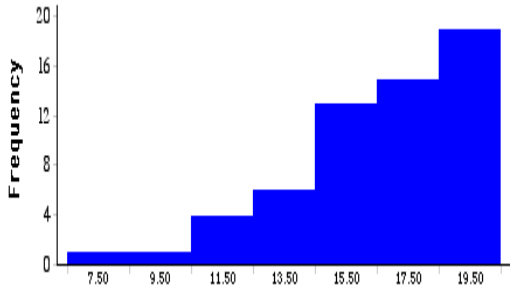
M 225 **Test 1** **B** Name_____
(1 point)

SHOW YOUR WORK FOR FULL CREDIT!

Problem	Max. Points	Your Points
1-14	14	
15	3	
16	5	
17	4	
18	4	
19	11	
20	9	
21	8	
22	16	
Total	75	

Multiple choice questions (1 point each)

1. Look at the following histogram.
   What shape would you say the data take?
   a) bimodal
   b) left-skewed
   c) right-skewed
   d) symmetric
   e) uniform



2. Which measures of center and spread are more appropriate for the distribution in the previous question?
   a) Median and interquartile range
   b) Mean and standard deviation
   c) Median and standard deviation
   d) Mean and interquartile range

3. If a distribution is skewed to the left,
   a) the median is less than the mean
   b) the mean is less than the median
   c) the mean and the median are equal

4. What percent of the observations in a distribution lie between the median and the third quartile Q_3?
   a) About 25%
   b) About 50%
   c) About 75%
   d) 100%

5. Which of the following is LEAST affected if an extreme outlier is added to your data?
   a) the mean
   b) the median
   c) the standard deviation
   d) range

6. Which of the following variables is categorical?
   a) Airfare from New York to Los Angeles
   b) Height of kindergarten kids
   c) Color of 2008 Volkswagen Beetle cars
   d) Time spent studying for this exam by the students in the class

7. Which of these statements are FALSE?
   a) There is a strong linear relationship between gender and height because we found a correlation of 0.54.
   b) Plant height and leaf height were found to be negatively correlated because the correlation coefficient is -1.45.
   c) Since the correlation between X and Y is 0, this means there is no relationship at all between these two variables.
   d) All of the above.
   e) None of the above.

8. What is the median of the following data: 1, 8, 4, 6, 9, 12, 16, 3, 7

a) 7
b) 12
c) 9
d) 22

9. What are all the values that a correlation r can possibly take?

a) $r \leq 0$
b) $0 \leq r \leq 1$
c) $-1 \leq r \leq 1$
d) $r \geq 0$

10. Several pieces of fruit from each tree in an orchard are selected. Identify the sampling technique.

a) Cluster sample
b) SRS
c) Stratified sample
d) Multistage sample

11. Using a local telephone book to select a simple random sample could introduce what type of bias?

a) Under-coverage bias
b) Non-response bias
c) Response bias
d) Question wording bias

12. Mr. Marino has compiled a list of 1,348 students in his high school. He has selected a sample of students by choosing every 14^{th} student on this list starting with a randomly selected student. Which type of sampling is he using?

a) random
b) stratified
c) cluster
d) systematic

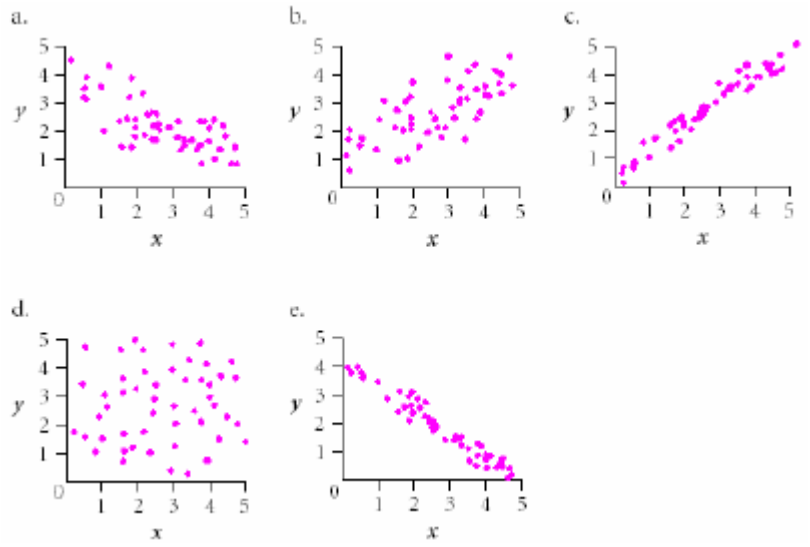
13. *Does drinking coffee tend to increase a students performance in school?* To answer this, I walked down to the Cafeteria and found 10 coffee-drinking undergraduates and 10 non-coffee drinking undergraduates. I asked them for their cumulative GPA and compared the GPAs for the two groups. Is this an experiment or an observational study?

a) experiment
b) observational study

14. Which one of the following statements is NOT true?

a) The only way the standard deviation can be 0 is when all the observations have the same value.
b) The correlation coefficient has the same units as the data.
c) The standard deviation has the same units as the data.
d) If the z-score for a value x is less than -2, the value is called an unusual

15. (3 points) Match each of the five scatterplots with its correlation.



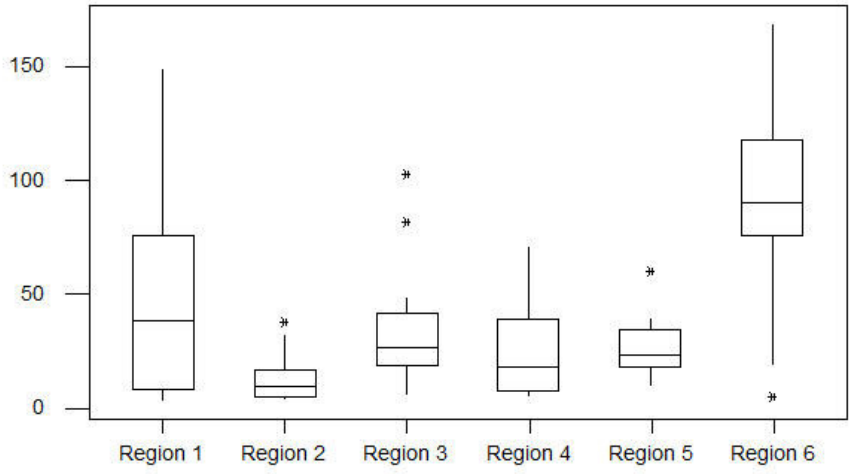
-0.5 **0.5** **0.95** **0** **-0.95**

a b c d e

16. TRUE or FALSE (1 point each)

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.

(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.

17. (4 points) Explain briefly the difference between observational study and experiment.

In an observational study we just observe the subject. We don't pose any treatments.
In an experiment we randomly assign treatments to the subjects. We can't establish cause and effect relationship using observational studies, but well designed experiments can establish cause and effect relationship. That's why we prefer to use them, if it's possible.

18. (4 points) What do we mean by double-blind experiment? Why do we use double-blind experiments at all?

Double-blind experiment means that the neither the subjects nor the experimenter know which group of subjects gets the treatment and which group gets the placebo. We like to use double-blind experiments because this is the best way to avoid bias.

19. The following data represent the price (in cents per pound) paid to 15 farmers for oranges.

17.2	19.6	16.4	19.1	18.0	17.4	17.3	20.1	19.0	17.5
18.6	17.6	18.4	17.7	19.8					

a. (1 point) Is the variable quantitative or categorical?

Quantitative

b. (1 point) Which of the following graphical displays is appropriate to for these data--stemplot or bar graph?

Stemplot

c. (4 points) Create the graph you picked in the previous part. Describe the shape of the distribution.

16.	4
17.	2 3 4 5 6 7
18.	0 4 6
19.	0 1 6 8
20.	1

The shape of the distribution is fairly symmetric and bimodal.

d. (5 points) Find the five-number summary, and check the data set for outliers using the 1.5(IQR) rule.

Five-number summary:
Min.: 16.4
Q1: 17.4
Med: 18
Q3: 19.1
Max: 20.1

IQR = Q3 – Q1 = 19.1 – 17.4 = 1.7

Q1 – 1.5(IQR) = 17.4 – 1.5(1.7) = 14.85 No data below 14.85, so no low outliers.
Q3 + 1.5(IQR) = 19.1 +1.5(1.7) = 21.65 No data above 21.65, so no high outliers.

No outliers.

20. The heights of men aged 20 to 29 is approximately Normal with mean 72 inches and standard deviation of 2.7 inches.

a. (3 points) How tall are those men who are in the middle 95%?

The middle 95% is two standard deviations below and above the mean:
72 – 2(2.7) = 66.6
72 + 2(2.7) = 77.4

Thus, the height of men in the middle 95% is between 66.6 and 77.4 inches.

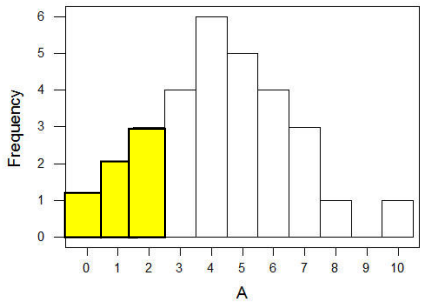
b. (3 points) What percent of men in this age group are taller than 74.7 inches?

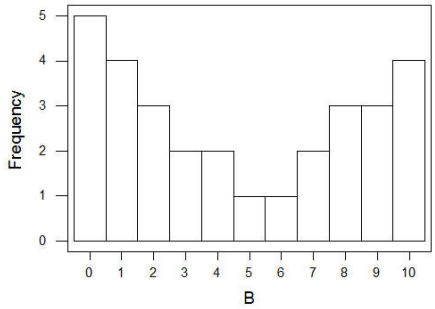
Since 74.7 is one standard deviation above the mean, the upper tail is 16%.
Thus, 16% of the men are taller than 74.7 inches.

c. (3 points) How tall are those men who are in the shortest 2.5%?

The shortest 2.5% is the lower tail below 2 standard deviations of the mean. That is 66.6 inches.
Thus, the shortest 2.5% of the men are 66.6 inches or shorter.

21. 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. (3 points) Which distribution has the higher standard deviation and why?

Distribution B has the higher standard deviation because most of the values are far from the mean. Only a few are around the mean.

b. (2 points) What percent of households have two houseplants or less?

3 households have two plants, 2 households have one plant, and 1 household has no plants. That is 6 out of 30, 6/30 = 0.2 = 20%
(See yellow bars on the graph)

c. (2 points) Select the statement below that gives the most complete and correct statistical description of the graph A.

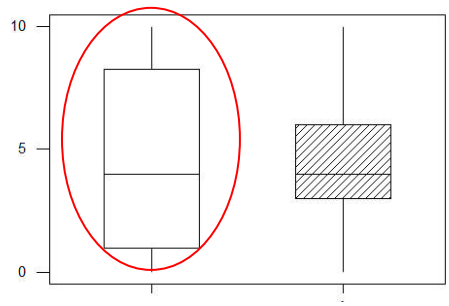
A. The bars go from 0 to 10, increasing in height to 4, then decreasing to 10. The tallest bar is at 4. There is a gap between 8 and 10.

B. The distribution is 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 outlier at 10. The typical number of houseplants owned is about 4, and the overall range is 10 plants.

d. (1 point) Which of these graphs is the boxplot for distribution B?



22. Data on Casino Employees (in thousands) and Crime Rate (number of crimes per 1,000 population) in Mississippi for the years 1998-2005 is given below.

Casino Employees	15	18	24	22	26	30	31	32
Crime Rate	1.35	1.63	2.33	2.38	2.63	2.24	3.41	3.26

$\bar{x} = 24.75 \quad s_x = 6.20$

$\bar{y} = 2.40 \quad s_y = 0.711$

$r = 0.892$

a. (2 points) Identify the explanatory and response variables.

Explanatory variable: Casino employees
Response variable: Crime rate

b. (4 points) Display the data in a scatter plot, and describe the plot

Form: linear
Direction: positive
Strength: moderately strong
Outliers: maybe one

c. (3 points) Find the equation of the least squares line, and sketch the line on the plot. Use three decimal digits in your answers.

Y = - 0.125 + 0.102X

d. (2 points) Predict the crime rate for a casino having 20,000 employees.

Y = - 0.125 + 0.102(20) = 1.92
Using the regression line, we can predict that the crime rate for a casino having 20,000 employees will be about 1.92 per 1,000 population.

e. (2 points) One observation greatly affects the apparent relationship. Circle it, and indicate which of the following is the most likely value of *r* if this point is removed:

–.94 –.65 .82 .98

g. (2 points) Would it be OK to use the regression line to predict the crime rate for a casino with 40,000 employees? Explain.

No, it wouldn't be OK. 40 is out of the range of the explanatory variable (which is 15 to 32 from the table), so it's not reliable to use the regression line for this prediction. That would be extrapolation.