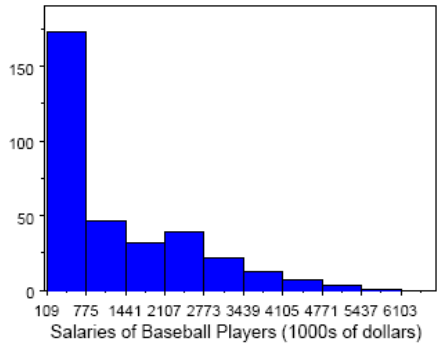
M 225 **Test 1** **A** Name_____
(1 point)

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 for salaries of baseball players. What shape would you say the data take?
   a) bimodal
   b) left-skewed
   c) right-skewed
   d) symmetric
   e) uniform

2. For the distribution of major league baseball players' salaries in the previous question, which measures of center and spread are more appropriate?
   a) Mean and standard deviation
   b) Median and interquartile range
   c) Mean and interquartile range
   d) Median and standard deviation

3. Which one of the following variables is NOT categorical?
   a) whether or not an individual has a cell phone
   b) the color of Reese's Pieces candy
   c) the airfare to a selected city from LAX
   d) the occupational background of a Civil War general

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

5. What percent of the observations in a distribution lie between the first quartile Q_1 and the third quartile Q_3?
   a) About 25%
   b) About 50%
   c) About 75%
   d) 100%

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

7. Which of these statements are FALSE?
   a) There is a strong linear relationship between gender and height because we found a correlation of 0.55.
   b) Plant height and leaf height were found to be negatively correlated because the correlation coefficient is -1.41.
   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, 5, 4, 8, 11, 3, 9, 15, 2
a) 4
b) 5
c) 4
d) 29

9. What are all the values that a correlation r can possibly take?
a) $r \geq 0$
b) $0 \leq r \leq 1$
c) $-1 \leq r \leq 1$
d) $r \leq 0$

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

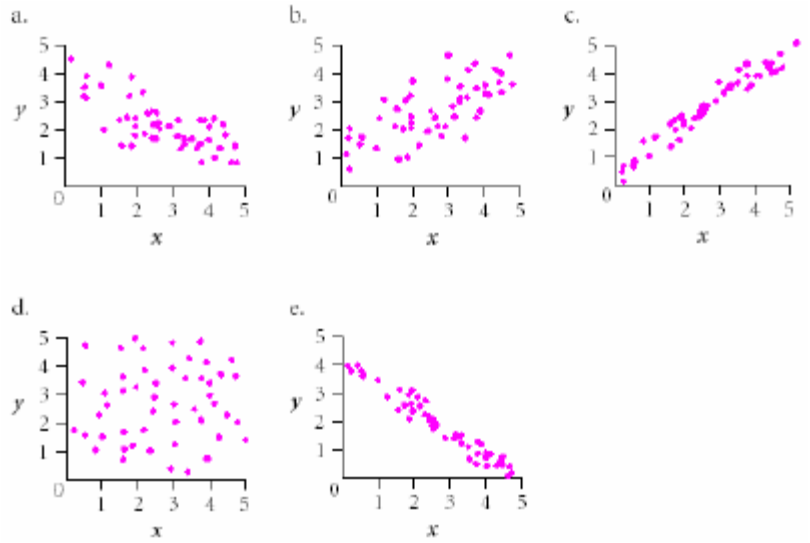
11. A sample of households in a community is selected at random from the telephone directory. In this community, 4% of households have no telephone and another 35% have unlisted telephone numbers. The sample will certainly suffer from
a) Nonresponse
b) Undercoverage
c) False response

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 value.

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

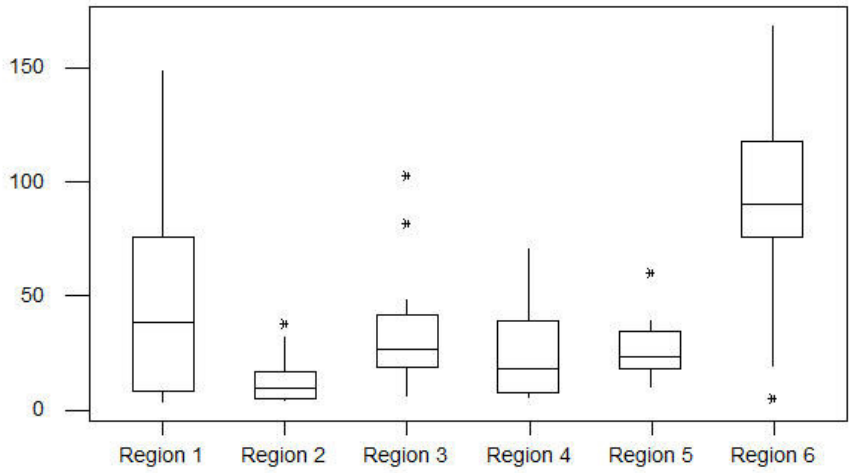


0.5	**-0.5**	**-0.95**	**0**	**0.95**
b	a	e	d	c

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 apples.

19.2 19.6 16.4 17.1 19.0 17.4 17.3 20.1 19.0 17.5
17.6 18.6 18.4 17.7 19.5

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. | 1 3 4 5 6 7
18. | 4 6
19. | 0 0 2 5 6
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: 18.4
Q3: 19.2
Max: 20.1

IQR = Q3 – Q1 = 19.2 – 17.4 = 1.8

Q1 – 1.5(IQR) = 17.4 – 1.5(1.8) = 14.7 — No data below 14.7, so no low outliers.
Q3 + 1.5(IQR) = 19.2 +1.5(1.8) = 21.9 — No data above 21.9, so no high outliers.

No outliers.

20. The heights of women aged 20 to 29 is approximately Normal with mean 64 inches and standard deviation of 2.7 inches.

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

The middle 95% is two standard deviations below and above the mean:
64 – 2(2.7) = 58.6
64 + 2(2.7) = 69.4

Thus, the height of women in the middle 95% is between 58.6 and 69.4 inches.

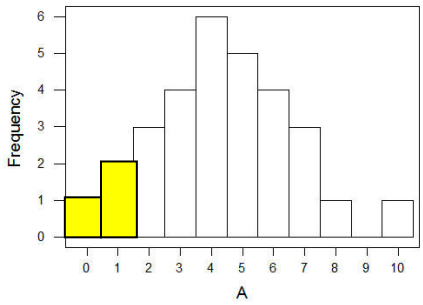
b. (3 points) What percent of women in this age group are taller than 66.7 inches?

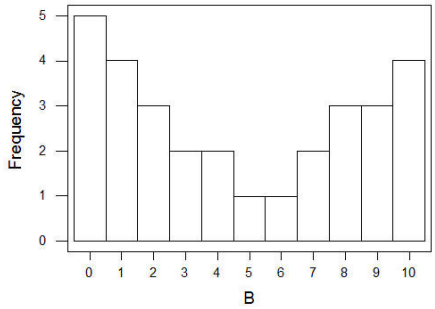
Since 66.7 is one standard deviation above the mean, the upper tail is 16%.
Thus, 16% of the women are taller than 66.7 inches.

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

The shortest 2.5% is the lower tail below 2 standard deviations of the mean. That is 58.6 inches.
Thus, the shortest 2.5% of the women are 58.6 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 lower standard deviation and why?

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

b. (2 points) What percent of households have one houseplant or none?

2 households have one plant, and 1 household has no plants. That is 3 out of 30, 3/30 = 0.1 = 10%
(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.

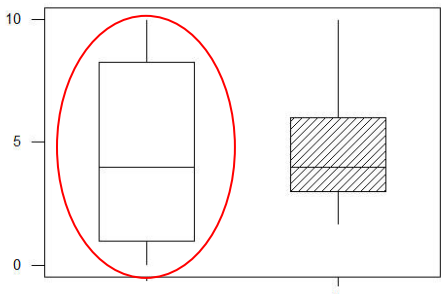
A. The bars go from 0 to 10, increasing in height to 4, then decreasing to 10. The tallest bar is at 4. There is a gap between 8 and 10.

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 a possible 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. The ages (in weeks) and the numbers of hours slept in a day by eight infants are given below.

Age	8	10	22	31	36	42	39	45
Hours slept	14.9	14.5	13.1	13.7	14.0	13.4	13.7	13.0

$\bar{x} = 29.13 \quad s_x = 14.29$

$\bar{y} = 13.79 \quad s_y = 0.66$

$r = -0.747$

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

Explanatory variable: Age
Response variable: Hours slept

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

Form: linear
Direction: negative
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 = 14.792 – 0.034X

d. (3 points) Predict the number of hours slept for a 15 weeks old infant.

Y = 14.792 – 0.034(15) = 14.275

Using the regression line, we can predict that a 15-week old infant sleeps about 14.275 hours a day.

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 .42 .78

g. (2 points) Would it be OK to use the regression line to predict the number of hours slept for a 60-week old infant? Explain.

No, it wouldn't be OK. 60 is out of the range of the explanatory variable (which is 8 to 45 from the table), so it's not reliable to use the regression line for this prediction. That would be extrapolation.