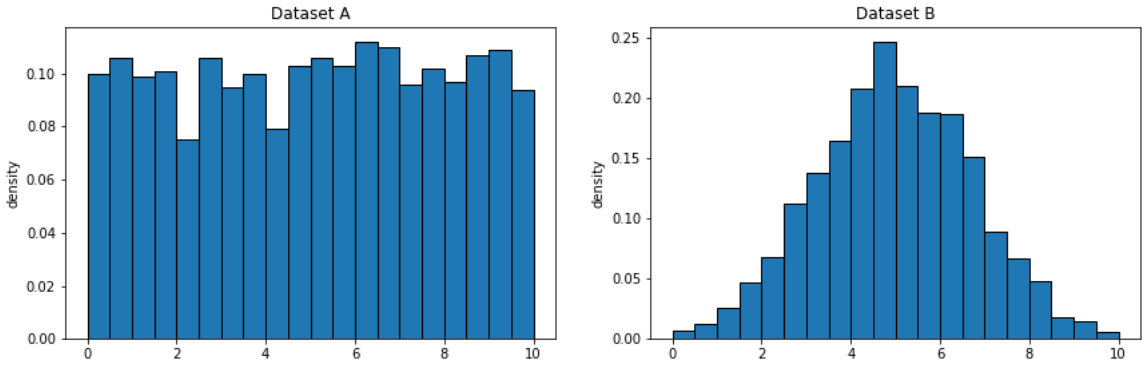
Statistics for Data Science, Final Assessment

Answer the questions in a text document named <yourname>_stats_assessment.txt.
Email completed assessments to michael.holloway@nashvillesoftwareschool.com

1. When performing exploratory data analysis, which of the following types of plots might be useful for understanding the distribution of a single variable in a dataset? There may be multiple correct choices.
    A. A bar plot
    B. A box plot
    C. A histogram
    D. A scatterplot

2. While calculating descriptive statistics for a dataset, you notice that the mean is significantly higher than the median. Which of the following might explain this discrepancy? There may be multiple correct choices.
    A. There is at least one large outlier.
    B. The dataset is approximately normally distributed.
    C. The dataset is right-skewed.
    D. The dataset is left-skewed.

3. Consider the following two histograms:



    (a) Choose the correct statement about the **range** of these two datasets.
        A. The range of dataset A is likely larger than the range of dataset B.
        B. The range of dataset A is likely smaller than the range of dataset B.
        C. The range of dataset A is about the same as the range of dataset B.
        D. It's impossible to say anything about the range based on just the histogram.

    (b) Choose the correct statement about the **standard deviation** of these two datasets.
        A. The standard deviation of dataset A is likely larger than the standard deviation of dataset B.
        B. The standard deviation of dataset A is likely smaller than the standard deviation of dataset B.
        C. The standard deviation of dataset A is about the same as the standard deviation of dataset B.
        D. It's impossible to say anything about the standard deviation based on just the histogram.

4. You conduct a poll of Nashvillians and ask, among other questions, what their annual income is and how many times they have attended the Grand Ole Opry.
   (a) Is annual income a *discrete* or *continuous* variable?
   (b) Is the number of times a person has attended the Grand Ole Opry a *discrete* or a *continuous* variable?

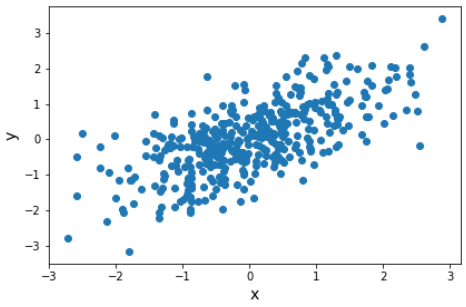
5. A study shows that the distribution in heights for both European women and American women are both approximately normal with the same mean. However, the standard deviation in European women's heights is larger than the standard deviation in American women's heights. Based on the results of this study, which of the following are correct. There may be multiple correct answers.
   A. The average European woman is taller than the average American woman.
   B. There is a higher proporion of European women who are very tall (say, six feet or taller) than the proportion of American women who are very tall.
   C. There is a higher proportion of American women who are very short (say, under 4.75 feet) than the proportion of European women who are very short.
   D. If we take a sample of 25 European women and a sample of 25 American women, we can expect that the sample mean for the European women will be larger than the sample mean for the American women.

6. When studying a dataset covering the years 1980 through 2020, you notice that cell phone usage and brain cancer rates are positively correlated.
   (a) Which of the following does this positive correlation imply?
      A. Years with above average cell phone usage tended to also have above average rates of brain cancer.
      B. Years with above average cell phone usage tended to have below average rates of brain cancer.
      C. A positive correlation does not imply any kind of association between cell phone usage and rates of brain cancer.
   (b) Based on this correlation, you conclude that people should limit cell phone usage to decrease chances of brain cancer. Describe what might be wrong with this reasoning.

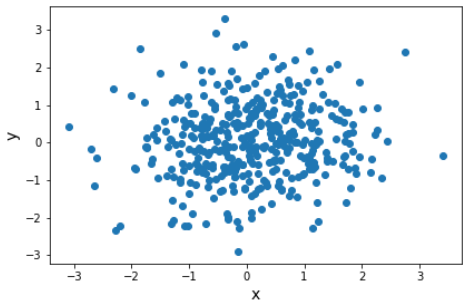
7. You have a dataset containing variable A and variable B. You calculate the correlation between variable A and variable B and get a value of 0.06. Can you conclude that there is likely no relationship at all between these variables? Why or why not?

8. For one of the following plots, the correlation between x and y of -0.7. Which one?

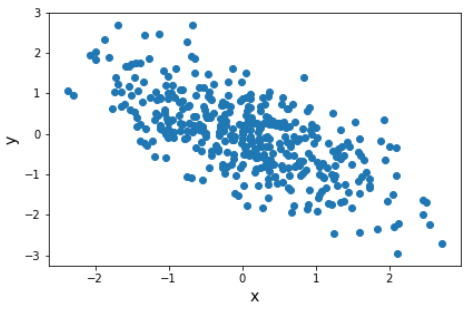
A.



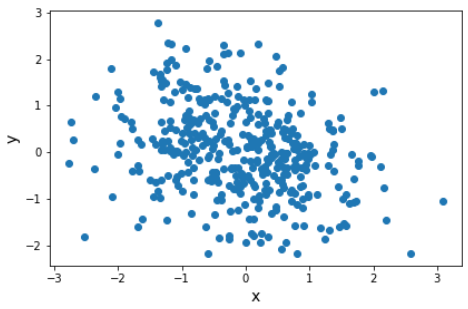
B.



C.



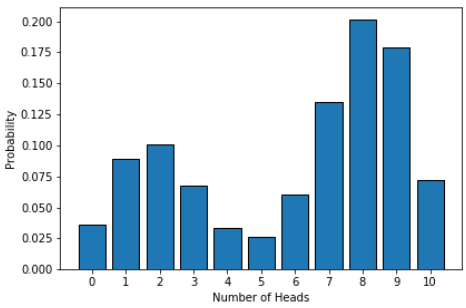
D.



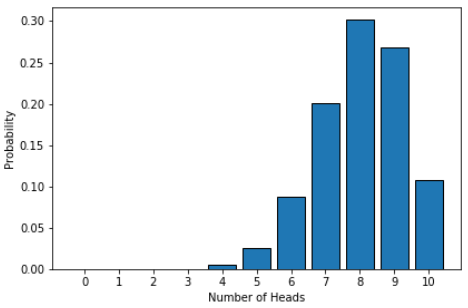
9. You have a sample from a population which you think is (approximately) normally distributed. How can you check the plausibility that the population is normally distributed?

10. You have a **bent** coin which has a probability of landing on heads of 0.8. Which of the following plots gives the probability mass function for the number of times the coin lands on heads in 10 flips?

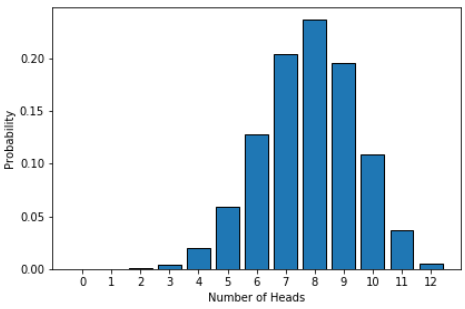
A.



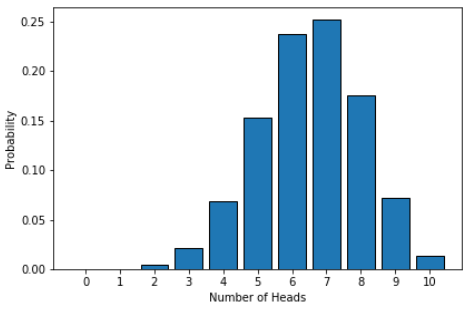
B.



C.



D.



11. For a given discrete random variable, what is the difference between its probability mass function and its cumulative distribution function?

12. **True or False:** (You do not need to compute the probabilities to answer this question.) The probability of a fair coin landing on heads 6 or more times out of 10 flips is approximately the same as the probability of a fair coin landing on heads 60 or more times out of 100 flips. Justify your answer.

13. The following snippet comes from a Chattanooga Times Free Press article entitled "Polls find majority support among Tennesseans for voting by mail".

    Source: https://www.timesfreepress.com/news/local/story/2020/jun/10/polls-find-majority-support-among-tennesseans/525018/

> On Tuesday, Vanderbilt University released its annual spring survey. Conducted May 5-25, the poll of 1,000 registered Tennessee voters found 57% strongly or somewhat supported voting by mail. Another 42% strongly or somewhat said they opposed it.
>
> The poll's margin of error is plus or minus 3.8%.

(a) Assume that the margin of error is based on a 95% confidence interval. If the true proportion of Tennesseans who strongly or somewhat supported voting by mail was actually 60%, would the result of the poll be considered unusual (meaning that the chance of it occurring was less than 5%)?

(b) Assume that the margin of error is based on a 95% confidence interval. If the true proportion of Tennesseans who strongly or somewhat supported voting by mail was actually 62%, would the result of the poll be considered unusual (meaning that the chance of it occurring was less than 5%)?

(c) Based on the results of this poll, can we conclude that it is impossible that a majority of Tennesseans *oppose* voting by mail? Why or why not?

14. When constructing a confidence interval for the mean, how would the margin of error differ if using a sample size of 100 versus a sample size of 1000? Bonus: by what factor would you expect the margin of error to change when going from a sample size of 100 to one of 1000?

15. You think that there is a chance that people who have pets tend to live longer than those who do not. You plan to run a hypothesis test.

    (a) What should your null and alternative hypotheses be?

    (b) You conduct a hypothesis test with a significance level of 0.05. After calculating your test statistic, you find that the p-value is 0.018. Which of the following is true? There may be more than one correct choice.

        A. The probability that the null hypothesis is true is 0.018.
        B. If the null hypothesis were true, only 1.8% of samples would be at least as extreme as the one you obtained.
        C. The null hypothesis is false.
        D. You have proven that the alternative hypothesis is true.
        E. You can reject the null hypothesis at the 0.05 significance level.

16. The following snippet comes from an article entitled "Researchers suggest this blood type is most susceptible to coronavirus" posted on silive.com.

    Source: https://www.silive.com/coronavirus/2020/06/researchers-suggest-this-blood-type-is-most-susceptible-to-coronavirus.html

    > "Our genetic data confirm that blood group O is associated with a risk of acquiring Covid-19 that was lower than that in non-O blood groups, whereas blood group A was associated with a higher risk than non-A blood groups," the researchers wrote. That data is consistent with other studies that have shown O blood types seem to be less susceptible to the virus.

    The researchers indicated that there is a statistically significant risk reduction for coronavirus for those with Type O blood.

    (a) Based on this study, can we conclude that it is safe for people with type O blood to take less precautions against the coronavirus (social distancing, mask wearing, etc.)? What additional information might you like to know before giving such a recommendation?

    (b) Would knowing how small the p-value obtained in this study was help us decide how large the effect is? Meaning, for example, that if we knew the p value was 0.00000038, could would conclude that there is a large difference between the risk for those with Type O blood and those without? Why or why not?

17. For each situation, state whether *linear* regression or *logistic* regression would be more appropriate.

    Situation 1: You want to predict whether a person will have a heart attack in the next year based on body weight, calorie intake, fat intake, and age.

    Situation 2: You want to predict the number of violent crimes in an area based on median income, police budget, number of guns per capita, and population density.

18. A hydrologist builds a linear regression model to predict the volume flow of a stream (y) in gallons/minute based on the daily rainfall (x) in inches. This results in the following linear regression equation:

$$y = 1.6 + 29 \cdot x$$

    If rainfall increases by 1 inch, by how much can we expect the volume flow of the stream to increase?