Objective Questions:

1. What is the main reason for the occurrence of Simpson's Paradox?

    a. Confounding variables
    b. Small sample size
    c. Outliers
    d. Non-random sampling

2. Dr. Goel wants to understand whether open workspaces are better than traditional office spaces and asks volunteers to indicate the degree to which they agree that open workspaces boost productivity by choosing from the following options: strongly agree, agree, neutral, disagree, strongly disagree. What type of variable best captures these responses?

    a. Nominal
    b. Ratio
    c. Ordinal

3. A disease test is 98% accurate. If a person has the disease, the probability of testing positive is 0.98 (98%). If a person does not have the disease, the probability of testing negative is 0.90 (90%). In a certain population, 1% of the people have the disease. A person is selected at random and tests positive. What is the probability that the person has the disease?

    a. 0.98
    b. 0.10
    c. 0.01
    d. 0.098

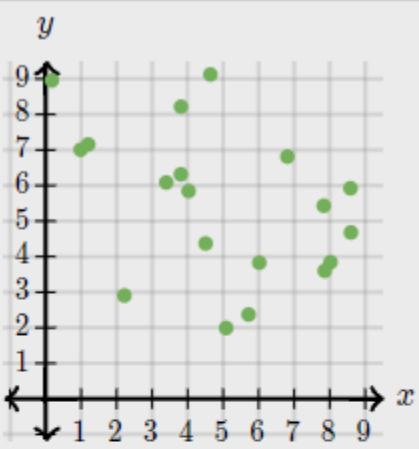
    Note: Full marks given to anyone who chose b or d

4. A person's blood pressure is measured using a sphygmomanometer (blood pressure cuff) on two separate occasions, with a period of time in between the two measurements (e.g. one week apart). The results of the two blood pressure measurements would be compared to assess the consistency of the person's blood pressure over time. Which kind of reliability measure does this test?

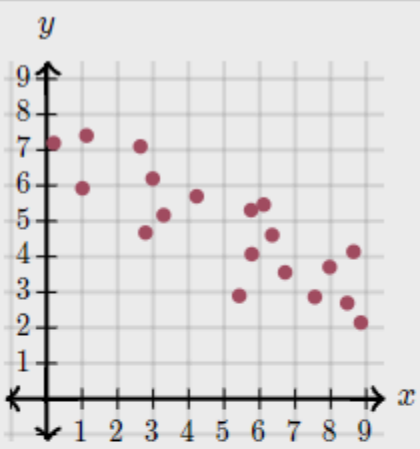
    a. Test-Retest Reliability
    b. Inter-Rater Reliability
    c. Parallel Forms Reliability
    d. Internal Consistency

5. Given the scatterplots given below, match them with their correlation coefficients:

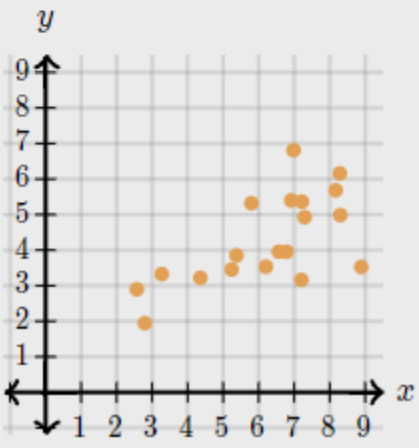
Scatterplot A



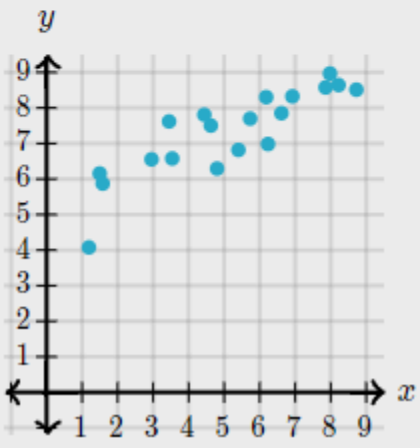
Scatterplot B



Scatterplot C



Scatterplot D



a. Scatterplot A	1. r = -0.42 (A)
b. Scatterplot B	2. r = -0.86 (B)
c. Scatterplot C	3. r = 0.65 (C)
d. Scatterplot D	4. r = 0.85 (D)

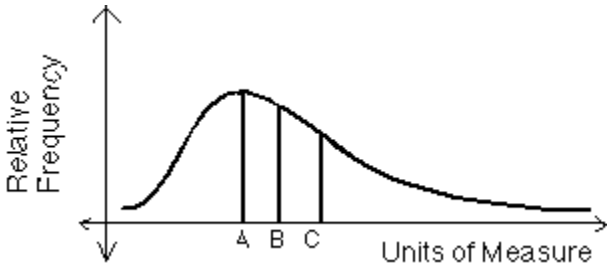
6. A histogram is a graph in which values of observations are plotted on the horizontal axis, and their density is plotted on the vertical axis.

    a. True
    b. False

7. Imagine we took a group of cigarette addicts. We recorded the number of cigarettes they consumed each day, whether they considered cutting down their consumption or not, and then split them randomly into one of two 4-week interventions; ‘hypnosis’ or ‘nicotine patches.’ After four weeks, we again recorded how many cigarettes they consumed each day. We subtracted this number from the number of cigarettes they each consumed pre-intervention to produce an intervention success score for each participant. Out of the following options, which would be the best method of looking at which intervention was the most successful, considering whether the participant wanted to cut down their consumption or not?

    a. Pie Chart Plot
    b. Funnel Chart Plot
    c. Simple Histogram
    d. Boxplot

8. Which of the following orders correctly represents the measures of central tendency for the distribution shown here?



    a. A: median, B: mode, C: mean

    b. A: mode, B: median, C: mean

    c. A: mean, B: median, C: mode

    d. A: mode, B: mean, C: median

    e. A: median, B: mean, C: mode

    f. None of these orders are correct

9. A student is conducting research on NAFLD (Non-Alcoholic Fatty Liver Diseases), where they are trying to find if specific genes are more likely to predict the onset of this disease. However, they are unable to produce satisfactory results with significant p-values. Now, instead of reporting these negative findings, they decide not to publish a paper altogether and focus on some other area. What kind of bias is this?

    a. Belief Bias
    b. Publication Bias
    c. Selection Bias
    d. Experimenter Bias

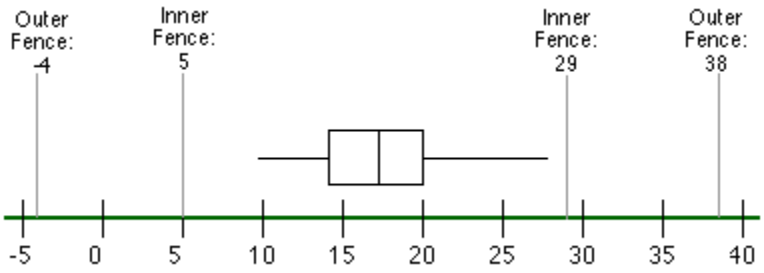
10. In which situation is it appropriate to use the mode as the preferred measure of central tendency?

    a. in reporting marks in a class quiz
    b. in reporting average selling price for homes in a community
    c. in determining what size shoes to reorder in a retail establishment
    d. when the distribution is significantly skewed to the left or to the right

11. Consider the following data set: 13, 18, 28, 28, 31, 26, 35, 20. Which measure of central tendency would change the most if the “20” would have been a “48”?

    a. median
    b. mode
    c. mean

12. Although not shown in the plot, which of the following values would be considered an outlier in this data set?



    a. -5
    b. 6
    c. 25
    d. 27
    e. None of the above

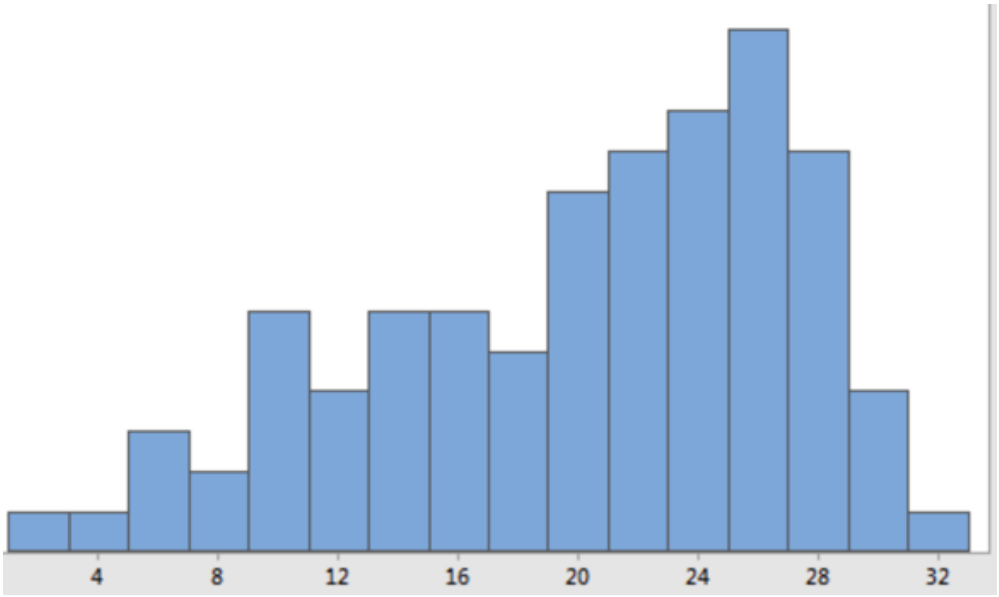
13. Data visualization tools provide an accessible way to see and understand _____ in data.

    a. Trends
    b. Outliers
    c. Patterns
    d. All of the above

14. Many laboratory memory studies use artificial lists of words. However, when analogous experiments are done using more naturalistic stimuli in real-world ecologically valid scenarios, researchers have found that the memory theories developed in the lab based on artificial lists of words generalize well to real-world scenarios. Based on this, we can conclude that artificial word list-based lab experiments have (pick the BEST option):

    a. High internal validity
    b. High external validity
    c. High ecological validity
    d. High face validity

15. Given this histogram of the population distribution of the age of chess players on a small island (assume mean = 24, standard deviation = 8), if you draw a sample of 100 ages from this population, and plot the distribution of the means of 10,000 such samples, what would you expect the resulting distribution to be?



    a. A positively skewed distribution with mean around 8
    b. A Normal distribution with mean around 8
    c. A Normal distribution with mean around 24
    d. A negatively skewed distribution with mean around 24

16. A student was conducting research about the Canonicalization of Vector Fields. Given the complexity of this topic and the student's workload, the student regularly skipped breakfast. After a few months, the student gets their paper published. Seeing this, one of their friends started skipping breakfast in hopes of getting their paper published too. Which of these is true?

    a. The friend is very likely to get their paper published
    b. The friend is not at all likely to get their paper published
    c. No inference can be drawn from this information

17. Which of the following best explains the difference between correlation and causation in the context of ice cream sales and crime rates?

    a. Correlation exists when ice cream sales and crime rates are positively correlated, while causation exists when increased ice cream sales cause a decrease in crime rates.
    b. Correlation exists when ice cream sales and crime rates are positively correlated, while causation exists when increased crime rates cause a decrease in ice cream sales.
    c. Correlation exists when ice cream sales and crime rates are positively correlated, while causation exists when there is no relationship between ice cream sales and crime rates.
    d. Correlation exists when ice cream sales and crime rates are positively correlated, while causation exists when a third variable, such as temperature, affects both ice cream sales and crime rates.

18. The much-awaited Magnus Carlsen vs. Hans Niemann chess rematch has been announced. You have been chosen to poll the audience to find whom the majority supports. However, people are ashamed to publicly support Neimann because of his questionable tactics. To get an accurate poll, you decide to use a coin to add a random chance to allow plausible deniability for the pollees:

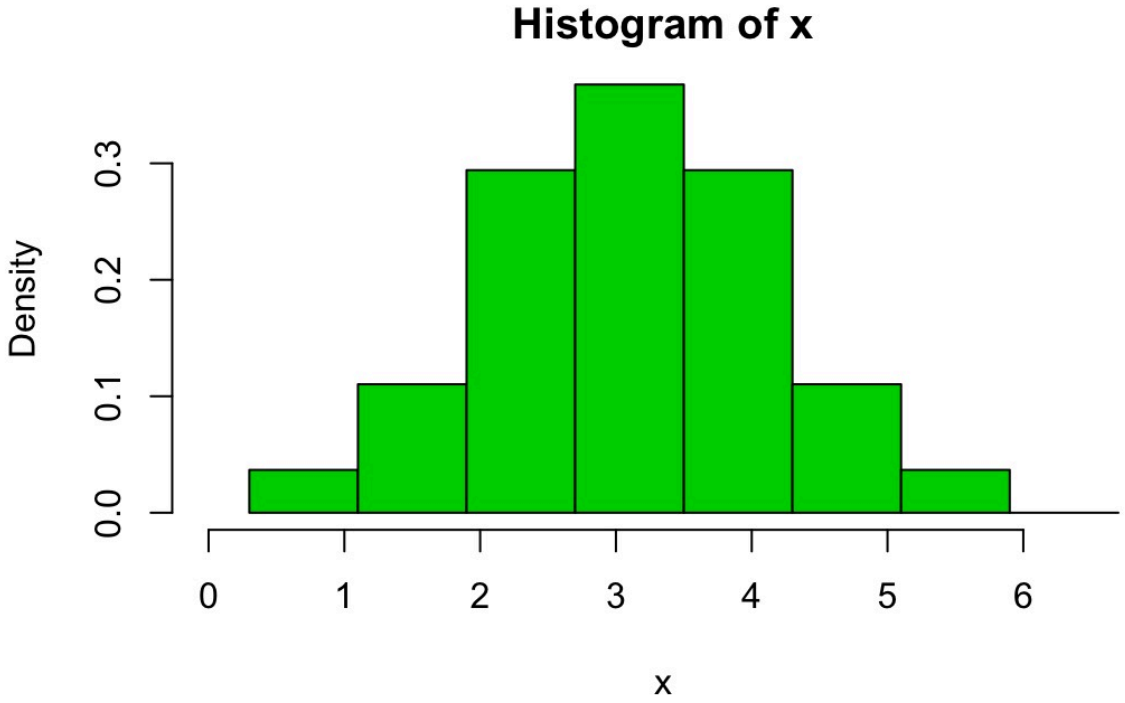
    The pollees toss the coin privately. If it lands on heads, they say Neimann regardless of their preference. If it lands on tails, they reveal their true preference.

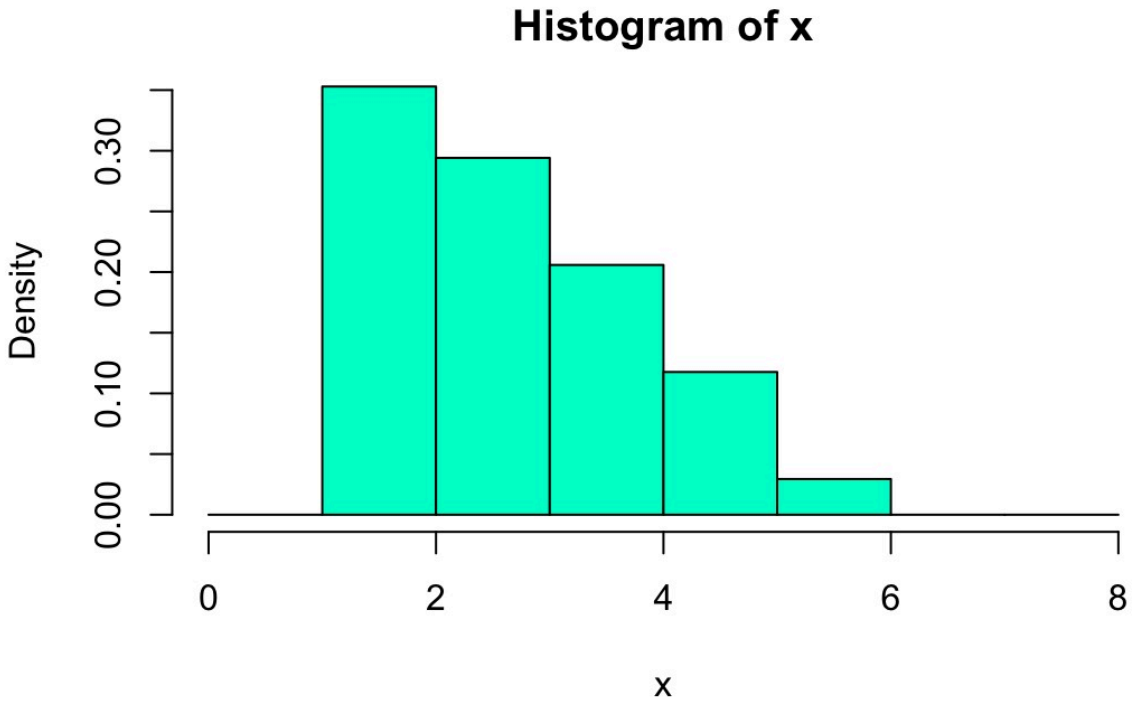
    You conduct a poll for 200 people using this method and receive the following results: 130 people vote for Niemann , while 70 people vote for Carlsen. What change should you make to this distribution to get a projection as close as possible to the actual votes?

    a. Subtract 100 from Niemann's votes
    b. Add 100 to Carlsen's votes
    c. Subtract 100 from Carlsen's votes
    d. Subtract 50 from Niemann's votes

Descriptive Questions:

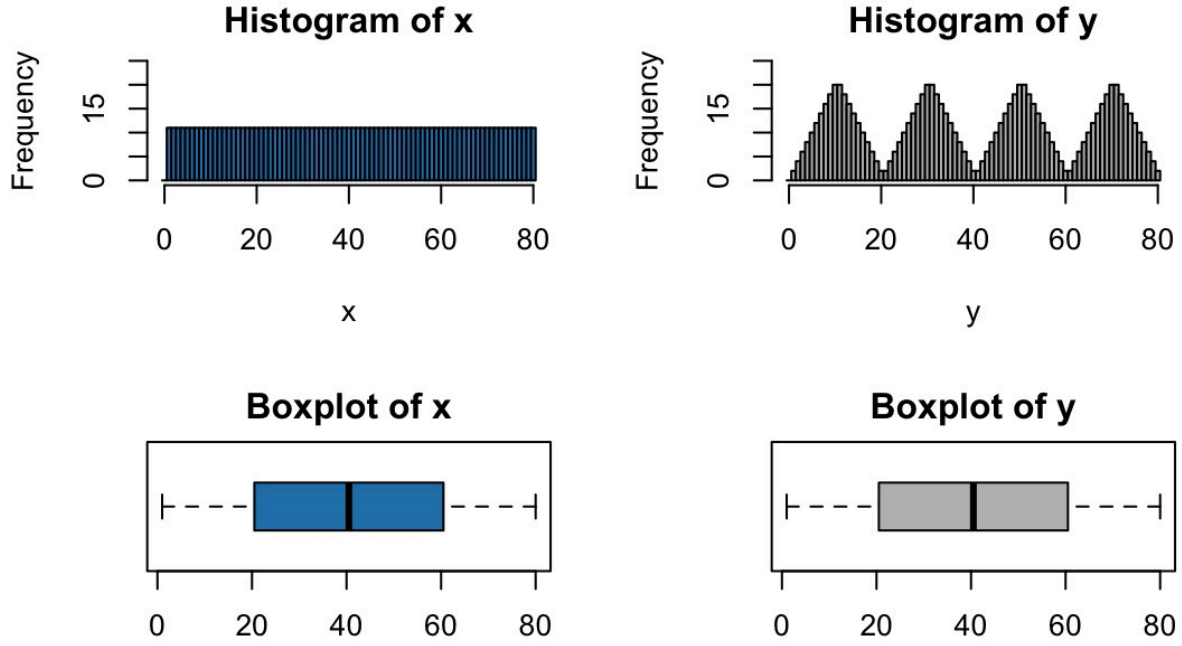
Q1. The following histograms have been generated from the same data. However, they show different distributions. Why?





Ans: The bin width has been changed. The first graph has a bin width of 0.8 while the second one has a bin width of 1.

Q2. Two groups of data (x and y) have been visualized as a boxplot and histogram. As can be seen, while the histograms show a different distribution, the boxplots are the same. Explain why.



Ans: Boxplots capture the minimum, first quartile, median, third quartile, and maximum of the data which is same for both groups of data here. Hence we have identical boxplots. However, histograms capture the frequency distribution which is different.