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) A study investigating the association between smoking and lung cancer by analyzing existing medical records of patients.

   (b) A study examining the effectiveness of a new medication by randomly assigning participants to either the medication group or a placebo group.

   (c) 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.

   (d) 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.

   (e) 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.

   (a) Number of outbreaks of pneumonia at UC Davis.

   (b) The amount of money you can physically hand to another person.

   (c) The shape of a particular cell.

   (d) The socioeconomic status of individuals.

3. A random sample of 100 students was taken, and the number of times a week the student exercised was recorded: That is, 20 students did not exercise, 40 exercised 1 time a week, 24 exercised twice, etc.

# of times exercised	0	1	2	3	10
Frequency	20	40	24	14	2

   (a) Find the average number of times a student exercised.

   (b) Find the median of the number of times a student exercised.

   (c) Find the variance of the number of times a student exercised.

   (d) Find the standard deviation of the number of times a student exercised.

   (e) Calculate the first quartile of the number of times a student exercised.

   (f) Calculate the third quartile of the number of times a student exercised.

   (g) Calculate the lower fence for outliers.

   (h) Calculate the upper fence for outliers.

   (i) Identify all outliers in the dataset.

4. Answer the following questions with TRUE or FALSE. It is good practice to explain your answers.

   (a) The standard deviation must always be larger than the mean.

   (b) Outliers do not have a strong influence on the range of a dataset.

(c) The 90th percentile is the value for which 10% of the data lies above it.

(d) Outliers have a strong influence on the mean of a dataset.

5. 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

(a) Estimate the probability that a randomly selected mouse was species 1.

(b) Estimate the probability that a randomly selected mouse was infected.

(c) Estimate the probability that a randomly selected mouse was both infected and species 1.

(d) Estimate the probability that a randomly selected mouse was not infected and species 2.

6. Continue with the data from Problem 5.

(a) If a mouse was species 1, what is the estimated probability they were infected?

(b) If a mouse was species 2, what is the estimated probability they were infected?

(c) What is the estimated probability that an infected mouse was species 1?

(d) What is the estimated probability that an infected mouse was species 2?

(e) Are the events that a mouse is species 1 and a mouse was infected independently?

7. For a particular disease, the probability of the disease is 0.04. If someone has the disease, the probability they test positive is 0.95. If they do not have the disease, the probability they test negative is 0.99.

(a) Estimate the probability someone both tests positive and has the disease. Hint: Rule (7).

(b) Estimate the probability that someone tests positive. Hint: Rule (8).

(c) Estimate the probability that if someone tested positive, they have the disease.

(d) Estimate the probability that if someone tests negative, they do not have the disease. Hint: Rule (7) and Rule (8).

8. A random variable X describes the number of hairs Yidong finds in his hairbrush each morning after brushing. X has the probability distribution given below.

i	1	2	3	4	5
$P(X = i)$	0.05	0.1	0.5	0.3	0.05

(a) How many hairs can Yidong expect to lose on a given morning? What is the standard deviation?

(b) Over 10 days, what is the probability that Yidong loses at least 2 hairs per day? Assume each of the days are independent. Hint: Rule (6).

(c) Over those 10 days, what is the probability that Yidong loses 5 hairs at least one day over 10 days? Hint: Setup a binomial random variable.

9. In the United States, 37% of the population has type $O+$ blood. Consider taking a simple random sample of size 12. Let X denote the number of persons in the sample with type $O+$ blood. Find

(a) $P(X = 0)$.

(b) $P(X \geq 11)$.

(c) $P(4 < X \leq 6)$.

(d) The mean and variance of X.

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. Attach source codes and any plots you produce to your homework submission. You may write down your numerical results.

    (a) 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.

    (b) Find the mean, standard deviation, and variance for sepal length.

    (c) Find the ve number summary for sepal length. Also calculate the IQR based on this ve number summary.

    (d) 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.

    (e) 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.

    (f) Plot a histogram of sepal length over all species. Use the histogram to describe the skewness and modality of the sepal length data.