M 140 **Test 1** **A** Name____________________

SHOW YOUR WORK FOR FULL CREDIT!

| Problem | Max. Points | Your Points |
|---|---|---|
| 1-10 | 10 | |
| 11 | 3 | |
| 12 | 4 | |
| 13 | 3 | |
| 14 | 10 | |
| 15 | 14 | |
| 16 | 10 | |
| 17 | 7 | |
| 18 | 4 | |
| 19 | 4 | |
| Total | 60 | |

## 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 left.**

**1.** 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

**2.** 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

**3.** Which one of the following variables is NOT categorical?
- a) whether or not an individual has pierced ears
- b) religion
- [✓ c)] diameter of sequoia trees
- d) zipcode

**4.** Which one 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 possible relationship between the two variables.
- d) The correlation coefficient has no units.

**5.** X and Y are two categorical variables. The best way to determine if there is a relation between them is
- a) to calculate the correlation between X and Y.
- b) to draw a scatterplot of the X and Y values
- [✓ c)] to make a two-way table of the X and Y values
- d) all of the above

**6.** 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?
- a) -10
- b) 0
- c) 63
- [✓ d)] 13

**7.** 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 95% of the college students typically sleep between
- a) 5.8 and 8.2 hours per night.
- [✓ b)] 4.6 and 9.4 hours per night.

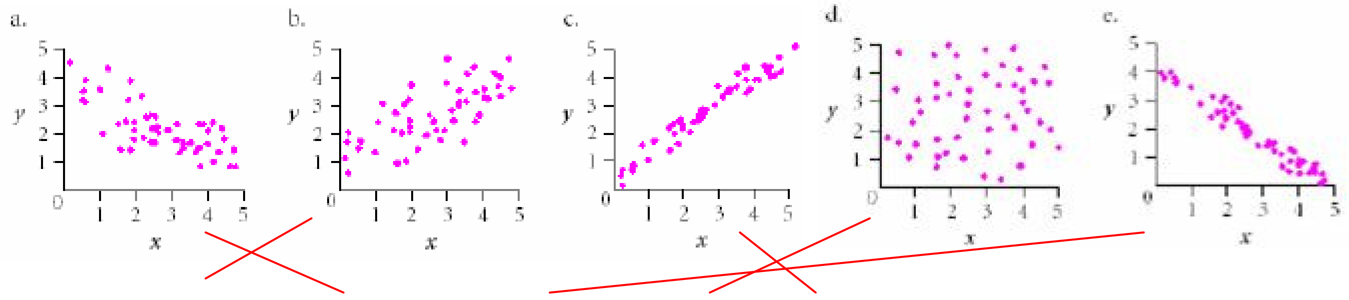
c) 6 and 8 hours per night.
d) It is impossible to determine the answer from the given information.

**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 that:
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 31 is an outlier.
b) only 2 is an outlier.
c) both 2 and 31 are outliers.
d. there are no outliers.

**11. (3 points) Match each of the five scatterplots with its correlation.**



**0.7** **-0.75** **-0.95** **0** **0.95**

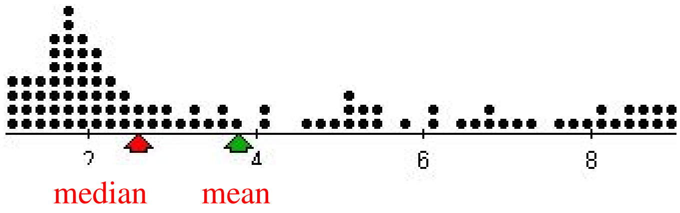
**12. (4 points) Answer the following questions:**

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

**b) Which graphical display shows the five-number summary?**
boxplot

**c) Which of the following statistical measures are *resistant*? Circle those that are:**
mean [✓ median] range [✓ IQR] standard deviation correlation coefficient

**13. (3 points) Given a dotplot below with two arrows. Indicate on the graph which one of the two arrows marks the value of the mean, and which one marks the median. Explain your decision.**



Since the mean is sensitive to outliers, it is always “pulled by the tail”.

**14. (10 points) For each of the situations described below, write the appropriate graphical display(s). You need to use ALL of these once (that means you need to write more than one graphs at some parts):**

**histogram scatterplot dotplot pie chart double bargraph stemplot**

**side-by-side boxplots bargraph**

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

Graphical display(s): ________scatterplot___________________________________.

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

Graphical display(s):________double bargaph_______________________________________.

**c) You want to explore the distribution of lengths of adult snakes living in California.**

Graphical display(s):______histogram, dotplot, stemplot___________________________.

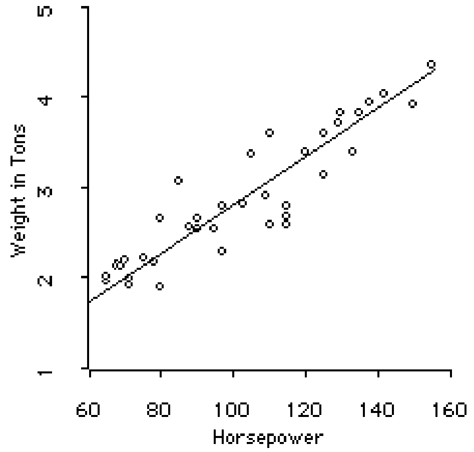
**d) You want to explore the relationship between work shift (morning, afternoon, night shift) and the number of accidents during the different shifts.**

Graphical display(s):______side-by-side boxplot___________________________________.

**e) You want to explore the distribution of taste preferences for a new soft drink (awful, OK, good, excellent)**

Graphical display(s):_______bargraph, pie chart__________________________________.

**15. (14 points) Given below is a scatterplot showing the relationship between the horsepower of a car and the weight of the car.**



| | Horsepower | Weight |
|---|---|---|
| Mean | 116 | 3.30 |
| Standard deviation | 24 | 0.38 |

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

Form: linear  Direction: positive
Strength: strong  Outliers: no outliers

b) Guess the correlation coefficient (circle the best answer):

$-0.94$  0.13  1.2  [✓ 0.9]  $-0.23$  1

c) The line drawn on the graph is called the __least squares regression line______________ .

Find the equation of the line (Y = a + bX) using the statistics given above in the table, and the correlation coefficient you guessed in part (b). Use two decimal digits in your answers.

$$b = r\frac{s_y}{s_x} = 0.9\frac{0.38}{24} = 0.01425 \approx 0.01$$

$$a = \overline{Y} - b\overline{X} = 3.30 - 0.01(116) = 2.14$$

The equation of the least squares regression line is $Y = 2.14 + 0.01X$

d) Provide an interpretation of the slope in this context.

The slope is 0.01. That is, when the horsepower of a car increases by one unit, the weight of the car increases by about 0.01 tons on average.

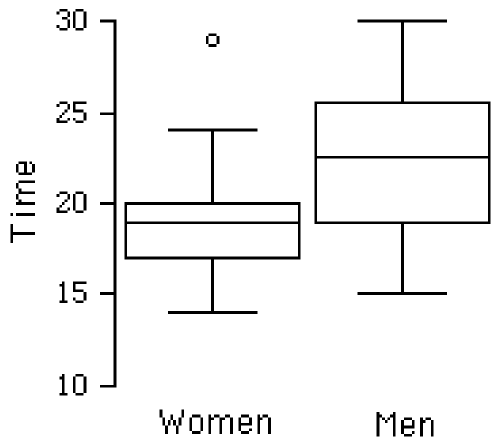
e) A car has 150 horsepower. Using the equation from part (c), predict the weight of the car.

$Y = 2.14 + 0.01(150) = 3.64$ tons

e) Would it be OK to use the line to predict the weight of a car with 50 horsepower ? Explain.

No, 50 is outside of the range of the explanatory variable. It is not reliable to use the regression line in this case—extrapolation.

**16. (10 points)** Students in Introductory Statistics were presented with a page containing 30 colored rectangles. Their task was to name the colors as quickly as possible; their times were recorded. We can compare the scores for the 16 men and 31 women who participated in the experiment by making separate box plots for each gender.



Are the following statements true or false?

[✓ T] F The median time for women is about the same as the lower quartile for men.

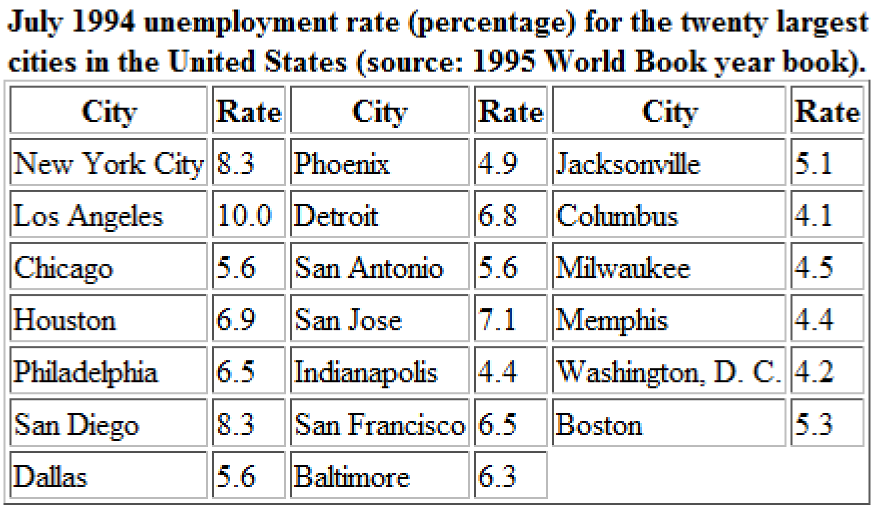
[✓ T] F The times women could name the colors is less variable than the times for men

[✓ T] F About 8 of the men needed 18-25 seconds to name the colors.

T [✓ F] About half of the women needed 14-17 seconds to name the colors.

[✓ T] F Since the distribution of the times for men is fairly symmetric, we can conclude that the mean time for men is about 23 seconds.

**17. (7 points) The unemployment rates for the twenty largest cities in the United States are given below.**



**July 1994 unemployment rate (percentage) for the twenty largest cities in the United States (source: 1995 World Book year book).**

| City | Rate | City | Rate | City | Rate |
|---|---|---|---|---|---|
| New York City | 8.3 | Phoenix | 4.9 | Jacksonville | 5.1 |
| Los Angeles | 10.0 | Detroit | 6.8 | Columbus | 4.1 |
| Chicago | 5.6 | San Antonio | 5.6 | Milwaukee | 4.5 |
| Houston | 6.9 | San Jose | 7.1 | Memphis | 4.4 |
| Philadelphia | 6.5 | Indianapolis | 4.4 | Washington, D. C. | 4.2 |
| San Diego | 8.3 | San Francisco | 6.5 | Boston | 5.3 |
| Dallas | 5.6 | Baltimore | 6.3 | | |

A student created a stemplot for the data:

```
 4 | 1 2 4 5 9
 5 | 1 3 6
 6 | 3 5 8 9
 7 | 1
 8 | 3
10 | 0
```

a. The stemplot is not correct. Find the mistakes, and correct them.

```
 4 | 1 2 4 4 5 9
 5 | 1 3 6 6 6
 6 | 3 5 5 8 9
 7 | 1
 8 | 3 3
 9 |
10 | 0
```

b. Describe the shape of the correct stemplot.

Skewed to the right

c. Find the range of the data.

Range = max. – min. = 10.0% – 4.1% = 5.9%

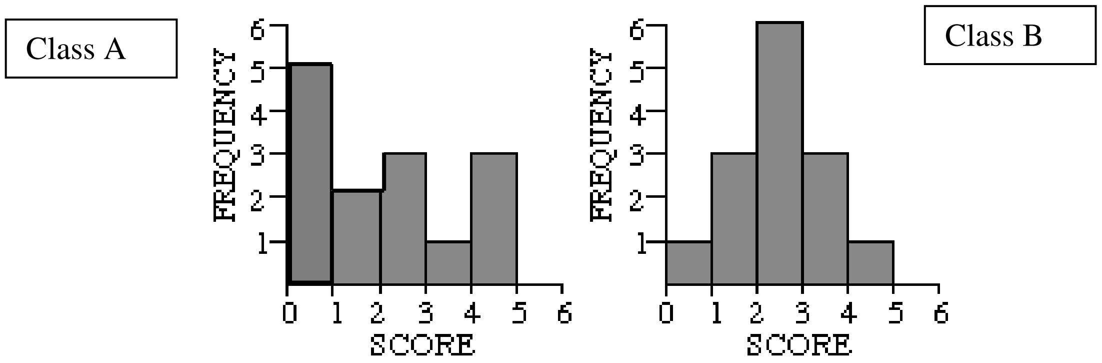
d. Find the interquartile range of the data, and interpret it in context.

Q1 = (4.5 + 4.9) / 2 = 4.7 Q3 = (6.8 + 6.9) /2 = 6.85

IQR = Q3 – Q1 = 6.85 – 4.7 = 2.15

The unemployment rate in the middle 10 of the 20 cities ranges between 4.7% and 6.85%. So the range of the unemployment rate for those 10 middle cities is 2.15%.

**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 lower standard deviation and **why**?

Class B has the lower standard deviation because on average the data points are closer from the mean.

**19. (4  points) There is a high correlation between ice cream sales and the number of shark attacks. Does this mean that we can lower the number of shark attacks by eating less ice cream? Explain using appropriate statistical terms.**

No, we can't lower the number of shark attacks by eating less ice cream. Even high correlation doesn't mean that one of the variables causes the changes in the other one. Correlation does not imply causation. The lurking variable connecting these two variables is temperature, or season. In summer, when the temperature is high many people eat ice cream, and many people swim in the ocean.