# STA 100 Homework 1

## Due 11:59 pm Friday, July 7 onto Gradescope

1. For each of the following cases, state whether the study should be observational or experimental. Why?
    - A study investigating the association between smoking and lung cancer by analyzing existing medical records of patients.
    - A study examining the effectiveness of a new medication by randomly assigning participants to either the medication group or a placebo group.
    - A study investigating the relationship between exercise and heart health by observing and measuring exercise habits and heart health indicators of a large population over a specific period.
    - A study evaluating the impact of a new teaching method by dividing a group of students into two classes: one using the new method and the other using the traditional method, and then comparing their academic performance.
    - A study examining the effects of environmental pollution on respiratory health by comparing the respiratory health outcomes of individuals living in polluted and non-polluted areas.
2. For the following random variables, specify if they are nominal, ordinal, continuous, or discrete.
    - Number of outbreaks of pneumonia at UC Davis.
    - The amount of money you can physically hand to another person.
    - The shape of a particular cell.
    - The socioeconomic status of individuals.
3. A trucking company records the lifetimes of a certain tire brand, measured in months with the following results:

    $$21, 36, 28, 28, 3, 48, 35, 36, 22, 15$$

    Use this sample data to solve the following (show all steps in your calculations):
    - Calculate the mean.
    - Calculate the median.
    - Calculate the variance.
    - Calculate the standard deviation.
4. Continue with the data in Problem 3.
    - Calculate the first quartile.
    - Calculate the third quartile.
    - Calculate the lower fence for outliers.
    - Calculate the upper fence for outliers.
    - Identify any outliers in the dataset.

5. A random sample of 100 students was taken, and the number of times a week the student exercised was recorded:

| # of Times Exercised | 0 | 1 | 2 | 3 | 10 |
|---|---|---|---|---|---|
| Freq | 20 | 40 | 24 | 14 | 2 |

That is, 20 students did not exercise, 40 exercised 1 time a week, 24 exercised twice, etc.

- Find the average number of times a student exercised.
- Find the median of the number of times a student exercised.
- Find the variance of the number of times a student exercised.
- Find the standard deviation of the number of times a student exercised.

6. Continue with the data in Problem 5.
   - Calculate the first quartile for time number of times a student exercised.
   - Calculate the third quartile for time number of times a student exercised.
   - Calculate the lower fence for outliers.
   - Calculate the upper fence for outliers.
   - Identify all outliers in the dataset.
7. Answer the following questions with TRUE or FALSE. It is good practice to explain your answers.
   - The standard deviation must always be larger than the mean.
   - Outliers do not have a strong influence on the range of a dataset.
   - The 90th percentile is the value for which 10% of the data lies above it.
   - Outliers have a strong influence on the mean of a dataset.
8. Consider the following contingency (frequency) table, in which two species of mice were tested for a specific parasite:

| | Infected | Not Infected |
|---|---|---|
| Species 1 | 38 | 16 |
| Species 2 | 20 | 35 |

   - Estimate the probability that a randomly selected mouse was species 1.
   - Estimate the probability that a randomly selected mouse was infected.
   - Estimate the probability that a randomly selected mouse was both infected and species 1.
   - Estimate the probability that a randomly selected mouse was not infected and species 2.
9. Continue with the data from Problem 8.
   - If a mouse was species 1, what is the estimated probability they were infected?
   - If a mouse was species 2, what is the estimated probability they were infected?
   - What is the estimated probability that an infected mouse was species 1?
   - What is the estimated probability that an infected mouse was species 2?
   - Are the events that a mouse is species 1 and a mouse was infected independent?

10. R is necessary for the remaining questions. We will be using R Studio to perform some basic data analysis. The dataset we will be exploring is the famous Edgar Anderson's Iris Data. It provides the measurements (in cm) of the variables sepal length, sepal width, petal length, and petal width for 50 flowers from each of 3 species of iris. The three species are iris setosa, versicolor, and virginica.

    - The Iris Data is included in R and is called iris. Visualize the structure of the Iris Data by using the command *head()* on the iris data. Report the sepal length, petal width, and species of the 6th observation in the dataset.

    Note: The head() command displays only the first six observations, but that does not mean there are only 6 observations in the dataset! As a reminder, you can use the print() command if you want to display all of the observations in the data.

    - Print any plots you produce, and include them with your homework submission. You may write down your numerical results.
        - Find the mean, standard deviation, and variance for sepal length.
        - Find the five number summary for sepal length. Also calculate the IQR based on this five number summary.
        - Plot a boxplot of sepal length. Make sure your boxplot is appropriately labeled and titled. Use this boxplot to determine if there are any outliers in the data.
        - Plot 3 side-by-side boxplots of sepal length, split by species. Describe how the species compare to one another. Also comment on the spread of the sepal length of each species, and point out any outliers.
        - Plot a histogram of sepal length over all species. Use the histogram to describe the skewness and modality of the sepal length data.