

M 140 Test 1 AName _____
(1 point)**SHOW YOUR WORK FOR FULL CREDIT!**

Problem	Max. Points	Your Points
1-10	10	
11	10	
12	3	
13	4	
14	18	
15	8	
16	7	
17	14	
Total	75	

Multiple choice questions (1 point each)

For questions 1-3 consider the following:

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

1. How would the mean and the median compare for this distribution?

- ☒ a) mean < median
- ☐ b) mean = median
- ☐ c) mean > median
- ☐ d) Can't tell without more information

When a distribution skewed to the left, the “tail” is on the left. Since the tail “pulls” the mean, the mean is less than the median.

2. Which set of measures of center and spread are more appropriate for the distribution of scores?

- ☐ a) Mean and standard deviation
- ☒ b) Median and interquartile range
- ☐ c) Mean and interquartile range
- ☐ d) Median and standard deviation

For skewed distributions the median and the interquartile range are the appropriate measures of center and spread because they are resistant to outliers.

3. What does this suggest about the difficulty of the test?

- ☒ a) It was an easy test
- ☐ b) It was a hard test
- ☐ c) It wasn't too hard or too easy
- ☐ d) It is impossible to tell

If the distribution is skewed to the left, then most of the values are to the right, toward the higher values. Thus, more students got higher scores, which suggests that the test was easy.

4. Which one of the following variables is NOT categorical?

- ☐ a) whether or not an individual has pierced ears
- ☐ b) religion
- ☒ c) length of a rattlesnake
- ☐ d) zipcode

All the other variables are categorical—the average doesn't make sense for those.

5. What percent of the observations in a distribution lie between the first quartile Q_1 and the third quartile Q_3 ?

- ☐ a) Approximately 25%
- ☒ b) Approximately 50%
- ☐ c) Approximately 75%
- ☐ d) Approximately 100%

Since Q1 is the first quartile (25%), and Q3 is the third quartile (75%), between them we have approximately 50% of the data.

6. Which of these statements are FALSE?

- a) There is a very weak linear relationship between gender and height because we found a correlation of -0.89 .
- b) Plant height and leaf height are strongly correlated because the correlation coefficient is -1.41 .
- c) If we change the units of measurements, the correlation coefficient will change.
- ☒ d) All of the above.
- e) None of the above.

All of the statements are false. a) is false because a correlation of -0.89 means a strong, not a very weak relationship. b) is false because the correlation coefficient cannot be less than -1 . c) is false because the correlation coefficient wouldn't change if we change the units of the measurement.

7. Consider two data sets A and B with more than 3 values. The sets are identical except that the maximum value of data set B is three times greater than the maximum value of set A. Which of the following statements is NOT true?

- a) The median of set A is the same as the median of set B.
- b) The mean of set A is smaller than the mean of set B.
- ☒ c) The standard deviation of set A is the same as the standard deviation of set B.
- d) The range of set A is smaller than the range of set B.

Standard deviation takes into consideration all the values. Thus, the data set B has higher standard deviation.

8. A vending machine automatically pours coffee into cups. The amount of coffee dispensed into a cup is roughly bell shaped with a mean of 7.6 ounces. From the Standard Deviation Rule we know that 95% of the cups contain an amount of coffee between 6.8 and 8.4 ounces. Thus, the standard deviation of the distribution is approximately equal to:

- a) 0.8 ounces
- ☒ b) 0.4 ounces
- c) 1.2 ounces
- d) 0.2 ounces

From the Standard Deviation Rule we know that 95% of the data is between two standard deviations below and above the mean. Thus, the standard deviation must be $(8.4 - 7.6)/2 = 0.4$

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 salaries for male and female professors are about the same 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

Simpson's paradox demonstrates that a great deal of care has to be taken when combining small data sets into a large one. Sometimes conclusions from the large data set are exactly the opposite of conclusion from the smaller sets. And this is the case here. The conclusion was reversed as we included the lurking variable, department.

10. Which of the following statements is FALSE?

- a) The only way the standard deviation can be 0 is when all the observations have the same value.
- b) If you interchange the explanatory variable and the response variable, the correlation coefficient remains the same.
- c) If the correlation coefficient between two variables is 0, that means that there is no relationship between the two variables.
- d) The correlation coefficient has no units.
- e) The standard deviation has the same units as the data.

c) is false because if $r = 0$, that means that there is no linear relationship between the two variables. It could be a very strong non-linear relationship.

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

T **F** Consider the following data set: 2 3 5 7 8 10 12 15 20 31. According to the 1.5(IQR) rule, 31 is not an outlier.

$Q1 = 5$, $Q3 = 15$, so $IQR = 10$. According to the 1.5(IQR) rule, $Q3 + 1.5(IQR) = 15 + 1.5(10) = 30$. Since 31 is higher than 30, it is considered as an outlier according to this rule.

T **F** Two variables have a negative linear correlation. That means that the response variable increases as the explanatory variable increases.

No, negative linear correlation means that as the explanatory variables increases, the response variable decreases.

T **F** If in each of the three schools in a certain school district, a higher percentage of girls than boys gave out Valentines, then it must be true for the district as a whole that a higher percentage of girls than boys gave out Valentines.

It could be, but it is not necessarily true. This goes back to Simpson's paradox.

T **F** We can display one categorical variable graphically with a histogram, a stemplot, or a boxplot.

No, these are used to display a quantitative variable.

T **F** A correlation of 0.8 indicates a stronger relationship than the correlation of -0.8 .

No, assuming linear relationship, 0.8 is as strong as -0.8 . The sign only shows the direction.

For the following five statements consider this:

Infant mortality rates per 1000 live births for the following regions are summarized with side-by-side boxplots: (www.worldbank.org)

Region 1: Asia (South and East) and the Pacific

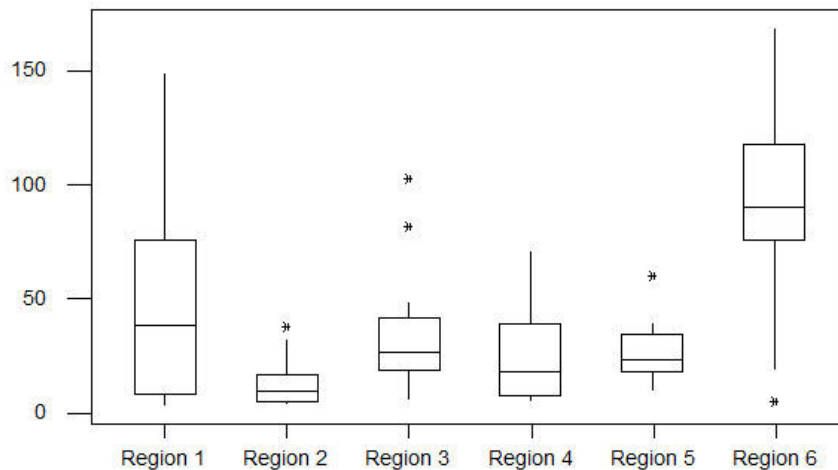
Region 2: Europe and Central Asia

Region 3: Middle East and North Africa

Region 4: North and Central America and the Caribbean

Region 5: South America

Region 6: Sub-Saharan Africa



T **F** About 75% of the countries in Region 1 have an infant mortality rate less than about 75, while about 75% of the countries in Region 6 have an infant mortality rate more than about 75.

T **F** The distribution of infant mortality rates in Region 4 is left skewed.

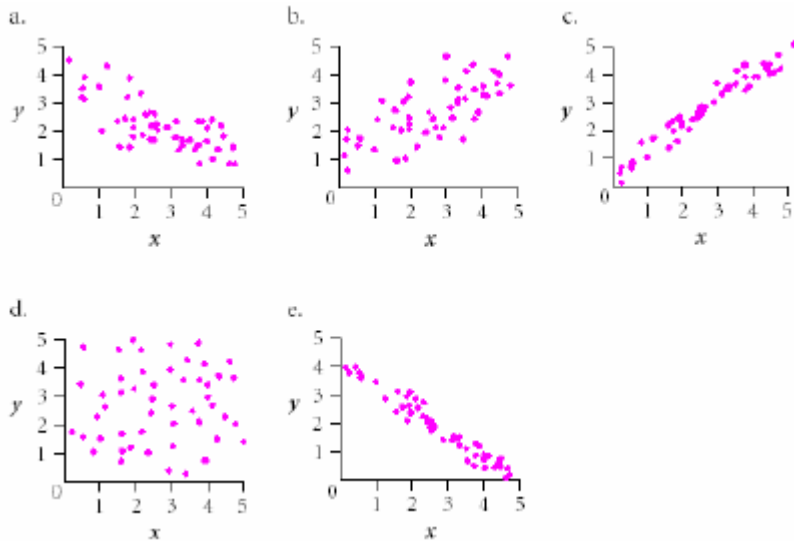
No, it's right skewed. You can see the longer "tail" in the higher values.

T **F** The median infant mortality rate in Region 4 is about the same as the first quartile of Region 3.

T **F** The variability in the infant mortality rate is the smallest in Regions 2 and 5.

T **F** The infant mortality rate in all countries except for one in Region 2 is lower than the mortality rate of 50% of the countries in Region 1.

12. (3 points) Match each of the five scatterplots with its correlation.



0.7 -0.75 -0.95 0 0.95
b. a. e. d. c.

13. (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

14. (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 between nationality and IQ scores.

The explanatory variable is: nationality .

The response variable is: IQ scores .

The explanatory variable is

Categorical

Quantitative

The response variable is

Categorical

Quantitative

Therefore, this is an example of Case I .

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

b) You want to explore the relationship between gender and party affiliation.

The explanatory variable is: gender .

The response variable is: party affiliation .

The explanatory variable is	<u>Categorical</u>	Quantitative
The response variable is	<u>Categorical</u>	Quantitative

Therefore, this is an example of Case II .

An appropriate graphical display would be: double bar graph .

c) You want to explore the relationship between the weight of the car and its mpg (miles per gallon).

The explanatory variable is: weight of the car .

The response variable is: mpg .

The explanatory variable is	Categorical	<u>Quantitative</u>
The response variable is	Categorical	<u>Quantitative</u>

Therefore, this is an example of Case III .

An appropriate graphical display would be: scatterplot .

15. (8 points) Thirty students in an experimental psychology class used various techniques to train rats to move through a maze. The running times of the rats (in minutes) are recorded at the end of the experiment. The following stemplot shows the results, where 3 | 1 represents 3.1 minutes. The summary statistics are also provided.

```
0|6
1|0167799
2|0047
3|123678
4|02455
5|236
6|0
7|6
8|0
9|1
```

N = 30
Mean = 3.698
Median = 3.45
StDev = 2.118
Minimum = 0.6
Maximum = 9.1
Q1 = 1.9
Q3 = 4.5

a) Briefly describe the shape of the distribution from the stemplot.

Skewed to the right

- b) Check the mean and the median. What do you think, did I switch them to trick you, or do they seem to be correct? Explain.

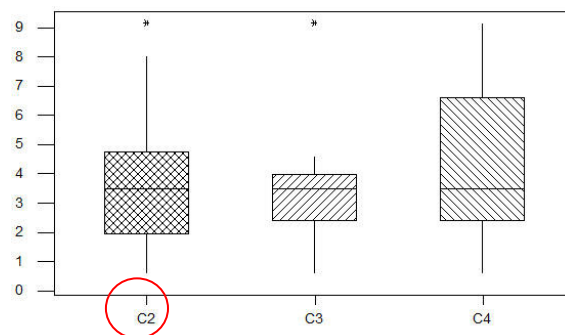
They are correct. When the distribution is skewed to the right, the mean is usually higher than the median. So I didn't switch them.

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

$$\text{IQR} = Q3 - Q1 = 4.5 - 1.9 = 2.6$$

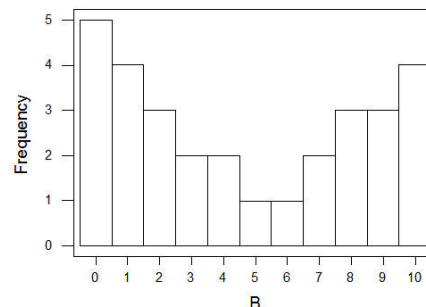
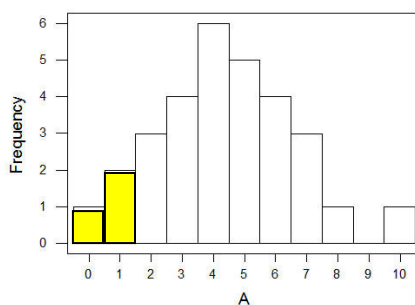
Interquartile range is the range of the middle 50% of the data. In this example, it means that 16 of the rats completed the maze between 1.9 and 4.5 minutes. Seven rats completed the maze below 1.9 minutes, and seven completed it in more than 4.5 minutes.

- d) Which one of the following boxplots represents the above data? Circle your answer.



9.1 is a suspected outlier. According to the 1.5(IQR) rule it is an outlier. $Q3 + 1.5(\text{IQR}) = 4.5 + 1.5(2.6) = 8.4$. Since 9.1 is above 8.4, it is an outlier. That leaves C2 or C3 to be the correct answers. But since the next highest number in the list is 8.0 minutes, the correct boxplot must be C2.

16. (7 points) Consider the following two distributions. The first one (A) shows the distribution of the number of houseplants owned by a sample of 30 households in Los Angeles. The second one (B) shows for a sample of 30 freshmen the distribution of the number of girlfriends/boyfriends they have ever had.



- a) What percent of households have one houseplant or none?

One household has no plants, and two households have 1 plant (see yellow highlights on the graph). That is 3 out of 30 households, $3/30 = 10\%$

- b) Which distribution has the lower standard deviation and why?

A has the lower standard deviation because the values are closer to the mean on average. Most of the values are around the center, less and less are farther from the center.

- c. Select the statement below that gives the most complete and correct statistical description of the graph A.
- A. The bars go from 0 to 10, increasing in height to 4, then decreasing to 10. The tallest bar is at 4. There is a gap between 8 and 10.
 - B. The distribution is approximately normal, with a mean of about 4 and a standard deviation of about 1.
 - C. Most households seem to have about 4 houseplants, but some have more or less. One household has 10 plants.
 - D. The distribution of the number of houseplants is somewhat symmetric and bell-shaped, with one possible outlier at 10. The typical number of houseplants owned is about 4, and the overall range is 10 plants.**

A description of a distribution should always be in context. So answers A and B are out. Between C and D, D is the most complete and correct description.

17. The number of hours 21 students spent studying (in hours) for a test and their scores on that test recorded.

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

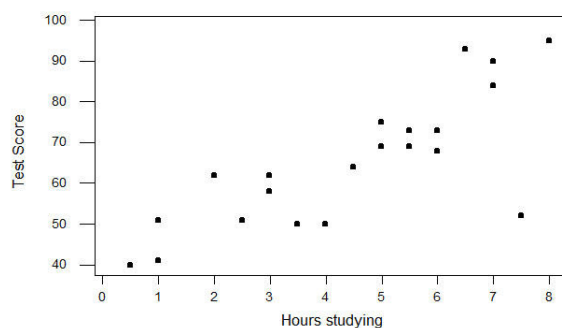
The mean hours spent studying is: 4.48 hours

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

The mean test scores 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.

Form: linear

Direction: positive

Strength: moderately strong

Outliers: maybe one at about (7.5, 50)

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

The equation of the LSRL is $Y = a + bX$, where $b = r \frac{s_y}{s_x}$, and $a = \bar{Y} - b\bar{X}$.

In this case, $b = r \frac{s_y}{s_x} = 0.79 \frac{16.15}{2.25} = 5.670$ and $a = \bar{Y} - b\bar{X} = 65.24 - 5.670(4.48) = 39.838$

Thus, the equation of the LSRL is $Y = 39.838 + 5.670X$

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

The slope is 5.670 scores/hour of study. That means that students who studied an extra hour for the test got 5.670 more points on average on their tests.

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

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

Using the LSRL, we can predict that a student who studied 1.5 hours for the test got about 48 points on the test.

- 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. That would be extrapolation.

- 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.