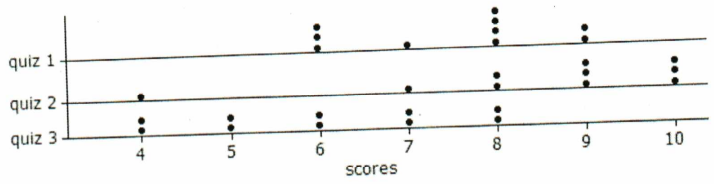
Name: _____ Date: _____

Unit 6: Data and Statistics

Review

1) The scores of three quizzes are shown in the following data plot for a class of 10 students. Each quiz has a maximum possible score of 10. Possible dot plots of the data are shown below.



a. On which quiz did students tend to score the lowest? Justify your choice. Quiz 3

Its median are lowest

No students scored above an 8

b. Without performing any calculations, which quiz tended to have the most variability in the students' scores? Justify your choice based on the graphs.

Quiz 3 Un. Form & spread out

Quiz 1 & Quiz 2 are clustered

c. If you were to calculate a measure of variability for Quiz 2, would you recommend using the interquartile range or the standard deviation? Explain your choice.

IQR → median St. dev. mean

Since data is skewed median is a better measure of center so IQR is better

d. For Quiz 3, move one dot to a new location so that the modified data set will have a larger standard deviation than before you moved the dot. Be clear which point you decide to move, where you decide to move it, and explain why. State the standard deviation for the original data set, and for your new data set.

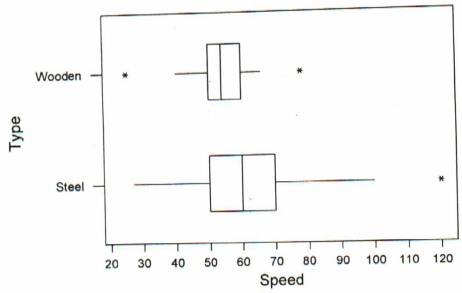
Need a larger spread move 6 to 10.

Moving the point further from the $\bar{x}$ would make the standard deviation higher

Name: _____ **Review**

Unit 6: Data and Statistics

2) The box plots below display the distributions of maximum speed for 145 roller coasters in the United States, separated by whether they are wooden coasters or steel coasters.



Based on the box plots, answer the following questions or indicate that you do not have enough information.

Which type of coaster has more observations?
A. Wooden
B. Steel
C. About the same
D. Cannot be determined

Explain your choice:
Specific parts not displayed in a box plot.

a. Which type of coaster has a higher percentage of coasters that go faster than 60 mph?
A. Wooden
B. Steel
C. About the same
D. Cannot be determined

Explain your choice:
50% of steel coasters have speed > 60mph vs 25% of wooden coasters.

b. Which type of coaster has a higher percentage of coasters that go faster than 50 mph?
A. Wooden
B. Steel
C. About the same
D. Cannot be determined

Explain your choice:
1st Quartile is in the same place for both.

c. Which type of coaster has a higher percentage of coasters that go faster than 48 mph?
A. Wooden
B. Steel
C. About the same
D. Cannot be determined

Explain your choice:
between the min & 1st Quartile for both data sets

d. Write 2–3 sentences comparing the two types of coasters with respect to which type of coaster normally goes faster.
A large % (50% vs 25%) of steel coasters go faster. The [illegible] coaster is an outlier

]]]

Unit 6: Data and Statistics — Review

3) Fifty moviegoers were surveyed about their favorite movie types.

- 15 men and 6 women chose "Action" as their favorite type
- 9 men and 10 women chose "Drama" as their favorite type
- 6 men and 4 women chose "Comedy" as their favorite type

a) Use the table below to construct a two-way frequency table.

Favorite Movie Types				
	Action	**Drama**	**Comedy**	**Total**
Men	15	9	6	30
Women	6	10	4	20
Total	21	19	10	50

b) Find the relative frequencies to compare and describe the survey.

Favorite Movie Types				
	Action	**Drama**	**Comedy**	**Total**
Men	$\frac{15}{50}$ 0.30	$\frac{9}{50}$ 0.18	$\frac{6}{50}$ 0.12	$\frac{30}{50}$ 0.6
Women	$\frac{6}{50}$ 0.12	$\frac{10}{50}$ 0.20	$\frac{4}{50}$ 0.08	$\frac{20}{50}$ 0.4
Total	$\frac{21}{50}$ 0.42	$\frac{19}{50}$ 0.38	$\frac{10}{50}$ 0.20	$\frac{50}{50}$ 1.

c) Find the row conditional relative frequencies to compare and describe the survey.

Favorite Movie Types				
	Action	**Drama**	**Comedy**	**Total**
Men	$\frac{15}{30}$ 0.5	$\frac{9}{30}$ 0.3	$\frac{6}{30}$ 0.2	$\frac{30}{30}$ 1.0
Women	$\frac{6}{20}$ 0.3	$\frac{10}{20}$ 0.5	$\frac{4}{20}$ 0.2	$\frac{20}{20}$ 1.0
Total	$\frac{21}{50}$ 0.42	$\frac{19}{50}$ 0.38	$\frac{10}{50}$ 0.2	1.0

d) Find the column conditional relative frequencies to compare and describe the survey.

Favorite Movie Types				
	Action	**Drama**	**Comedy**	**Total**
Men	$\frac{15}{21}$ 0.71	$\frac{9}{19}$ 0.47	$\frac{6}{10}$ 0.6	$\frac{30}{50}$ 0.6
Women	$\frac{6}{21}$ 0.29	$\frac{10}{19}$ 0.53	$\frac{4}{10}$ 0.4	$\frac{20}{50}$ 0.4
Total	$\frac{21}{21}$ 1.0	$\frac{19}{19}$ 1.0	$\frac{10}{10}$ 1.0	$\frac{50}{50}$ 1.0

e) What movie type do the respondents prefer? How do you know?

42% of people surveyed prefer action. Also almost at 38% is drama
So% of males prefer action while 50% women like drama best
of genders specifically

f) Is there an association between movie type and gender? How do you know? Does gender cause an individual to have a particular movie preference? Explain.

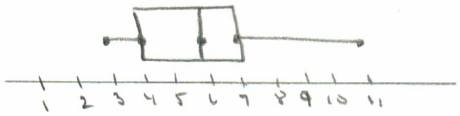
Not for comedies both 20% of men & 20% of women
prefer comedies

6) Twenty-two students were surveyed about the number of days they played outside in one month. The results of this survey are shown below.

3, 3, 4, 4, 4, 4, 5, 5, 5, 6, 6, 6, 6, 6, 7, 7, 7, 8, 9, 9, 10, 11

a. Use your calculator to find the Mean, Median, Q1 and Q3. On the grid below, create a **dot plot** based on the data. Be sure to check for outliers. # days play outside

$\bar{y} = 6.14$
$Q_1 = 4$
$Med = 6$
$Q_3 = 7$
max 11
min 3



Outliers: $1.5 \times IQR$
$1.5(7-4) = 4.5$
$1^{st}Q - 1.5IQR = 4 - 4.5 = -0.5$ ✗
$3^{rd}Q + 1.5IQR = 7 + 4.5 = 11.5$ ✗
no data points are outside these #

b. Identify the typical number of days spent outside by the twenty-five students.

median → 6 days

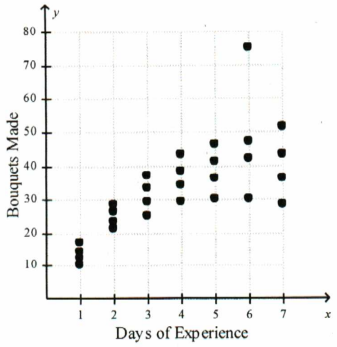
c. Use the statistical features of your calculator to find the standard deviation of the data set (round to the nearest hundredth).

Standard Deviation: 2.21
Sx on calculator

d. Write a sentence Interpreting Standard Deviation in the context of the question.

Typically the # of days played outside will vary 2.21 days from the mean (68% of data will fall between 6.14 ± 2.21)

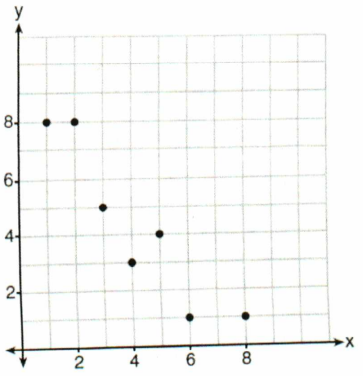
4) A floral delivery company conducts a study to measure the effect of worker experience on productivity. Tell whether the scatter plot appears to have a linear or non-linear pattern of association.



non-linear

5) Multiple Choice – Pick the BEST answer.

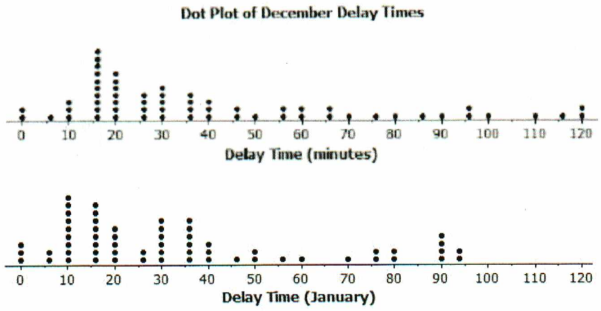
What is the correlation coefficient of the linear fit of the data shown below, to the *nearest hundredth*?



Negative as x ↑ y ↓ not perfect since all the points do NOT lie on the line

(1) 1.00
(2) 0.93
(3) −0.93
(4) −1.00

7) Transportation officials collected data on sixty flight delays in the month of December and sixty flight delays in the month of January.



Which measure of variability would be appropriate in describing the typical flight delays in December and January (pick from the Standard Deviation of the Interquartile Range)? Why? Explain the meaning of this measure of variability in the context of the question.

Delay time skewed right → median ∴ IQR
(measure of center) (measure of variability)

8) Find the standard deviation BY hand for the data given below.

10, 15, 18, 19, 21, 26, 28

$n = 7$ $\frac{137}{7} = 19.57 = \bar{x}$

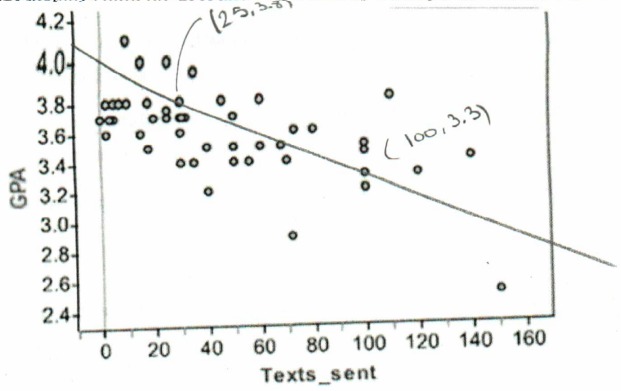
$$= \sqrt{\frac{(10-19.57)^2 + (15-19.57)^2 + (18-19.57)^2 + (19-19.57)^2 + (21-19.57)^2 + (26-19.57)^2 + (28-19.57)^2}{7-1}}$$

$$= \sqrt{\frac{229}{6}}$$

$$= 6.18$$

Review

...high school students:



a) There appears to be a linear relationship between the variables. Describe the relationship (strong positive, strong negative, etc.) moderate negative

b) Draw what you think may be a good line of best fit on the graph above AND find the equation of your line BY HAND. Answers will vary (25, 3.8) (100, 3.3)

$m = \frac{3.8-3.3}{25-100} = \frac{.5}{-75} \approx -0.007$

$y - 3.8 = -0.007(x - 25)$

$y = 0.007x + 3.975$

← not modeling exact line - estimation error

c) What would you estimate is the correlation coefficient for the data set. Explain why you picked your estimate. ~ -0.5 or -0.6 not a strong relationship

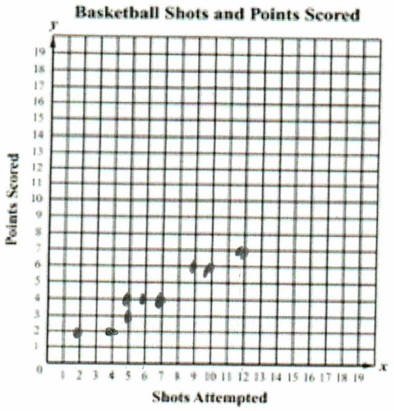
d) Explain what a residual is and what it means. What does the sign (positive or negative) of the residual tell you about the actual versus the predicted value? What does the sign tell you about the location of the actual value relative to the regression line?

Residual is the actual data value − predicted value (using ...

The table below shows the number of shots attempted by a basketball player in ten games and the number of points scored as a result of the attempted shots.

Basketball Shots and Points Scored

Shots Attempted (x)	4	7	12	5	2	5	10	6	9	5
Points Scored (y)	2	4	7	3	2	3	6	4	6	4



a) Make a scatterplot of the data.

b) Use your calculator to find the linear regression equation for this data set. Write the equation below.

$\hat{y} = 0.55x + 0.52$

c) Calculate the residual for the least squares regression line for an x value of 7.

predicted = 4.37 actual 4

residual = actual − predicted
= 4 − 4.37
= −0.37 below line

c) What is the slope of the linear regression equation? Interpret the slope in context.

0.55 For every shot attempted 0.55 add'l points are scored. ≈ 2 attempts results in 1 more point.

d) Use your calculator to find the correlation coefficient. Interpret the correlation coefficient.

$r = 0.966$ Very strong, almost perfect, positive relationship between # shots and # points scored

e) Does the number of shots attempted have a causal relationship with the number of points earned? Explain. The # of shots taken is highly correlated to the # of points scored. It isn't necessarily causal other factors - defense of other team...