SHOW YOUR WORK FOR FULL CREDIT!

<table>
<thead>
<tr>
<th>Problem</th>
<th>Max. Points</th>
<th>Your Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-13</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>75</strong></td>
<td></td>
</tr>
</tbody>
</table>
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. For the distribution 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 quantitative?
   a) running time of a Hitchcock movie
   b) the weight of elephants in a herd
   c) the airfare to a selected city from LAX
   d) the color of M&Ms

4. If a distribution is skewed to the left,
   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 median and the third quartile Q₃?
   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. What is the median of the following data: 1, 5, 4, 8, 3
   a) 4
   b) 5
   c) 3
   d) 8
8. 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) \( r \leq 0 \)

9. Several pieces of fruit from each tree in an orchard are selected. Identify the sampling technique.
   a) Multistage sample
   b) SRS
   c) Stratified sample
   d) Cluster sample

10. 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

11. 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

12. A research study has reported that there is a correlation of \( r = -0.59 \) between the eye color (brown, green, blue) of an experimental animal and the amount of nicotine that is fatal to the animal when consumed. This indicates:
   a) nicotine is less harmful to one eye color than the others.
   b) the lethal dose of nicotine goes down as the eye color of the animal changes.
   c) one must always consider the eye color of animals in making statements about the effect of nicotine consumption.
   d) the researchers need to do further study to explain the causes of this negative correlation.
   e) the researchers need to take a course in statistics because correlation is not an appropriate measure of association in this situation.

13. 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.
14. (3 points) Match each of the five scatterplots with its correlation.

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

15. (8 points) Let’s suppose you are majoring in Child Development. As your project, you need to estimate the proportion of elementary school students in a small district who believe in Santa Claus.

If you could ask every elementary student in the district, that would be called a(n) ____census____.

But you don’t have the time to ask each and every elementary kid about Santa Claus, so you take a random sample of 500 of them.

a. Clearly identify in words the population of interest, the parameter, the sample, and the statistic:

Population: ALL the elementary kids in that district

Parameter: the proportion of ALL elementary kids in that district who believe in Santa Claus.

Sample: the 500 elementary students you randomly selected

Statistic: the proportion of the 500 elementary kids in the sample who believe in Santa Claus.

b. There are five elementary schools in the district, and each school has grades from Kindergarten to 5th grade. There are many ways you can pick those 500 students. Identify the sampling method described in each case below:

From each elementary school you randomly select 100 students.

Sampling method: ______stratified sample_______________________________
You randomly pick two elementary schools, and then in these two schools you randomly select three grade levels. Then from these grades you randomly select 500 students.

Sampling method: __multistage sampling_________________

From a list of all of the elementary school children in the district, you randomly select 500 students.

Sampling method: ________SRS__________________________

16. (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.

17. 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 for these data--stemplot or bar graph?  stemplot

c. (3 points) Create the graph you picked in the previous part. Describe the shape of the distribution.

\[
\begin{array}{cccccccc}
16. & 4 & 17. & 2 & 3 & 4 & 5 & 6 & 7 \\
18. & 0 & 4 & 6 & 19. & 0 & 1 & 6 & 8 \\
20. & 1 & & & & & & & \\
\end{array}
\]

Shape: roughly symmetric, bimodal

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

Min = 16.4  Q1 = 17.4  M= 18.0  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  
Q3 + 1.5(IQR) = 19.1 + 1.5(1.7) = 21.65  
No data below 14.85 \(\Rightarrow\) no low outliers  
No data above 21.65 \(\Rightarrow\) no high outliers

18. The heights of men aged 20 to 29 is approximately Normal with mean 72 inches and standard deviation of 2.7 inches. Use the Empirical rule to answer the following questions:

a. (3 points) What proportion of men are between 66.6 in and 77.4 in?

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.

19. 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?

c. 
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. (1 point) Which of these graphs is the boxplot for distribution A?

![Boxplot Image]

e. (1 point) What can you say about the median number of houseplants and the median number of girlfriends/boyfriends?

Looking at the graph, it seems that they are equal (the line in the boxes (the median) are at the same height).

20. (3 points) Human body temperatures have a mean of 98.20 °F and a standard deviation of 0.62°F. Convert the following temperature to a z-score and determine whether it is usual or unusual.

\[
z = \frac{x - \text{mean}}{s.d.} = \frac{96.80 - 98.20}{0.62} = -2.258
\]

Circle the correct answer: usual unusual

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

<table>
<thead>
<tr>
<th>Age</th>
<th>Hours slept</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>14.9</td>
</tr>
<tr>
<td>10</td>
<td>14.5</td>
</tr>
<tr>
<td>22</td>
<td>13.1</td>
</tr>
<tr>
<td>31</td>
<td>13.7</td>
</tr>
<tr>
<td>36</td>
<td>14.0</td>
</tr>
<tr>
<td>42</td>
<td>13.4</td>
</tr>
<tr>
<td>39</td>
<td>13.7</td>
</tr>
<tr>
<td>45</td>
<td>13.0</td>
</tr>
</tbody>
</table>

\[
\bar{x} = 29.13 \quad s_x = 14.29
\]
\[
\bar{y} = 13.79 \quad s_y = 0.66
\]
\[
r = -0.747
\]
a. (2 points) Identify the explanatory and response variables.

Explanatory: Age  
Response: Hours slept

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

![Scatter plot]

Y = 14.792 – 0.034X

d. (3 points) Predict the number of hours slept for a 20 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

The correlation coefficient is -0.747, but if we remove that point, the correlation will be stronger, so closer to -1. Thus it must be -0.94.
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.

22. (2 points) What are the two main differences between a bargraph and a histogram?

Bargraphs are used to represent one categorical variable, and the bars are not touching each other.

Histograms are used to represent one quantitative variable, and the bars are touching each other.