# Exam 1 Review Section Worksheet Solutions

This is a worksheet for review for Exam 1. You do not need to turn this in.

1. **Parameter or Statistic?** According to UCSB's Office of Institutional Research, 31% of all new transfer students in Fall 2022 are first-generation college students. The 31% is a
   a. **parameter.**
   b. statistic.

---

2. **Parameter or Statistic?** The average financial aid package amount for 25 randomly selected UCSB undergraduate students is $21,000. The $21,000 is a
   a. parameter.
   b. **statistic.**

---

3. **Podcasts Gain in Popularity** According to an April 2023 Pew Research report, 49% of U.S. adults listened to a podcast in the past year. These results were from a random sample of 5132 U.S. adults. The population for this poll appears to be:
   a. **all U.S. adults.**
   b. all U.S. adults who listened to a podcast in the past year.
   c. the percentage of all U.S. adults who listened to a podcast in the past year. (this is the parameter)
   d. the 49% of U.S. adults who listened to a podcast in the past year. (this is the statistic)
   e. the 5132 U.S. adults. (this is the sample)

---

**Use the following information for Questions 4-5:**
The weights (in grams) of California scrub jays can be modeled with a normal distribution with mean 85 g and standard deviation 5 g.

4. Approximately 99.7% of weights for California scrub jays fall between which of the following values?
   a. 65 to 105 grams
   b. **70 to 100 grams**
   c. 75 to 95 grams
   d. 80 to 90 grams

   Almost all (99.7%) of the values in a normal distribution will be within 3 standard deviations of the mean:
   $$\mu - 3\sigma = 85 - 3(5) = 85 - 15 = 70$$
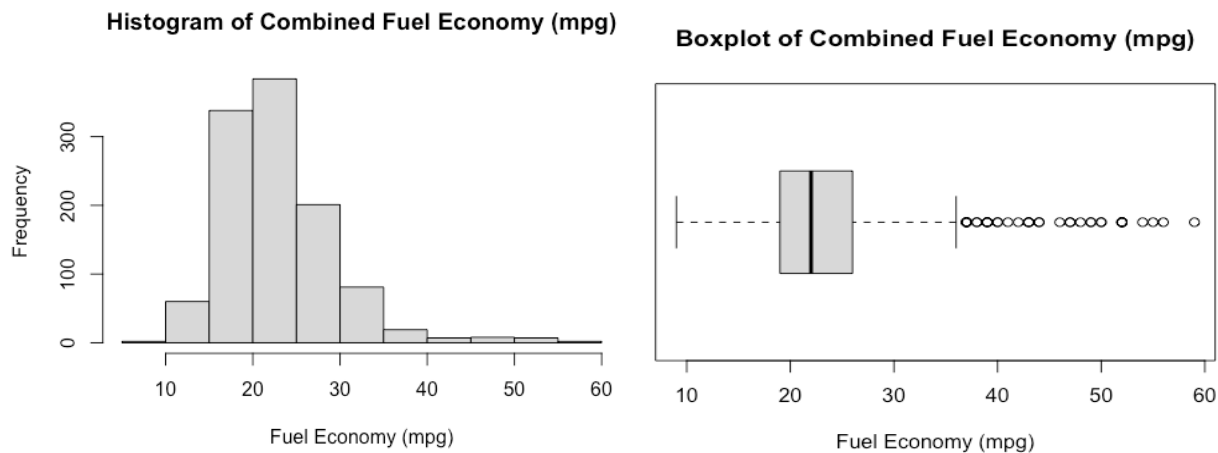   $$\mu + 3\sigma = 85 + 3(5) = 85 + 15 = 100$$

5. One California scrub jay has a weight at the 97.5th percentile for California scrub jays. Which of the following is a close approximation for this scrub jay's weight?

   a. 70 grams
   b. 75 grams
   c. 80 grams

   d. 90 grams
   **a. 95 grams**
   b. 100 grams

   A California scrub jay in the 97.5th percentile for weights is 2 standard deviations above the mean weight. This is because approximately 95% of all weights will be within 2 standard deviations of the mean. That leaves 5% in the two tails combined. Since normal distributions are symmetric, 2.5% is in each tail. A California scrub jay in the 97.5th percentile has a weight Almost all (99.7%) of the values in a normal distribution will be within 3 standard deviations of the mean:

   $$\mu + 2\sigma = 85 + 2(5) = 85 + 10 = 95$$

---

**Use the following information for Questions 6-9:**
The Office of Energy Efficiency & Renewable Energy provides fuel economy data to the public. Interestingly, the data come from vehicle testing at National Vehicle and Fuel Emissions Laboratory. The following histogram and box plot show the combined city and highway gas mileage (combined fuel economy) for the 1109 vehicles in the 2022 data set.



Histogram of Combined Fuel Economy (mpg)



Boxplot of Combined Fuel Economy (mpg)

6. We used R to calculate summary statistics for the combined fuel economies. The two measures of center reported by R are, in increasing order, 22.00 mpg and 23.27 mpg. Which is the mean, and which is the median?

   a. mean = 22.00 mpg; median = 23.27 mpg
   **b. mean = 23.27 mpg; median = 22.00 mpg** (the distribution is skewed to the right, so the mean will be greater than the median)
   c. It is impossible to determine without more information.

7. We used R to calculate summary statistics for the combined fuel economies. Which of the following is the most reasonable value for the standard deviation for the combined fuel economies?
   a. 3 mpg
   b. **7 mpg**
   c. 15 mpg
   d. 25 mpg

8. Approximately what percentage of vehicles have a combined fuel economy of at most 20 mpg?
   a. **36%** (there look to be about 400 vehicles in the first two bars, and 400/1109 = 0.3613)
   b. 50%
   c. 64%

9. Refer to the histogram and box plot above. The boxplot indicates that fuel economies at or above about 36 mpg might be outliers. Should we remove these points from the data analysis? Clearly state "yes" or "no" and briefly (in 1-2 sentences) explain.
   No. We should never remove outliers unless there is a clear reason to do so. There is nothing that suggests that these fuel economies are not valid data points, so we need to keep them in the data set.
   **Common mistakes:**
   - Commented about the number of potential outliers (e.g., too many so cannot remove them or small part of data set so can remove them)
   - Said we shouldn't remove because removing the potential outliers would change the measures of center and variability (true, but not the reason)

---

**Use the following information for Questions 10-12:**
The standard therapy for dogs with inflammatory bowel disease (IBD) is diet plus prednisone. Gut microbe researchers wanted to determine if probiotics along with the standard therapy would help keep intestinal mucosa from becoming inflamed. Thirty-four (34) dogs with IBD were randomized into two treatment groups: one group got the standard diet therapy; the other group got the standard diet therapy plus the probiotic Visbiome. (Luna is very familiar with Visbiome these days!) At the beginning of the study, the clinical disease severity (CIBDAI) was determined to be moderate to severe in both treatment groups. Both treatment groups had significant reductions in CIBDAI scores in follow-up visits. Additionally, 83% of the dogs on the standard diet therapy and 86% of the dogs on the standard diet therapy plus Visbiome were in remission from IBD.

10. This is an
    a. observational study.
    b. **experiment.**

11. The randomization of dogs into the two treatment groups means that
    a. the sample of dogs should be representative of the population.
    b. **the treatment groups should be as similar as possible other than the treatment.**
    Common mistake: Remember that random assignment has nothing to do with how the dogs were sampled from the population.

12. Another group of researchers would like to do a follow-up study to determine if dog breed is related to the efficacy of Visbiome for dogs with IBD. In that study, the researchers will create groups by dog breed and will then randomize the dogs within each breed to the two treatment groups. In this follow-up study, the variable `breed` is a/an
    a. confounding variable.
    b. explanatory variable.
    c. **blocking variable.**
    d. response variable.

---

13. **Relative Temperature** Evan and Jamie are siblings who compete about everything. One day Evan mentioned how hot it was in was in Santa Barbara, CA. As soon as Jamie hear that, Jamie commented about how much hotter it was in Raleigh, NC. Let's see who is correct. June temperatures in Santa Barbara average 72 °F with standard deviation 2.7 °F. June temperatures in Raleigh average 76 °F with standard deviation 3.3 °F. On the day in question, it was 85 °F in Santa Barbara and 89 °F in Raleigh.

    Relatively speaking, was it hotter in Santa Barbara or in Raleigh?
    a. **Santa Barbara**
    b. Raleigh
    c. The cities were equally as hot.

    Explain your choice briefly and include work to support your reasoning.

    Santa Barbara: $z = \frac{x-\mu}{\sigma} = \frac{85-72}{2.7} = \frac{13}{2.7} = 4.815$

    Raleigh: $z = \frac{x-\mu}{\sigma} = \frac{89-76}{3.3} = \frac{13}{3.3} = 3.939$

    The temperature in Santa Barbara was 4.815 standard deviations above average, while the temperature in Raleigh was 3.939. Relatively speaking, the temperature in Santa Barbara was hotter, since the $z$-score was farther from 0 than the $z$-score for Raleigh.

---

14. **Support for Nuclear Energy** In 2022, 51% of U.S. adults favored the use of nuclear energy as one of the ways to provide electricity for the United States. Is the percentage of adults in the United States favored the use of nuclear energy as one of the ways to provide electricity for the United States different in 2023? In a random sample of 1009 U.S. adults in March 2023, 552 said that they favor the use of nuclear energy as one of the ways to provide electricity for the United States.

a. Write the hypotheses we should use to test whether the proportion of U.S. adults who favor the use of nuclear energy has changed from the proportion in 2022. Be sure to define the parameter.
$$H_0: p = 0.51 \text{ versus } H_A: p \neq 0.51$$
where $p$ represents the population proportion U.S. adults who favor the use of nuclear energy as one of the ways to provide electricity for the United States in 2023

b. Calculate the proportion of U.S. adults in the sample who favor the use of nuclear energy as one of the ways to provide electricity for the United States in 2023. Be sure to include the correct notation.
$$\hat{p} = \frac{552}{1009} = 0.547$$

c. We will use blue poker chips to represent U.S. adults who favor the use of nuclear energy and yellow poker chips to represent U.S. adults who oppose the use of nuclear energy. We will use a total of 100 poker chips to simulate the null hypothesis model.
   i. How many of the 100 poker chips should be blue?
      51 (we need the proportion of blue to match what we assume in the null hypothesis)

   ii. What does drawing one poker chip from the bag represent?
      One poker chip represents one U.S. adult (and whether they favor or oppose the use of nuclear energy as one of the ways to provide electricity for the United States).

   iii. Now that we have the poker chip allocation taken care of, we want to complete one repetition of the simulation. We will begin by mixing our poker chips inside the bag. Based on what was observed in the study, how many poker chips should we draw from the bag to complete one repetition of the simulation?
      We need to draw 1009 poker chips from the bag (with replacement) to represent the 1009 U.S. adults in the sample.

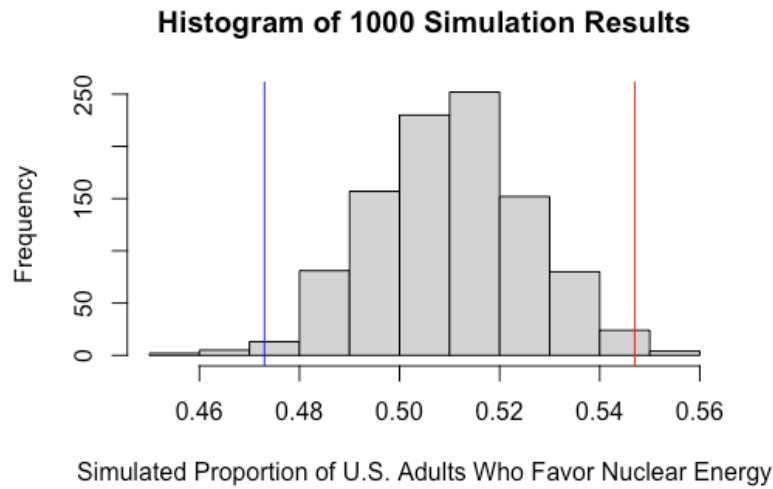d. Assuming the null hypothesis is true, where should the null distribution be centered?
   i. 0.50
   ii. **0.51** (the null distribution/distribution of the simulated proportions will always be centered at the proportion we assume to be true in the null hypothesis)
   iii. 0.52
   iv. 0.547

e.  Here is a histogram of 1000 repetitions of the simulation. The number of simulations with values at or to the left of the blue line is 9; the number of simulations with values at or to the right of the red line is 6.

### Histogram of 1000 Simulation Results



Simulated Proportion of U.S. Adults Who Favor Nuclear Energy

What is the estimated $p$-value for the hypothesis test?

The estimated $p$-value is the area in both tails since we have a not equal to alternative hypothesis. Simulated proportions at or above 0.547 and at or below 0.473 represent getting results as extreme as or more extreme than our observed results. The estimated $p$-value is

$$\frac{9+6}{1000} = \frac{15}{1000} = 0.015$$

f.  Based on the results of the hypothesis test and the estimated p-value above, write a conclusion for this hypothesis test at the $\alpha = 0.05$ level. Be sure to include:
   • Your decision to **reject** or **fail to reject** the null hypothesis (and why)
   • Whether or not the results are statistically significant
   • A clear interpretation of the result in the context of the study

Since the estimated p-value of 0.015 is less than $\alpha = 0.05$, we reject the null hypothesis. The results are statistically significant. Our analysis suggests that the percentage of U.S. adults who favor the use of nuclear energy as one of the ways to provide electricity for the United States in 2023 is different from the 51% reported in 2022.