Student Name: ____________________ Date: __________ Class Period: ________

EOCT Review – Statistics and Probability –

You need to be able to answer questions about each of the following concepts:

A. Measures of center and variability

- **Calculate mean, mean absolute deviation, median, quartiles, interquartile range, range**
- **compare 2 or more data sets using any of the above calculations**
- **read and understand boxplots and histograms – to obtain any of the calculations above**

B. Read and interpret a 2-way frequency table

- **Calculate marginal and relative frequencies**

C. Describe the correlation of 2 variables

- **Discuss correlation versus causation**
- **Approximate the correlation value, r, for given data or scatterplot**

D. Approximate a line of best fit

- **Determine if a model for data is linear or exponential**
- **Make predictions based on the line of best fit**
- **Calculate residuals**

A.

Mean = $\frac{\textit{sum of all values}}{\textit{total \# of values}}$; **MAD** = $\frac{\sum|x_i-\bar{x}|}{n}$

Example: Calculate the mean and mean absolute deviation (MAD). Data: 5, 6, 11, 3, 8

Median, 1st quartile, 3rd quartile – list # in order
Median = middle # when listed in order; if there are 2 #'s in the middle find the mean of those 2 #.
1st quartile, Q_1 = middle of the bottom half
3rd quartile, Q_3 = middle of the top half
IQR = $Q_3 - Q_1$

Example: Calculate the median, upper and lower quartiles, and the interquartile range.
Data: 62, 75, 77, 80, 81, 85, 87, 91, 94

Example: **These lists show the heights, in inches, of 12 players on a basketball team and 12 players on a soccer team.**

Basketball Players:
73, 74, 70, 72, 75, 73, 72, 70, 75, 77, 68, 72

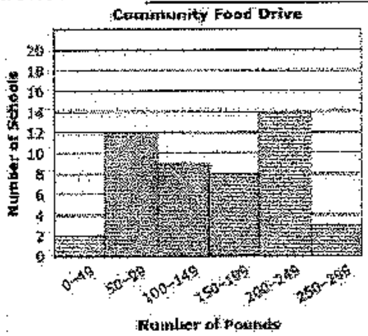
Soccer Players:
68, 70, 69, 71, 73, 72, 75, 70, 71, 68, 70, 69

Which statement BEST explains the variability of the two sets?

A. The heights of the soccer players are more variable because the median of the soccer team is greater.
B. The heights of the basketball players are more variable because the median of the basketball team is greater.
C. The heights of the soccer players are more variable because the interquartile range of the soccer team is greater.
D. The heights of the basketball players are more variable because the interquartile range of the basketball team is greater.

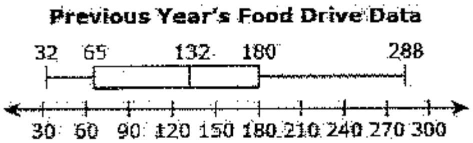
Example: All of a city's schools participated in a food-collection drive to stock the community food pantry. Each school reported the number of pounds of food collected. The histogram shows the results of the community food drive.

Student Name: ______________________ Date: ____________ Class Period: __________



a) In which interval is the median number of pounds collected by the schools located? Explain your answer.

b) The same schools participated in the food-collection drive the previous year. The box plot gives the results for the previous year.



Estimate the difference, in pounds, between the medians of the two sets of data. Could the two medians be equal? Justify your answer.

B.

Two-way frequency table – way to organize data that can be categorized by two variables.

Joint relative frequencies – values in each category divided by the total # of values in data set

Marginal relative frequencies – adding joint relative frequencies in each row and column

Conditional relative frequencies – dividing the joint relative frequency by the marginal relative frequency

Example: The two-way frequency table below shows the data from 20 adults and children asked if they like ice cream.

	Yes	No
Children	3	8
Adults	7	2

a) What is the probability of someone liking ice cream given that they are a child?

b) What is the probability that someone is an adult given that they like ice cream?

c) What is the probability that someone (child or adult) does not like ice cream?

C.

Correlation – how closely a data set models a linear relationship.

r = correlation coefficient

if r is close to +1, strong positive correlation

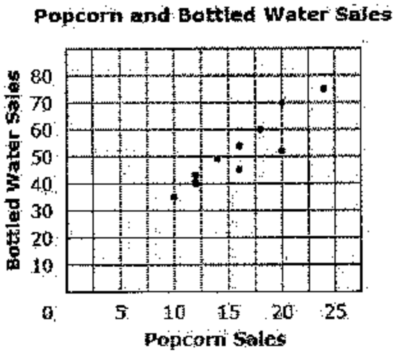
if r is close to – 1, strong negative correlation

if r is close to 0, weak correlation (pos or neg)

*** In order to show causation, an experiment must be conducted. It does not matter if there is a strong correlation. Correlation does not imply causation unless you have conducted a study to indicate it.**

Example:

The scatter plot shows the relationship between the popcorn sales and the bottled water sales for each basketball game.



Which statement can be concluded from the data?

A. The data has a weak correlation and demonstrates causation.

B. The data has a strong correlation and demonstrates causation.

C. The data has a weak correlation and does not demonstrate causation.

D. The data has a strong correlation and does not demonstrate causation.

D.

Line of best fit – look for characteristics of the line to approximate the equation (slope, y-intercept).

- Be able to determine when the scatterplot indicates that an exponential function is the best model instead of a linear function.

- Make predictions for a model by substituting values into the regression equation.

- To calculate residuals, determine the predicted values using the given equation.

Residual = observed – predicted

Example: The table shows the amounts Anita earned for babysitting.

Student Name: ______________________ Date: ____________ Class Period: __________

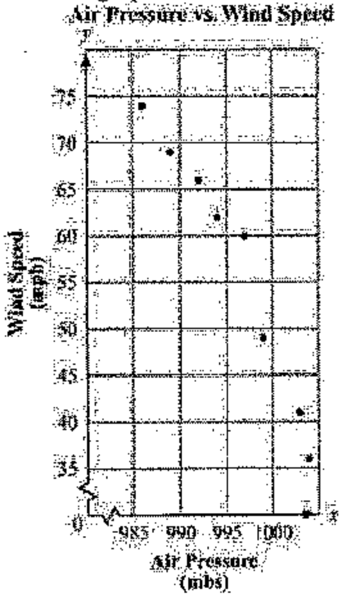
Anita's Babysitting Data

Hours (x)	Amount Earned (y)
2	$15
3	$25
5	$40
7	$55

a) A linear model for the data is $y = 7.88135x + 0.25423$. What do the slope and y -intercept of the linear model represent in the context of the given data?

b) Analyze the residuals for the data shown in the table. Show your work and explain your answer.

Example: A meteorologist recorded the wind speed and air pressure at different times during a storm in 2002. The data are shown in this graph.

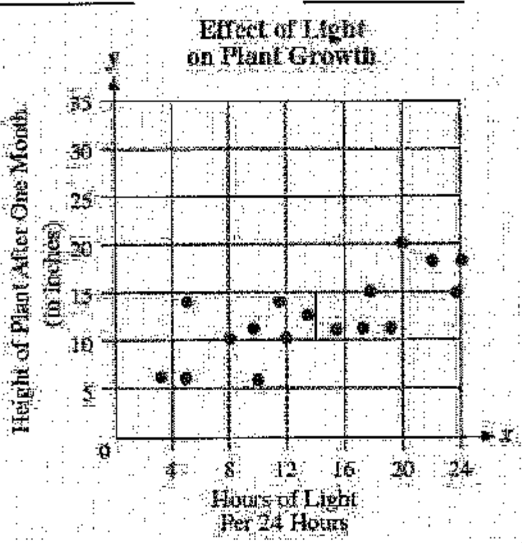


Which linear equation BEST models these data?

A. $y = -1.292x + 1347$ B. $y = -1.435x + 1489$

C. $y = -2.070x + 2118$ D. $y = -3.099x + 3148$

Example: Jenny studied the effect of light on plant growth. She graphed a scatterplot to represent her data.

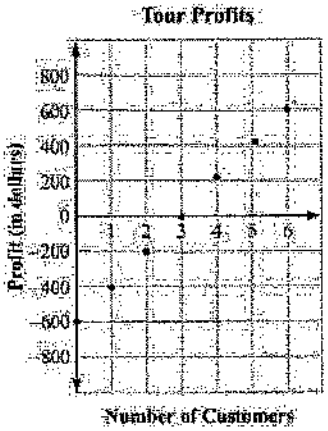


Which of the following **best** represents the equation for the line of best fit for the data shown?

A. $y = -0.4x + 5$ B. $y = 0.4x + 5$

C. $y = -4x + 5$ D. $y = 4x + 5$

Example: This graph shows the relationship between a touring company's profit and the number of customers on a tour for up to 6 customers.



What does the graph's x-intercept represent in this situation?

A. the rate of change of the company's profit

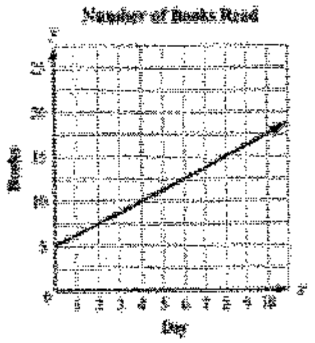
B. the amount of money the company spent on the tour

C. the number of customers needed for the company to break even

D. the number of customers needed for the company to make a profit

Additional notes:

Juan and Patti decided to see who could read the most books in a month. They began to keep track after Patti had already read 5 books that month. This graph shows the number of books Patti read for the next 10 days.



If Juan has read no books before the fourth day of the month and he reads at the same rate as Patti, how many books will he have read by day 12?

A. 5 B. 10

C. 15 D. 20

Data Analysis

This table shows the average low temperature, in °F, recorded in Macon, GA and Charlotte, NC, over a six-day period.

Day	1	2	3	4	5	6
Temperature, in °F, in Macon, GA	71	72	66	69	71	73
Temperature, in °F, in Charlotte, NC	69	64	68	74	71	75

Which conclusion can be drawn from the data?

A. The interquartile range of the temperatures is the same for both cities.
B. The lower quartile for the temperatures in Macon is lower than the lower quartile for the temperatures in Charlotte.
C. The mean and median temperatures of Macon were higher than the mean and median temperatures in Charlotte.
D. The upper quartile for the temperatures in Charlotte was lower than the upper quartile for temperatures in Macon.

A reading teacher recorded the number of pages read in an hour by each of her students. The numbers are shown here: 44, 49, 39, 43, 50, 44, 45, 49, 51. For this data which summary statistic is NOT correct?

A. The minimum is 39. B. The lower quartile is 44.
C. The median is 45. D. The maximum is 51.

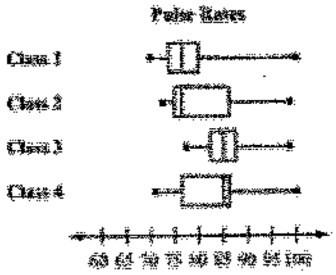
A school was having a coat drive for a local shelter. A teacher determined the median number of coats collected per class and the interquartile ranges of the number of coats collected per class for the freshmen and sophomores.

- The freshmen collected a median number of coats per class of 10, and the interquartile range was 6.
- The sophomores collected a median number of coats per class of 10, and the interquartile range was 4.

Which range of numbers includes the third quartile of coats collected for both classes?

A. 4 to 14 B. 6 to 14 C. 8 to 15 D. 12 to 15

A science teacher recorded the pulse rates for each of the students in her classes after the students had climbed a setoff stairs. She displayed the results, by class, using the box plots shown.



Which class had the highest pulse rates after climbing the stairs?

A. Class 1
B. Class 2
C. Class 3
D. Class 4

Peter went bowling, Monday to Friday, two weeks in a row. He only bowled one game each time he went. He kept track of his scores below.

Week 1: 70, 70, 70, 73, 75
Week 2: 72, 64, 73, 73, 75

What is the best explanation of why Peter's Week 2 mean score was lower than his Week 1 mean score?

A. Peter received the same score three times in Week 1.
B. Peter had one very bad score in Week 2.
C. Peter did not improve as he did the first week.
D. Peter had one very good score in Week 1.

A teacher determined the median scores and interquartile ranges of scores for a test she gave to two classes.

- In Class 1, the median score was 70 points, and the interquartile range was 15 points.
- In Class 2, the median score was 75 points, and the interquartile range was 12 points.

Which range of numbers includes only third quartile of scores for both classes?

A. 70 to 87 points
B. 70 to 85 points
C. 75 to 87 points
D. 75 to 85 points

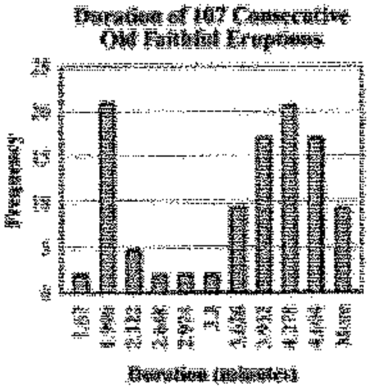
This table shows admission price for various museums in the same city.

Museum Prices				
$9.00	$12.00	$9.75	$8.25	$11.25

Which is the mean absolute deviation for this set of data?

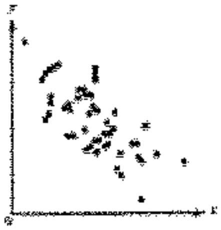
A. $1.26
B. $6.30
C. $10.05
D. $10.13

This histogram shows the frequency distribution of duration times for 107 consecutive eruptions of the Old Faithful geyser. The duration of an eruption is the length of time, in minutes, from the beginning of the spewing of water until it stops. What is the **BEST** description for the distribution?



A. bimodal
B. uniform
C. multi-outliers
D. skewed to the right

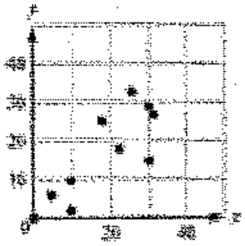
How would you describe the correlation of the two variables based on the scatter plot?



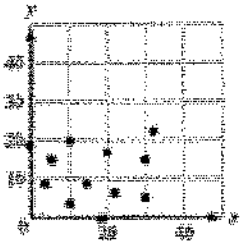
A. positive, strong linear
B. negative, weak linear
C. negative, fairly strong linear
D. little or no correlation

Which graph displays a set of data for which a linear function is the model of best fit?

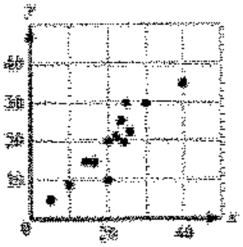
A.



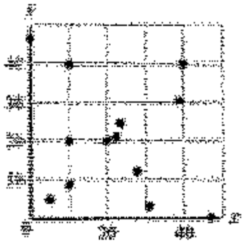
B.



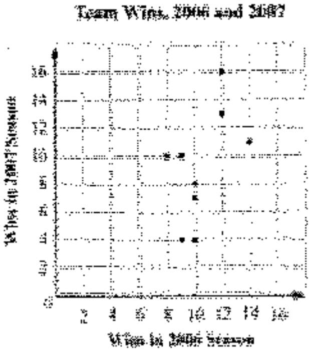
C.



D.



This graph plots the number of wins in the 2006 and 2007 seasons for a sample of professional football teams.



Which equation BEST represents a line that matches the trend of this data?

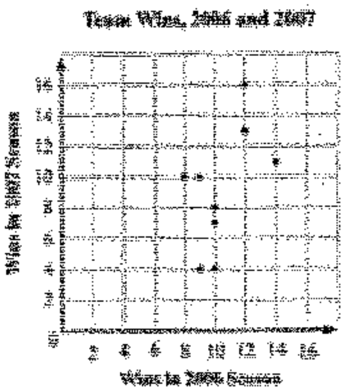
A. $y = \frac{1}{2}x$

B. $y = \frac{1}{2}x + 8$

C. $y = 2x - 6$

D. $y = 2x - 12$

This graph plots the number of wins in the 2006 and 2007 seasons for a sample of professional football teams.



Based on the regression model, what is the predicted number of 2007 wins for a team that won 5 games in 2006?

A. 3

B. 4

C. 5

D. 6