For the following questions please use the sat\_act.csv file. You can download this dataset for from Canvas under Files > Lab & Homework Assignments > Week 3 > Homework. It contains 50 observations on the following variables

* gender
* education
* age (in years)
* ACT: composite scores from the ACT exam
* SATV: SAT verbal scores
* SATQ: SAT quantitative score

Provide your answers and the code you used for each question in the text boxes below. You may submit answers as text or screenshots. Note: The text has been set to blue inside the text boxes. This is intentional and will make it easier for the TA’s to see your answers.

1. Download the sat\_act.csv file and read it into R. What are the first 3 values in the education column? What are the fourth, fifth, and sixth values in the SATV column? **[2 pt.]**

Answer:

Education: "some college/university" "some college/university" "some college/university"

SATV: 550 600 640

Code/Syntax:

sat <- readr::read\_csv("sat\_act.csv")

sat$education[1:3]

sat$SATV[4:6]

2. Create a histogram of people’s ACT scores, and describe the distribution (what is its shape / does it appear normally distributed, and do there appear to be any outliers?). Be sure to give your histogram a title, and an informative x-axis label.

Answer:

The ACT scores distribution is close to a normal distribution, with a slight negative skewness. There appear to be no outliers since the distribution range is across reasonable values.

A graph of a graph

Description automatically generated

Code/Syntax:

hist(sat$ACT,main = "Histogram of ACT scores",

xlab = "ACT")

3. What is the average ACT score for people who people whose education is “graduate/professional”? **[1 pt.]**

Answer:

31.8

Code/Syntax:

sat |>

dplyr::filter(education == "graduate/professional") |>

dplyr::summarise(average\_act = mean(ACT, na.rm=TRUE))

4. What is the average ACT score for people who people whose education is “graduate/professional” and identified as “female”? **[2 pts.]**

Answer:

31.9

Code/Syntax:

sat |>

dplyr::filter(education == "graduate/professional",

gender == "female") |>

dplyr::summarise(average\_act = mean(ACT, na.rm=TRUE))

5. Compute the standard error of the mean of the SATQ scores. In your own words, explain what this value means. **[2 pts.]**

Answer:

34.36

The standard error of the mean is a measure of the sample mean variation, given the sample size, around the population mean, across repeated sampling from the population. Since we usually cannot know this true value, we estimate it from our sample.

Code/Syntax:

sd(sat$SATQ / sqrt(length(sat)))

6. Copy and paste the following function into RStudio, which is similar to the code we used in the lab this week for examining changes in the sampling distribution.

demo\_sampling\_dist <- function(num\_samples = 20,

sample\_size = 10,

pop\_mu = 0,

pop\_sigma = 1) {

# num\_samples: the number of samples to take

# sample\_size: the sample size for each sample

# pop\_mu: the mean of the \*population\*

# pop\_sigma: the standard deviation of the \*population\*

means\_of\_samples <- rep(NA, num\_samples)

for (i in 1:num\_samples) {

sample\_i <- rnorm(sample\_size, pop\_mu, pop\_sigma)

means\_of\_samples[i] <- mean(sample\_i)

}

xlims <- c(pop\_mu - (3\*pop\_sigma), pop\_mu + 3\*pop\_sigma)

hist(means\_of\_samples,

main = "Sampling Distribution of Means",

xlab = "Values",

xlim = xlims)

abline(v = pop\_mu, col = "red", lwd = 2)

abline(v = mean(means\_of\_samples), col = "steelblue", lwd = 2)

}

What does this code do? It creates a function called demo\_sampling\_dist, which takes as its arguments the number of samples you want to draw (num\_samples), the sample size of each sample (sample\_size), the mean and the standard deviation of the *population* distribution you are drawing from (pop\_mu and pop\_sigma). Every time you use this function, it will draw num\_samples samples of size sample\_size from a population with mean pop\_mu and SD pop\_sigma, and calculate the mean for each of those samples. These means are stored into a vector for you to use later.

The function then creates a histogram of those means (your sampling distribution!), and adds lines representing the population mean and average of all the sample means.

Once you’ve loaded this function, try out different values for each of num\_samples, sample\_size, and pop\_sigma. What do you notice about how each of these values affects the sampling distribution? (i.e., for each of these factors, tell us what happens as they increase or decrease?)

Remember that the blue line is the mean of the sampling distribution and that the red line is the population mean. **[3 pts.]**

Answer:

As the number of samples increases, the means distribution assumes a shape closer to the normal distribution.

Increasing the sample size has the effect of narrowing the distribution of means around the population mean.

Finally, with the increase in the population sigma, the sampling distribution gets more spread out.