**Free Response**

**Use the following to answer problems 1-4:**
The following is a list of the top 20 in total rebounds in all games for men's basketball in the Mountain West conference in the 2011-2012 season.

| 91 | 115 | 127 | 131 | 134 | 138 | 144 | 144 | 146 | 151 |
|----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 154 | 156 | 166 | 172 | 227 | 228 | 231 | 252 | 369 | 388 |

1. Calculate and label the 5-number summary for the data.

2. Are there any outliers in the data? **Show all work.**
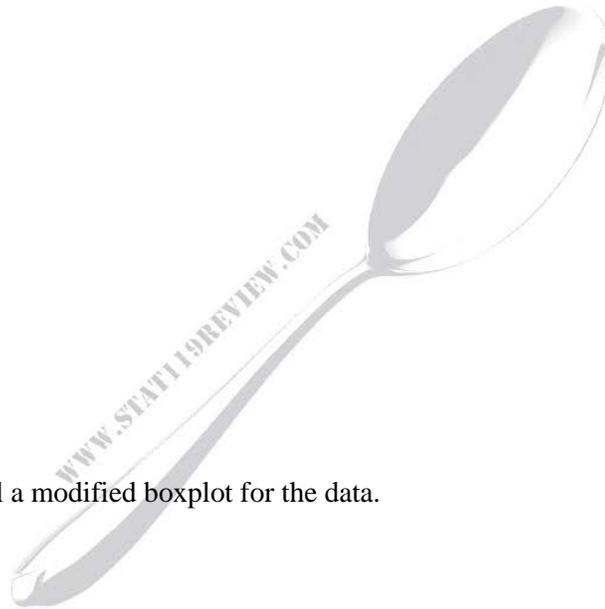
3. Construct and label a modified boxplot for the data.

```
_____
80  100  120  140  160  180  200  220  240  260  280  300  320  340  360  380
```

4. Identify the best measure of center and spread, based on the distribution (do **NOT** calculate). Then justify your choice of measures in one sentence:
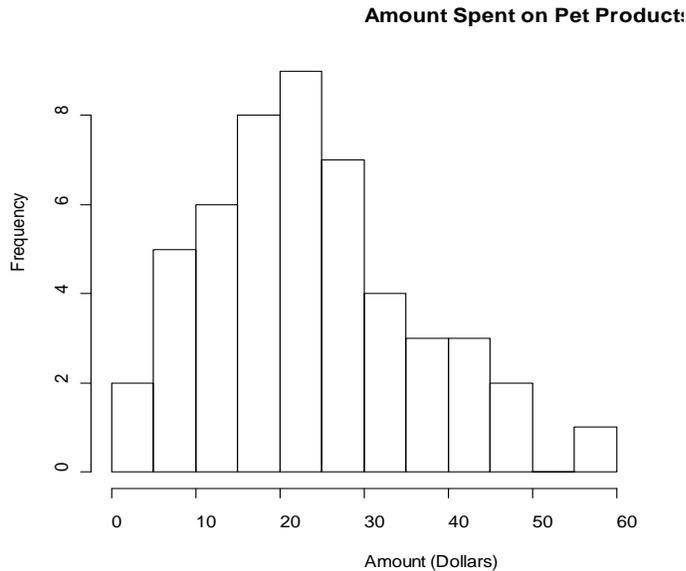
   Center:_____        Spread:_____

   Justification:

1

**Use the following information to answer questions 5-6**

A marketing consultant randomly surveyed 50 shoppers at a particular pet store to see how much they spent each month on pet products. A histogram of the survey data was constructed.

**Amount Spent on Pet Products**



5.  What percent of individuals spend more than $40?

6.  Which interval is Q3 contained in?

7.  Corepower Yoga offers a 200-hour yoga teaching certification. They offer teacher training at 4 of their locations: Point Loma, Pacific Beach, La Jolla and North Park. Suppose they want to sample 20 of the recent teacher training graduates to see what factors influenced them to take the training. Describe how they could collect this sample using the three types of sampling listed below, in one or two sentences. Assume that 20 people graduate from each location.

    Systematic:



    Stratified:



    Cluster:

**Use the following information to answer questions 8-13**

A sample of small cars was selected to attempt to use the horsepower (hp) of the car to predict the fuel efficiency (mpg). A researcher fitted a linear regression model.

$$\hat{y} = 44.0 - 0.150x \qquad\qquad R^2 = 68\%$$

8. Interpret the slope of the regression line.

9. If a particular car has a horsepower of 150hp and a fuel efficiency of 25mpg, what is the residual for this car?

10. Calculate the correlation coefficient.

11. How would you interpret the correlation coefficient found in (3)?

12. If a car with 120 horsepower has residual of 3.27, what was the actual fuel efficiency of the car?

13. The study also investigated the relationship between car weight and fuel efficiency and obtained the following information:

$$Fuel\ Efficiency = 48.71 - 8.365(Weight) \qquad\qquad R^2 = 78.2\%$$

Which is the better predictor of fuel efficiency, horsepower or weight? Explain your choice using STATISTICS from the two regression lines for credit!

14. Real estate ads suggest that 56% of homes for sale have garages, 50% have swimming pools, and 11% have both features.

a.) What is the probability that a house selected at random will not have a garage but will have a swimming pool?

b.) What is the probability, given the house does not have a garage, that it has a swimming pool?

c.) Are having a garage and swimming pool independent? Explain why or why not using probabilities.
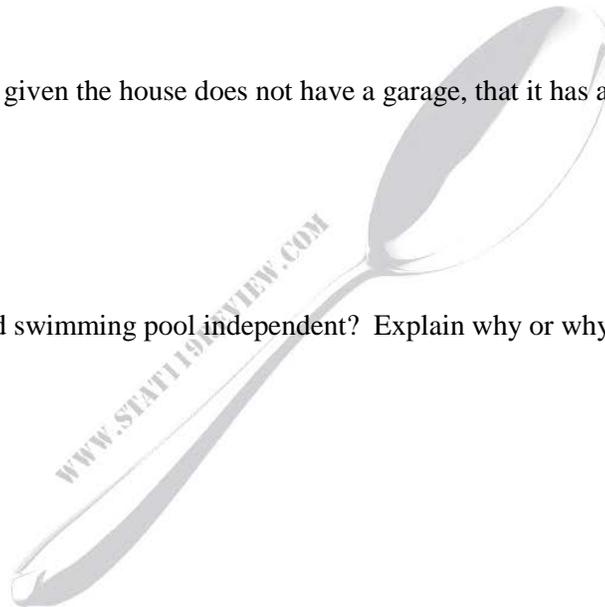
15. Suppose a cat has a litter of 8 kittens. Of those 8 kittens, 3 are black. Three kittens are selected at random without replacement.
   a.) What is the probability that all three kittens will be black?

b.) What is the probability that at exactly one kitten will be black?

16. Ten juniors and four seniors have applied for two open student council positions. School administrators have decided to randomly select the two new members. What is the probability that one junior and one senior are chosen for the new positions?

17. Assume we know that the probability that a college students owns an iPod is 0.75. Four students are selected at random (therefore we can assume independence). What is the probability that none of them owns an ipod?

18. In an intro stats class, 57% of students eat breakfast in the morning and 80% of students floss their teeth daily. If 91% of students eat breakfast in the morning or floss daily, what percent of students eat breakfast in the morning and floss their teeth daily?

19. Assume the probability that a freshman SDSU student is from San Diego is 0.5 and the probability that a freshman SDSU student is taking introductory statistics is 0.1. If these two events are independent, what is the probability that a student selected at random is taking introductory statistics or is from San Diego?

20. A survey of SDSU students is taken. It is found that the probability of an SDSU student being an undergraduate is 75%. Given a student is an undergraduate, the probability of that student being employed is 30%. Given a student is a graduate student, the probability of that student being employed is 60%.

   a.) Draw a tree diagram below: use correct notation to label events and fill in all appropriate probabilities (including intersections).
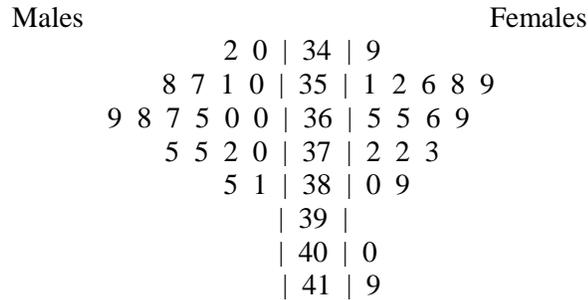
   b.) What is the probability of a student being unemployed?

   c.) If a student is unemployed, what is the probability that the student is an undergraduate?
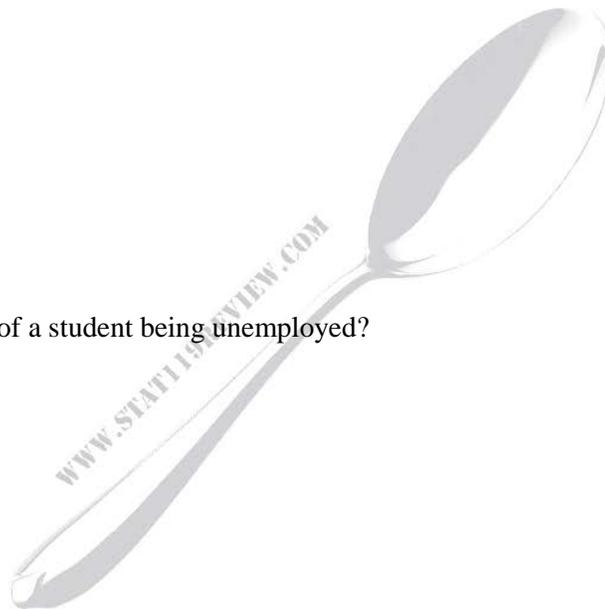
**Use the following information to answer questions 21-24**

The following back-to-back stemplot represents the amount spent by a sample of males and females on clothing in a four month period.

```
           Males                          Females
              2 0 | 34 | 9
          8 7 1 0 | 35 | 1 2 6 8 9
      9 8 7 5 0 0 | 36 | 5 5 6 9
          5 5 2 0 | 37 | 2 2 3
            5 1 | 38 | 0 9
                 | 39 |
                 | 40 | 0
                 | 41 | 9
```

21. What are the appropriate measures of center and spread for the **male** data? (1 point extra credit for each of the calculated values)

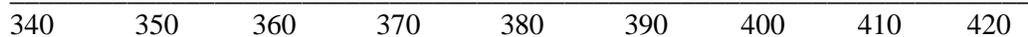    **Choice of Center:** _____        **Calculated value:** _____

    **Choice of Spread:** _____        **Calculated value:** _____

    **Justification of choice:** _____

22. What is the 5-number summary for the **female** data?

23. Check the **female** data for outliers, using the criterion discussed in class.

24. (4 points) Construct a modified boxplot for the **female** data.

```
    |_____|
  340     350     360     370     380     390     400     410     420
```

### Multiple Choice

1. A survey of students at SDSU is conducted. Which of the following variables is not categorical?
   A. Gender
   B. Telephone number
   C. Age
   D. Relationship status

### Use the following information to answer questions 3-4

A marketing consultant randomly surveyed 50 shoppers at a particular pet store to see how much they spent each month on pet products. A histogram of the survey data was constructed.

**Amount Spent on Pet Products**
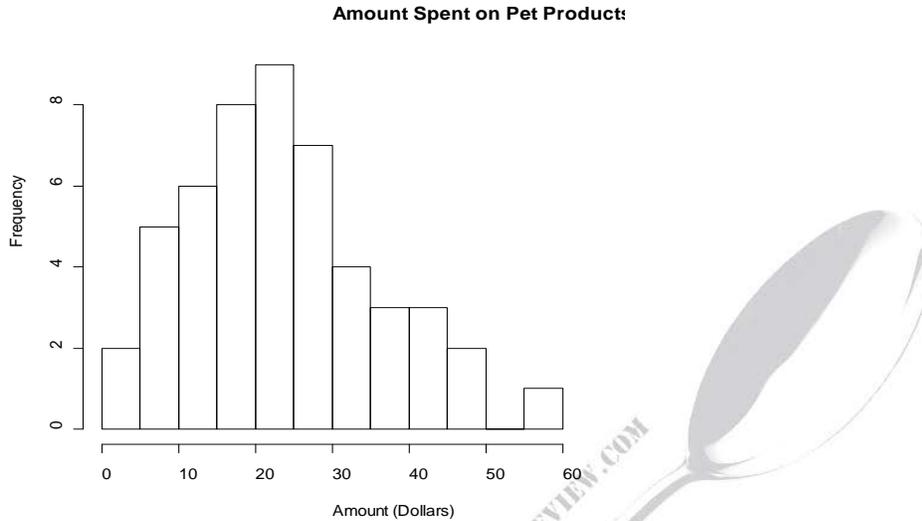


2. What percent of individuals spend between $30 and $50 per month?
   A. 12%          B. 19%          C. 24%          D. 38%

3. What interval is Q3 located in?
   A. [10,15)      B. [15,20)      C. [20,25)      D. [25,30)      E. [30,35)

4. SDSU's Dining Services conducted a survey to find out how on-campus first-year students felt about the current required meal plans. They randomly sampled 50 first-years from each of the 6 residence halls and them if they were happy with their meal plan. What type of sampling did they use?
   A. Systematic Sample
   B. Simple Random Sample
   C. Stratified Sample
   D. Cluster Sample

### Use the following to answer questions 5-6:

An exercise physiologist is doing a research study on post-menopausal women and bone density. The researcher considers a variety of variables that could have a role in the bone density of this group of women. He looks at a group of 1529 post-menopausal women and asks them about whether or not they had taken oral contraceptives prior to menopause and how many minutes they exercise.

5. What would be the appropriate graphical display for these two variables, respectively?

   A. bar chart and histogram            B. pie chart and bar chart
   C. histogram and boxplot              D. boxplot and pie chart

6. What is the population in this study?
    A. Women
    B. Post-menopausal women
    C. 1539 post-menopausal women
    D. It can't be determined from the information given

7. A high school track coach wants to determine if strength training will improve the times of the runners on the team. The team consists of 20 long distance runners and 20 sprinters. If the coach randomly assigns half of the long-distance runners and half of the sprinters to a strength training regime in addition to track practice and the other half continue with only track practices. What type of experimental design has the track coach used?
    A. Stratified
    B. Completely Randomized
    C. Block
    D. Matched Pairs

8. In a random sample of 50 college students it was determined that the average number of times they checked social media sites per day was 4.65. The value 4.65 is a:
    A. Population       B. Parameter       C. Sample       D. Statistic

## Use the following table to answer questions 9-12

Suppose the following table displays information from 200 new car purchasers showing which additional options they purchased along with whether they were a first time new car buyer or not:

|  | Extended Warranty Only | Maintenance Plan Only | Both | Neither |
|---|---|---|---|---|
| First Time Buyer | 65 | 0 | 32 | 23 |
| Not First Time Buyer | 24 | 1 | 19 | 36 |

9. What percentage of subjects purchased both the extended warranty and the maintenance plan?
    A. 16%              B. 25.5%           C. 32%              D. 51%

10. What percentage of subjects were not first time buyers and did not purchase any additional options?
    A. 18%              B. 19.2%           C. 45%              D. 61%

11. What percentage of subjects purchased the maintenance plan?
    A. 0.5%             B. 16%             C. 26%              D. 52%

12. Of those subjects who purchased an extended warranty only, what percent were first time buyers?
    A. 32.5%            B. 54.2%           C. 65%              D. 73%

13. In a study of the weights of 3rd graders, the data was found to follow a symmetric bell curve. Which of the following is the best measure of spread for this data?
    A. Standard deviation
    B. IQR
    C. Mean
    D. Median
    E. Cannot be determined.

14. A weather station recorded wind speeds every day for 50 days. When the researcher looked at the data using a histogram, there was one low outlier - a day where the wind speed was 0 mph. It was later determined that the instrument that measured wind speed malfunctioned on that day.

    This incorrect data point was removed from the data and a new histogram was made and all new statistics were calculated. Which of the following statements is true?

    When the incorrect data point is removed:
    A. The mean and standard deviation will both be smaller
    B. The mean and standard deviation will both be larger
    C. The mean will be smaller, but the standard deviation will be larger
    D. The mean will be larger, but the standard deviation will be smaller
    E. None of the above

15. The IFOCE, the governing body for competitive eating, wants to know what competitive eaters think about Kobayashi's contract dispute. Kobayashi is currently unable to compete in the Nathan's hot dog eating competitions because he wouldn't sign a contract limiting him to compete exclusively in IFOCE events. To see what other competitive eaters thought, the IFOCE asked each of the eaters competing at the Nathan's hot dog competition to fill out an anonymous survey. What is the most prominent source of bias in this survey?
    A. Voluntary Response Bias
    B. Nonresponse Bias
    C. Response Bias
    D. Undercoverage

16. The census found that in 2003, households headed by persons between the ages of 45 and 54 had a median household income of $61,111 and a mean household income of $77,634. What does this tell us about household incomes for households headed by persons between the ages of 45 and 54?
    A. Over half of the households earn less than $61,111
    B. Over half of the households earn more than $61,111
    C. Over half of the households earn less than $77,634
    D. Over half of the households earn more than $77,634

17. Below are the summary statistics for average daily wind speed for a region of Massachusetts:

| Min | Q1 | Median | Q3 | Max | Mean | Standard Deviation |
|-----|-----|--------|-----|------|------|--------------------|
| 0.20 | 1.15 | 1.90 | | 2.93 | 8.67 | 2.13 | 1.33 |

    Which of the following is a correct statement?
    A. The sample of average wind speeds is left skewed
    B. The sample of average wind speeds is right skewed
    C. The sample of average wind speeds is symmetric
    D. We cannot tell the shape of the distribution from summary statistics

18. According to a recent article, the mean hourly wage for general managers at a large firm was $92.33 and the median was $69. We can conclude
    A. More than half the managers earn less than $92.33/hour
    B. More than half the managers earn more than $69/hour
    C. The distribution of hourly wages is symmetric
    D. The distribution of hourly wages is left-skewed

19. In a study of unemployment rates for a sample of major cities, the distribution was found to be left skewed with two low outliers. What would be the most appropriate measure of center for this distribution?
    A. Mean
    B. Median
    C. Standard Deviation
    D. IQR
    E. Range

**Use the following boxplot to answer 20 & 21**

The following boxplots show risk scores for Alveolar Echinococcosis among case and control groups in a German study from 1999.
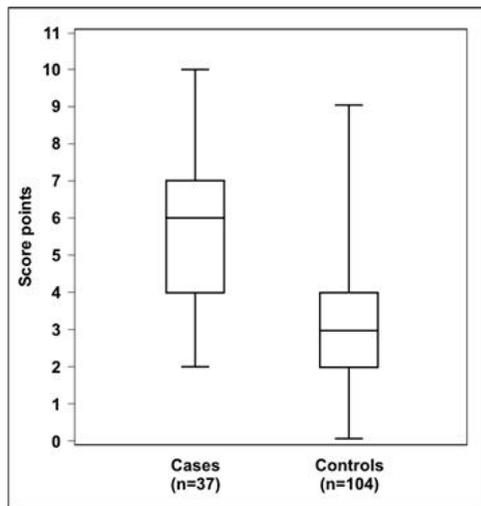


20. Which of the following statements is NOT true?
    A. 25% of the case group had scores of at least 7.
    B. 75% of the control group had scores under 4.
    C. The IQR for the case group is smaller than that for the control group.
    D. The median for the control group is smaller than that for the case group.

21. 75% of the control group had scores of at least:
    A. 2      B. 3      C. 4      D. 6      E. 7

22. Which of the following statements about the coefficient of determination is not true?
    A. It is also denoted as $R^2$.
    B. It is the measure of the percent of variability in the response variable that can be explained by linear regression of the response variable on the explanatory variable.
    C. It can be used to determine if the relationship between the explanatory and response variables is positive or negative.
    D. It can be used to determine the predictive power of a regression equation.

23. A dentist found that the coefficient of correlation between the time patients spend brushing their teeth and a numerical rating of whiteness of those teeth is 0.8. This suggests that
    A. There is no association between time spent brushing teeth and whiteness of teeth.
    B. 80% of people who brush their teeth have white teeth
    C. There is a strong, negative, linear relationship between time spent brushing teeth and the whiteness of teeth.
    D. The coefficient of determination for this relationship is 64%.

11

24. A researcher finds that the correlation coefficient between people's IQ and their favorite type of music is 0.67. What should he conclude?
   A. There is a strong non-linear relationship.
   B. There is a moderate positive linear relationship.
   C. Those who listen to classical music have a higher IQ.
   D. A correlation coefficient is not appropriate for this data.

25. All of the following are reasons for plotting your data on a scatterplot before doing regression except:
   A. Identifying non-linear trends in the data.
   B. Identifying possible influential points or outliers in the data
   C. Identifying possible grouping problems in the data.
   D. Identifying possible lurking variables.

26. In order to determine whether we can estimate the age of a tree simply from its diameter (inches), a forester measured 27 trees of the same species that had been cut down and counted the rings to determine the ages (years) of the trees. The data was analyzed to produce the following results:

$$Age = -6.22 + 2.56(Diameter) \qquad Correlation\ coefficient = 0.908$$

What percent of variability in age can be explained by regression of age on diameter?
   A. 82.45%
   B. 95.29%
   C. -95.29%
   D. 90.8%
   E. It cannot be determined from the information given.

27. A positive residual for a particular subject means
   A. The regression line underfit the prediction for the subject.
   B. The subject is an outlier.
   C. The regression line overfit the prediction for the subject.
   D. The subject has high leverage.

**Use the following to answer questions 28**
The following is a series of boxplots comparing the weights of chickens based on the feed they were given.



Chicken Weights by Feed
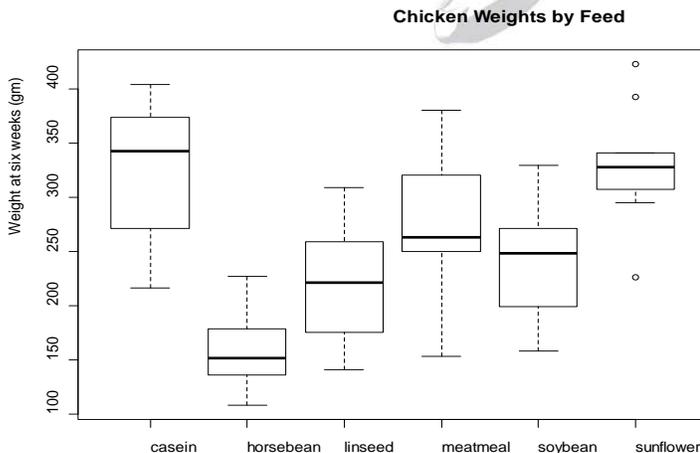
28. Which of the feeds listed below has the smallest median value?
   A. Casein          B. Meatmeal          C. Soybean          D. Sunflower

29. In an intro stats class, 57% of students eat breakfast in the morning and 80% of the students floss their teeth daily. 46% of student eat breakfast and floss their teeth. What is the probability that a student selected at random from this class eats breakfast or flosses their teeth?

   A.  0.34          B. 0.91          C. 1.37          D. 0.09          E. 0.11

30. A box contains 16 soccer jerseys: 4 medium, 7 large and 5 extra large. Two jerseys are drawn from the box, without replacement. What is the probability that exactly one medium jersey is selected?

   A.  0.375         B. 0.20          C. 0.05          D. 0.40

**Use the following information to answer questions 31-32**

A survey of real estate in New York classified homes into two price categories. Low: less than $250,000 and high: over $250,000. The survey found that 62% of houses were low priced (event A). 56% of the houses surveyed had at least 2 bathrooms (event B). 34% of the houses surveyed were high priced and had at least two bathrooms.

   31. (2 points) Fill in the four parts of the Venn diagram below with the appropriate probabilities.
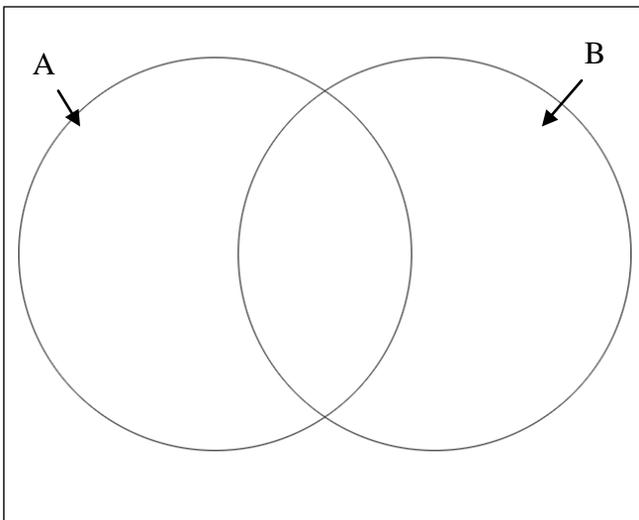
32. Given that a house is high priced, what is the probability it has at least two bathrooms?
A.  0.6071          B. 0.5484          C. 0.34          D. 0.8947

33. A dentist found that the coefficient of correlation between the time patients spend brushing their teeth and frequency of cavities is -0.8. This suggests that
   A.  80% of people who don't brush their teeth have cavities.
   B.  There is no association between the time spent brushing teeth and cavities.
   C.  There is strong negative association between the time spent brushing teeth and frequency of cavities.
   D.  The coefficient of determination for this relationship is 64%.

34. A Z-score of -3 means
   A.  The observation is 3 standard deviations smaller than the mean
   B.   The observation is 3 standard deviations larger than the mean
   C.  The observation is three times smaller than the mean
   D.  The difference between the observation value and the mean is 2

35. Suppose a student takes 2 exams, getting 55 in a verbal test and 60 in a numerical reasoning test. The class scores for each exam are normally distributed. For the verbal test, the mean is 50 and standard deviation 5 and for the numerical test, the mean is 50 and standard deviation is 12. On which test did the student score better on, relative to his classmates?
    A. The verbal test because the Z-score is higher.
    B. The numerical reasoning test because the Z-score is higher
    C. The verbal test because the Z-score is lower.
    D. The numerical reasoning test because the Z-score is lower.

**For problems 36-55: determine whether each statement is true or false.**

36. If the IQR is 0, all the data points must be the same.

37. If you add 5 to the values in a data set, the mean and median will also increase by 5.

38. When conducting an experiment, the control group is the one that receives some type of treatment, or stimulus.

39. If the mean is larger than the median for a data set, the distribution of that data set is likely skewed to the left.

40. Standard deviation is not resistant to outliers.

41. Outliers should be removed from the data before running any analysis because they will change our results.

42. As the variability in a data set decreases, the standard deviation will get closer to 0.

43. Removing the smallest data point will leave the IQR unchanged.

44. If two data sets have the same mean and median, then they must be identical.

45. If a data point has high leverage, it must be influential.

46. If a data point has high leverage, it must be an outlier.

47. If a data point is an outlier, it must have high leverage.

48. A correlation coefficient of 0 means there is no relationship between X and Y.

49. The closer the coefficient of determination is to 100%, the better the predictive power of our linear model.

50. If the correlation coefficient is -1, then the slope of the regression line is -1.

51. The closer the correlation coefficient to 0, the smaller the association between x and y.

52. Leverage indicates a data point has an x value far away from the mean of x.

53. Fanning in a residual plot indicates a linear model may not be appropriate for the data.

54. If $R^2$ is 81%, the correlation coefficient for the line must be .9.

55. Influential points are any points in the data set whose removal will result in a significantly different regression model.