

SHOW YOUR WORK FOR FULL CREDIT!

Problem	Max. Points	Your Points
1-10	10	
11	5	
12	4	
13	3	
14	4	
15	18	
16	9	
17	4	
18	4	
19	14	
Total	75	

Multiple choice questions (1 point each)

For questions 1 and 2 consider the following:

The distribution of scores on a certain statistics test is strongly skewed to the right.

- Which set of measures of center and spread are more appropriate for the distribution of scores?
 - Mean and standard deviation
 - Median and interquartile range
 - Mean and interquartile range
 - Median and standard deviation
- What does this suggest about the difficulty of the test?
 - It was an easy test
 - It was a hard test
 - It wasn't too hard or too easy
 - It is impossible to tell
- Which one of the following variables is NOT quantitative?
 - weight of an elephant
 - length of a Hitchcock movie
 - time spent studying for this test
 - color of an M&M candy
- What percent of the observations in a distribution lie between the median and the third quartile Q_3 ?
 - Approximately 25%
 - Approximately 50%
 - Approximately 75%
 - Approximately 100%
- X and Y are two categorical variables. The best way to determine if there is a relation between them is
 - to calculate the correlation between X and Y.
 - to draw a scatterplot of the X and Y values
 - to make a two-way table of the X and Y values
 - all of the above
- For an exam given to a class, the student's scores ranged from 35 to 98, with a mean of 74. Which of the following is the most likely value for the standard deviation?
 - 10
 - 0
 - 13
 - 63
- The typical amount of sleep per night for college students can be assumed to be normally distributed with a mean of 7 hours and a standard deviation of 1.2 hours. From the Standard Deviation Rule we know that about 68% of the college students typically sleep between
 - 5.8 and 8.2 hours per night.
 - 4.6 and 9.4 hours per night.
 - 6 and 8 hours per night.
 - It is impossible to determine the answer from the given information.

8. A study found a correlation of $r = -0.61$ between the gender of a worker and his or her income. You may correctly conclude:
- a) women earn less than men on the average.
 - b) women earn more than men on the average.
 - c) this is incorrect because r makes no sense here.
 - d) an arithmetic mistake was made. Correlation must be positive.
9. A study of the salaries of professors at Smart University shows that the median salary for female professors is considerably less than the median male salary. Further investigation shows that the median salary for female professors is more than the median male salary in every department (English, Physics, etc.) of the university. This apparent contradiction is an example of
- a) extrapolation
 - b) Simpson's paradox
 - c) causation
 - d) correlation
10. Consider the following data set: 2 3 5 7 8 10 12 15 20 31. According to the 1.5(IQR) rule,
- a) only 2 is an outlier.
 - b) only 31 is an outlier.
 - c) both 2 and 31 are outliers.
 - d. there are no outliers.

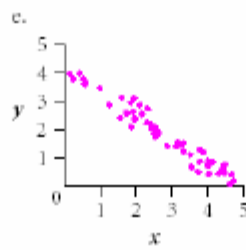
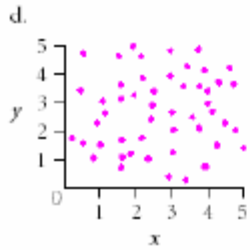
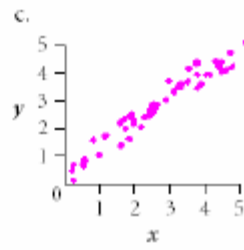
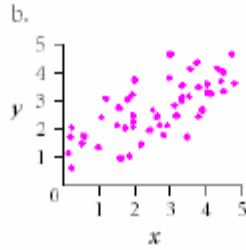
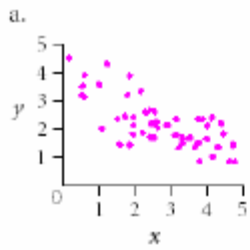
11. TRUE or FALSE? Circle either T or F for each statement below (1 point each)

- T F The only way the standard deviation can be 0 is when all the observations have the same value.
- T F If you interchange the explanatory variable and the response variable, the correlation coefficient changes.
- T F If the correlation coefficient between two variables is 0, that means that there is no possible relationship between the two variables.
- T F We can display one categorical variable graphically with a histogram, a stemplot, or a boxplot.
- T F Assuming linear relationships, correlation of 0.8 is indicates a stronger relationship than the correlation of -0.8 .

12. (4 points) A study of the size of jury awards in civil cases (such as injury, product liability, and medical malpractice) in Chicago showed that the median award was about \$8,000. But the mean award was about \$69,000. Explain how this great difference between these two measures of center could have occurred.

It must have been one or more high awards. Those high outliers greatly affect the mean, but not the median.

13. (3 points) Match each of the five scatterplots with its correlation.



-0.75
a

0.7
b

0.95
c

0
d

-0.95
e

14. (4 points) Answer the following questions:

a) Which measure of spread indicates variation about the mean?

The standard deviation

b) Which graphical display shows the median and data spread about the median?

boxplot

c) Which of the following statistical measures are *not resistant*? Circle those that are:

mean median range IQR standard deviation correlation coefficient

15. (18 points) For each of the situations described below, identify the explanatory variable and the response variable, and indicate if they are quantitative or categorical. Also, write the appropriate graphical display for each situation.

a) You want to explore the relationship nationality and party affiliation.

The explanatory variable is: nationality Categorical Quantitative

The response variable is: party affiliation Categorical Quantitative

Therefore, this is an example of Case II .

An appropriate graphical display would be: double bar chart .

b) You want to explore the relationship between work shift (morning, afternoon, night shift) and the number of accidents during the different shifts.

The explanatory variable is: work shift . Categorical Quantitative

The response variable is: number of accidents . Categorical Quantitative

Therefore, this is an example of Case I .

An appropriate graphical display would be: side-by-side boxplots .

c) You want to explore the relationship between weight of the brain and IQ scores.

The explanatory variable is: weight of brain . Categorical Quantitative

The response variable is: IQ scores . Categorical Quantitative

Therefore, this is an example of Case III .

An appropriate graphical display would be: scatterplot .

16. (9 points) In 2009, Joey Chestnut beat his previous record by eating 68 hot dogs & buns in 10 minutes, nine more than in 2008--setting new event, American, and world records. The stemplot below shows the number of hot dogs eaten by contestants in a recent hot dog eating contest. The summary statistics are also provided.

```
0|99
1|0167799
2|00479
3|123678
4|02
5|2
6|8
```

N = 24
Mean = 27.34
Median = 25.5
StDev = 14.49
Minimum = 9
Maximum = 68
Q1 = 17
Q3 = 36.5

a) How many hot dogs did the second place winner eat? **52**

b) Briefly describe the **shape** of the distribution from the stemplot.

Slightly skewed to the right

b) Check the mean and the median. Can you be pretty sure just by looking at the two values that I didn't switch them by mistake? **Explain** (Do NOT calculate them).

I did not switch them by mistake. Since the distribution is slightly skewed to the right, we can expect the mean to be slightly greater than the median. And since, $27.34 > 25.5$, they are correct.

c) Find the value of the interquartile range, and interpret it clearly in the context of the experiment.

$$IQR = Q_3 - Q_1 = 36.5 - 17 = 19.5$$

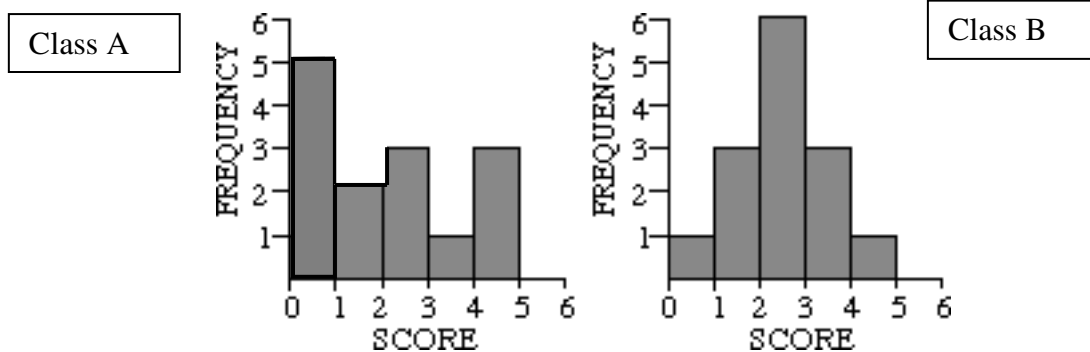
That's the range of the middle 50% of the data. Half of the contestants (the middle half) ate about between 17 and 36.5 hotdogs.

17. (4 points) Explain what the phrase “association does not imply causation” means, and give an example.

Even if two variables have a high correlation coefficient, it does not mean that the explanatory variable CAUSED the changes in the response variable.

One example: shoe size and spelling ability. Even though there is high correlation between the two variables, changing shoe size doesn't cause the changes in spelling ability. The lurking variable is age.

18. (4 points) Consider the following two distributions of scores on a quiz of 14 students in class A, and 14 students in class B.



Which distribution has the higher standard deviation and **why**?

The distribution of scores in class A has the higher standard deviation because most of the values are farther from the mean, whereas in class B most of the values are close to the mean.

19. The number of hours 21 students spent studying (in hours) for a test and their scores on that test are shown below:

Hours spent studying	0.5	1.0	2.5	4.0	4.5	5.0	5.5	5.0	6.0	6.5	7.0	7.0	8.0	3.5	3.0	1.0	2.0	5.5	6.0	3.0	7.5
Test scores	40	41	51	50	64	69	73	75	68	93	84	90	95	50	62	51	62	69	73	58	52

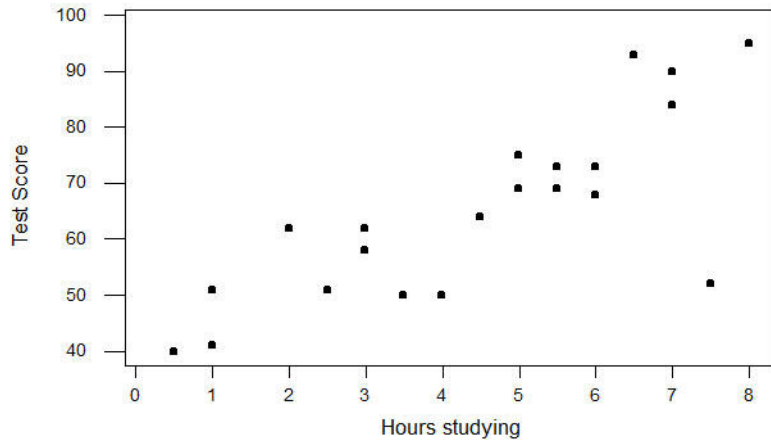
The mean hours spent studying is: 4.48 hours

The standard deviation of the hours spent studying is: 2.25 hours

The mean test score is: 65.24

The standard deviation of the test scores is: 16.15

The correlation coefficient is: 0.79



- a) Describe the scatterplot. Make sure you mention all four features.

Direction: positive

Form: linear

Strength: moderate

Outliers: one potential outlier

- b) Find the equation of the least squares line, and sketch the line on the plot. Use three decimal digits in your answers.

$$b = r \frac{s_y}{s_x} = 0.79 \frac{16.15}{2.25} = 5.670$$

$$a = \bar{Y} - b\bar{X} = 65.24 - 5.670(4.48) = 39.838$$

Equation of the least squares line: $Y = a + bX = 39.838 + 5.670X$

- c) Provide an interpretation of the slope in this context.

The slope is 5.670 scores/hour of study. That means that for every extra hour of studying the test increases by 5.670 points on average.

- d) Predict the test score for a student who studied 1.5 hours for the test.

$$Y = 39.838 + 5.670(1.5) = 48.343$$

- e) Would it be OK to use the regression line to predict the test score for a student who studied 10 hours for the test? Explain.

No, it would not be OK to use the regression line to predict the test score for a student who studied 10 hours for the test because 10 hours is out of the range of the explanatory variable. The predicted value would not be reliable since we don't know whether or not the LSRL would be a good predictor or not. This is called extrapolation. We can use the line to predict the response variable for the 0.5-8 hours of studying.

- f) Does the data support the notion that test scores are associated with the number of hours spent studying for the test? Explain briefly, giving all evidence that supports your contention.

Yes. It seems to be a positive association between the number of hours spent studying for the test and the test scores. The scatterplot shows a positive relationship, and the correlation coefficient is high. In fact, $r^2 = 0.62$, meaning that about 62% of the variation in the test scores can be explained by the number of hours spent studying, while 38% of the variation is attributable to factors other than number of hours studied.

- g) Can you name one lurking variable that might affect the test scores?

Maybe students' interest in the material. Number of hours slept. Test anxiety. Gender. IQ. Students' previous knowledge of the material, etc.