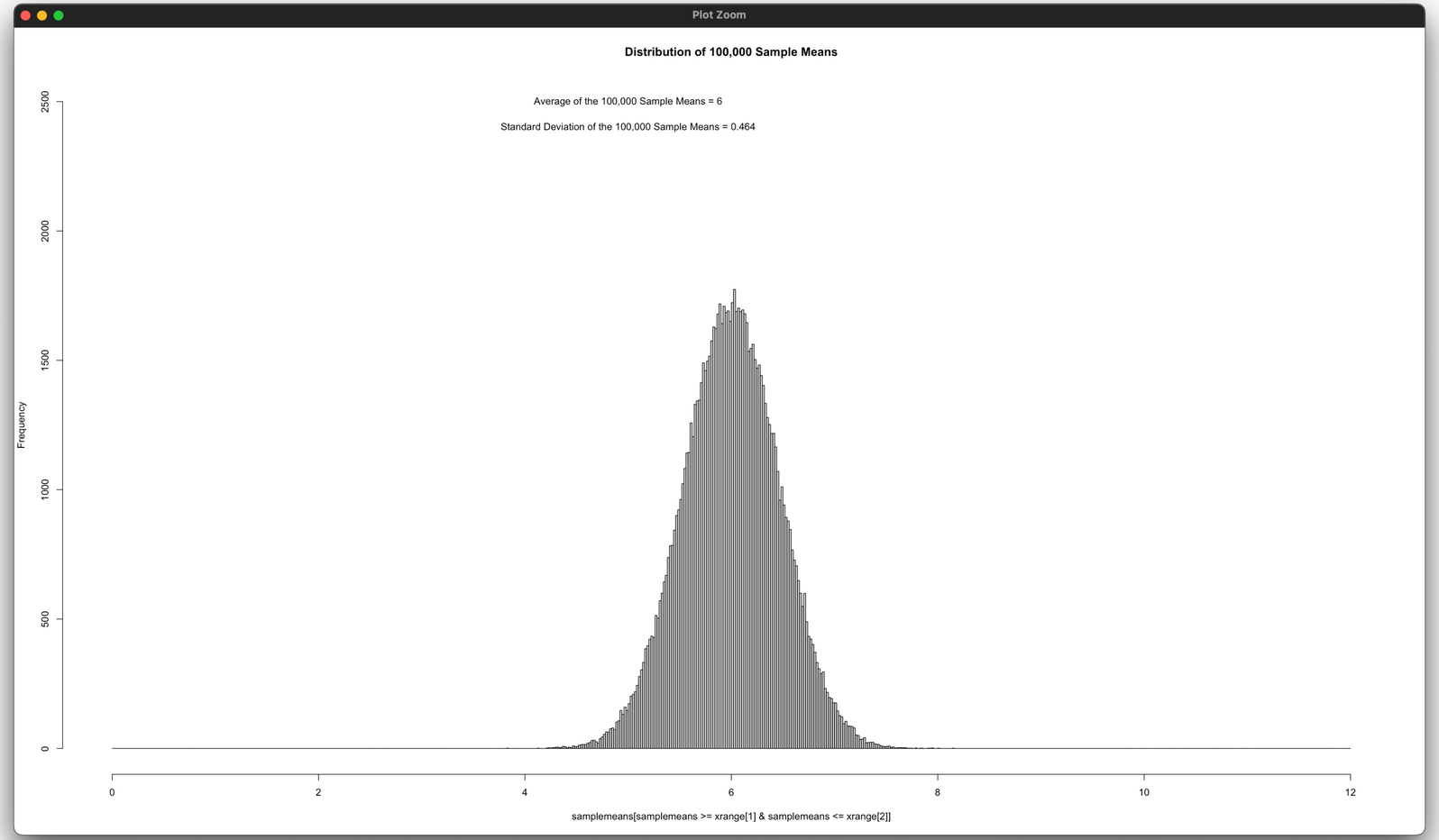
SIMULATION WITH 100,000 DATA SETS

- Running this code section (do not modify) generates 100,000 samples of data, computes the sample mean from each data set, then summarizes the 100,000 sample means

```
# SIMULATION WITH 100,000 RANDOM SAMPLES OF DATA ----

# initialize data set to hold 100,000 sample means
samplemeans <- rep(0, 100000)
# loop to draw 100,000 random samples
for(s in 1:100000){
  if(normal_data){ # normal data
    sampledata <- rnorm(n = sample_N, mean = pop_mean, sd = pop_sd)
  } else { # skewed data
    sampledata <- pop_mean + ((rchisq(sample_N, pop_mean) - pop_mean)/sqrt(2*pop_mean)) * pop_sd
  }
  # store the sample mean in element of s
  samplemeans[s] <- mean(sampledata)
}
...
```

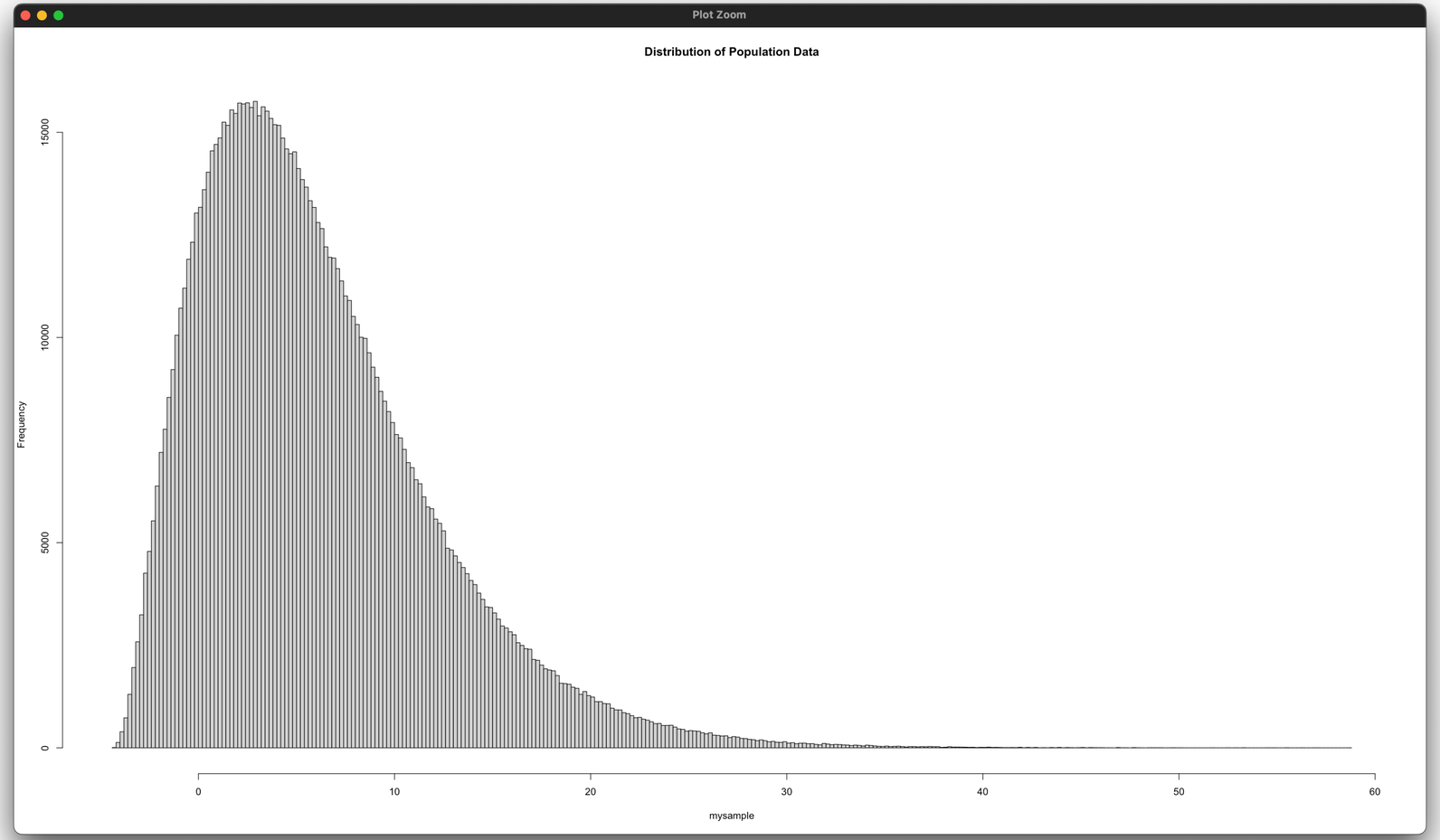
DISTRIBUTION OF 100,000 SAMPLE MEANS



SMALL GROUP EXERCISE

The sampling distribution of the means illustrates how samples vary around a true mean of five. The sample mean from the clinical trial was $\bar{X} = 5.53$. In groups of two or three, discuss how likely it is for a sample mean of 5.53 to originate from a population where the true mean is 5. Justify your conclusion using graphical evidence from the simulated sampling distribution.

SIMULATION 2: SKEWED POPULATION

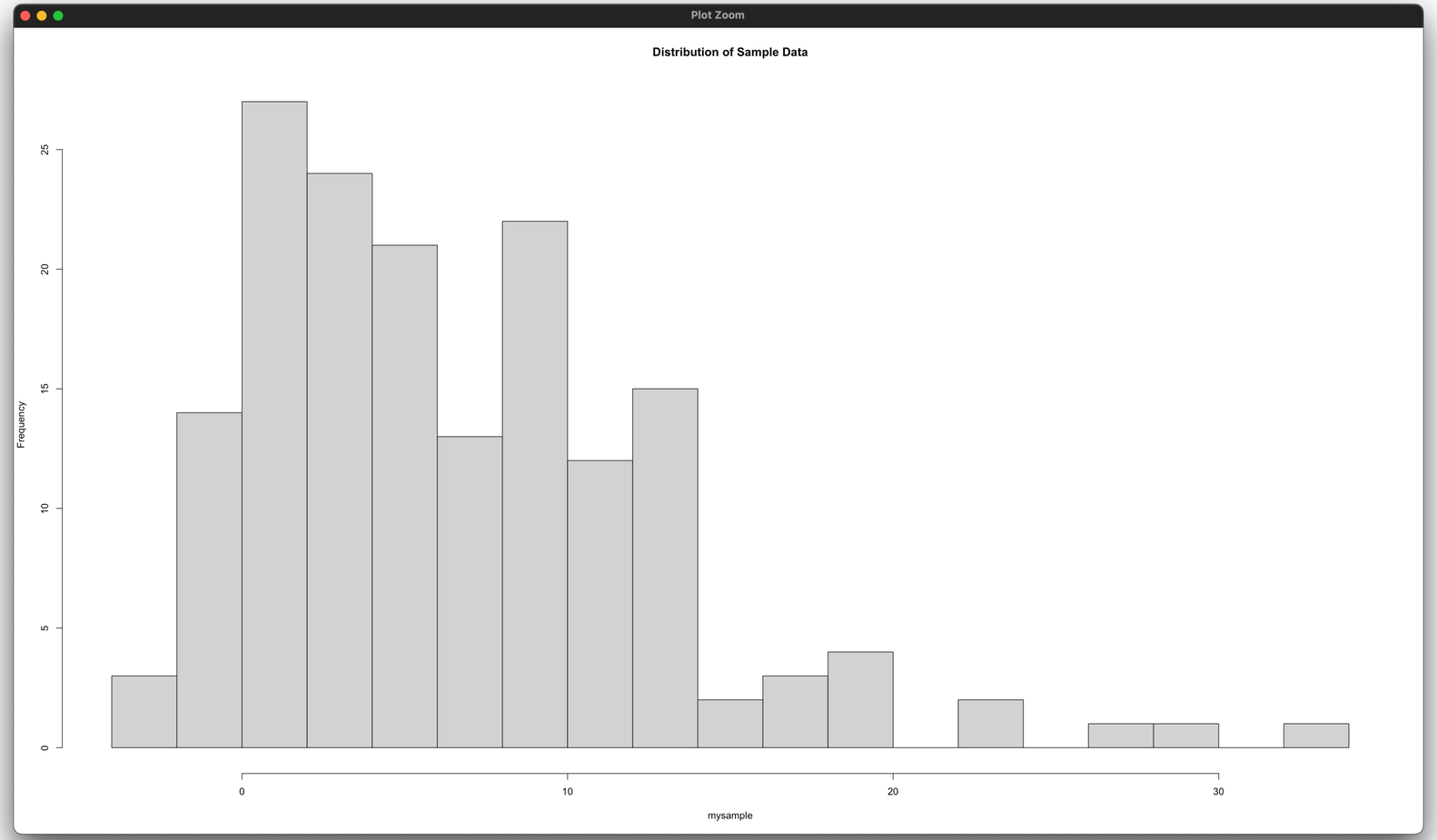


SIMULATION INPUTS

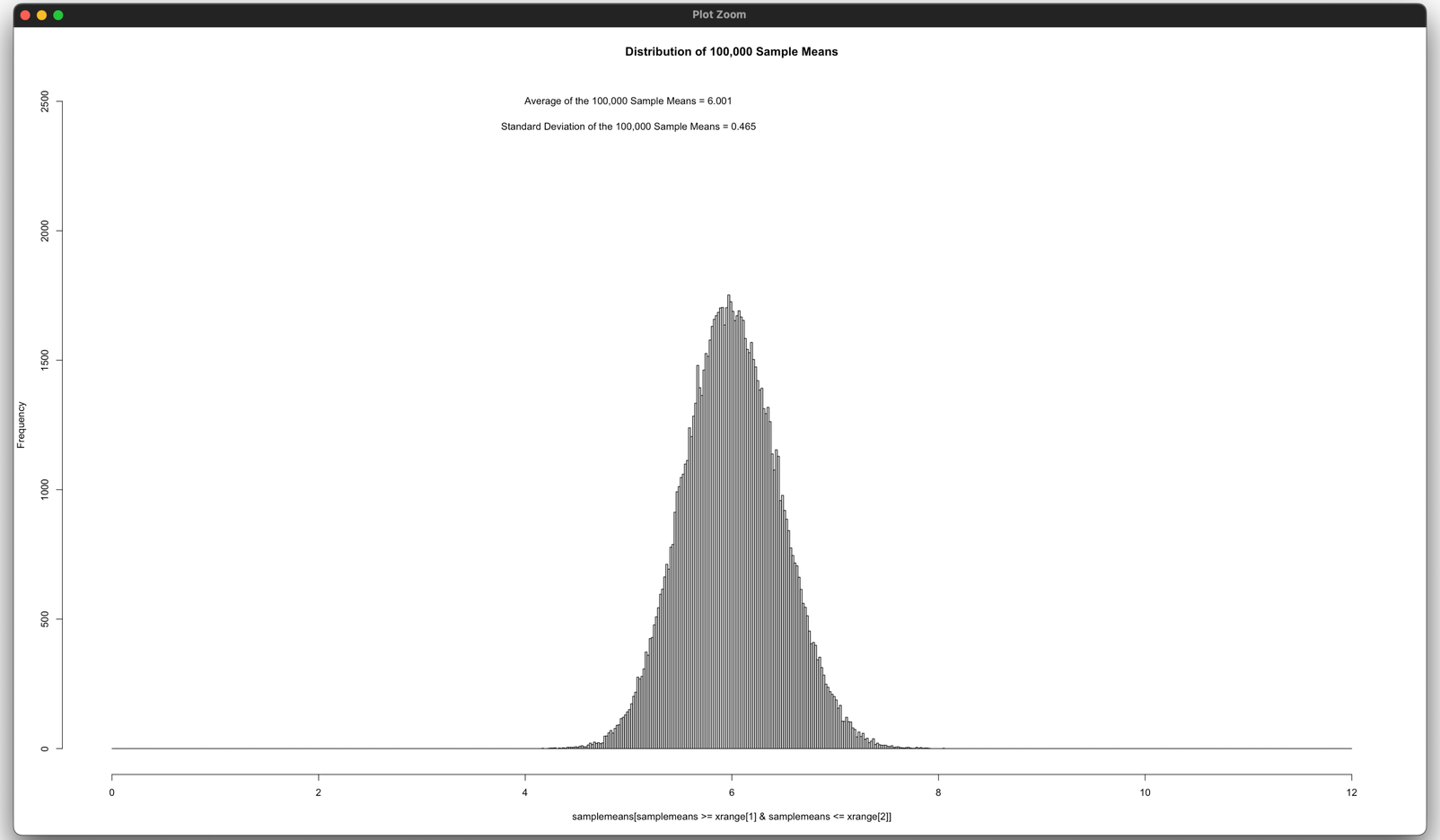
```
# SPECIFY POPULATION PARAMETERS ----

# simulation inputs
pop_mean <- 6           # population mean
pop_sd <- 5.96          # population standard deviation
sample_N <- 165         # sample size
normal_data <- FALSE    # TRUE = normal data, FALSE = skewed data
```

SCORE DISTRIBUTION FROM ONE SAMPLE



DISTRIBUTION OF 100,000 SAMPLE MEANS



SMALL GROUP EXERCISE

Download the “Week 3 Lab. Sampling Error.R” script from Bruin Learn. In groups of two or three, you will complete a series of computer simulation tasks that provide practice for the next assignment. There is no need to write code from scratch; instead, simply modify the simulation inputs at the top of the Lab file.

SMALL GROUP EXERCISE TASK 1

- Modify the computer simulation is to examine the impact of decreasing the sample size on the estimates from many random samples. In this simulation, you will draw 100,000 samples of 82 scores (roughly half the original sample size).

- Before running the simulation, make predictions about the sampling distribution. How will it's shape, center, and spread change as a result of decreasing the sample size?

SMALL GROUP EXERCISE TASK 2

- Evaluate your predictions. What discrepancies, if any, exist between your predictions and the sampling distribution. Did anything surprise you?

- Looking back at the standard error equation, did the sampling distribution's spread change in a way that is consistent with what you would expect when plugging a smaller number into the denominator of the standard error formula?

SMALL GROUP EXERCISE TASK 3

- Modify the computer simulation is to examine the impact of increasing the standard deviation on the estimates from many random samples. In this simulation, you will draw 100,000 samples of 165 scores (the original sample size) from a population with a standard deviation of 12 (roughly twice as large).

- Before running the simulation, make predictions about the sampling distribution. How will it's shape, center, and spread change as a result of increasing variability among its sample members?

SMALL GROUP EXERCISE TASK 4

- Evaluate your predictions. What discrepancies, if any, exist between your predictions and the sampling distribution. Did anything surprise you?

- Looking back at the standard error equation, did the sampling distribution's spread change in a way that is consistent with what you would expect when plugging a larger number into the numerator of the standard error formula?

SMALL GROUP EXERCISE TASK 5

- Modify the computer simulation is to examine the impact of decreasing the sample size to a very small number. In this simulation, you will draw 100,000 samples of 10 scores.

- Before running the simulation, make predictions about the sampling distribution. How will it's shape, center, and spread change as a result of decreasing the sample size?

SMALL GROUP EXERCISE TASK 6

- Evaluate your predictions. What discrepancies, if any, exist between your predictions and the sampling distribution. Did anything surprise you?

- Looking back at the standard error equation, did the sampling distribution's spread change in a way that is consistent with what you would expect when plugging a smaller number into the denominator of the standard error formula?