Name _____ Date _____

1. A recent social survey asked 654 men and 813 women to indicate how many "close friends" they have to talk about important issues in their lives. Below are frequency tables of the responses.

Number of Close Friends	0	1	2	3	4	5	6	Total
Males	196	135	108	100	42	40	33	654
Females	201	146	155	132	86	56	37	813

a. The shape of the distribution of the number of close friends for the males is best characterized as

A. Skewed to the higher values (right or positively skewed).
B. Skewed to the lower values (left or negatively skewed).
C. Symmetric.

b. Calculate the median number of class friends for the females. Show your work.

c. Do you expect the mean number of close friends for the females to be larger or smaller than the median you found in part (b), or do you expect them to be the same? Explain your choice.

d. Do you expect the mean number of close friends for the males to be larger or smaller than the mean number of close friends for the females, or do you expect them to be the same? Explain your choice.

2. The physician's health study examined whether physicians who took aspirin were less likely to have heart attacks than those who took a placebo (fake) treatment. The table below shows their findings.

	Placebo	Aspirin	Total
Heart attack	189	104	293
No heart attack	10,845	10,933	21,778
Total	11,034	11,037	22,071

Based on the data in the table, what conclusions can be drawn about the association between taking aspirin and whether or not a heart attack occurred? Justify your conclusion using the given data.

3. Suppose 500 high school students are asked the following two questions:

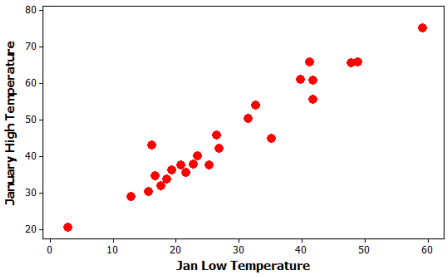
- What is the highest degree you plan to obtain? (check one)
  - ☐ High school degree ☐ College (Bachelor's degree)
  - ☐ Graduate school (e.g., Master's degree or higher)
- How many credit cards do you currently own? (check one)
  - ☐ None ☐ One ☐ More than one

Consider the data shown in the following frequency table.

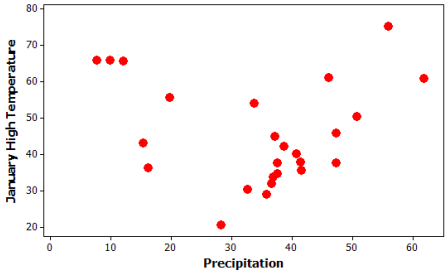
	No Credit Cards	One Credit Card	More than One Credit Card	Total
High school	?		6	59
College	120	240	40	394
Graduate school				47
Total		297		500

Fill in the missing value in the cell in the table that is marked with a "?" so that the data would be consistent with no association between education aspiration and current number of credit cards for these students. Explain how you determined this value.

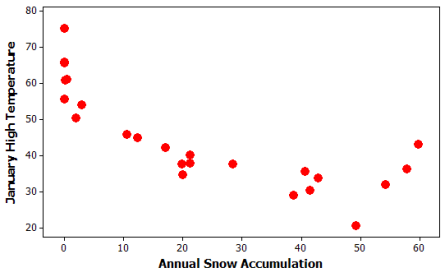
4. Weather data were recorded for a sample of 25 American cities in one year. Variables measured included January high temperature (in degrees Fahrenheit), January low temperature (in degrees Fahrenheit), annual precipitation (in inches), and annual snow accumulation. The relationships for three pairs of variables are shown in the graphs below (Jan. Low Temperature—Graph A; Precipitation—Graph B; Annual Snow Accumulation—Graph C).



Graph A



Graph B



Graph C

a. Which pair of variables will have a correlation coefficient closest to 0?

A. Jan. high temperature and Jan. low temperature
B. Jan. high temperature and precipitation
C. Jan. high temperature and snow accumulation

Explain your choice:

b. Which of the above scatterplots would be best described as a strong nonlinear relationship? Explain your choice.

c. Suppose we fit a least squares regression line to Graph A. Circle one word choice for each blank that best completes this sentence, based on the equation:

If I compare a city with a January low temperature of 30°F to a city with a higher January low temperature, then the __(1)__ January high temperature of the second city will __(2)__ be __(3)__.

(1) actual, predicted

(2) probably, definitely

(3) smaller, larger, the same, equally likely to be higher or lower

d. For the city with a January low temperature of 30°F, what do you predict for the annual snow accumulation? Explain how you are estimating this based on the three graphs above.

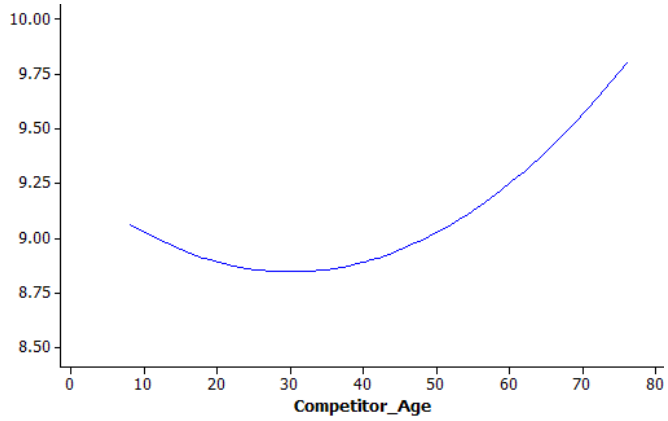
5. Suppose times (in minutes) to run one mile were recorded for a sample of 100 runners, ages 16–66 years, and the following least squares regression line was found.

   Predicted time in minutes to run one mile $= 5.35 + 0.25 \times (age)$

   a. Provide an interpretation in context for this slope coefficient.

   b. Explain what it would mean in the context of this study for a runner to have a negative residual.

   c. Suppose, instead, that someone suggests using the following curve to predict time to run one mile. Explain what this model implies about the relationship between running time and age, and why that relationship might make sense in this context.



d. Based on the results for these 100 runners, explain how you could decide whether the first model or the second model provides a better fit to the data.

e. The sum of the residuals is always equal to zero for the least squares regression line. Which of the following must also always be equal to zero?

A. The mean of the residuals
B. The median of the residuals
C. Both the mean and the median of the residuals
D. Neither the mean nor the median of the residuals

A Progression Toward Mastery					
Assessment Task Item		**STEP 1** Missing or incorrect answer and little evidence of reasoning or application of mathematics to solve the problem	**STEP 2** Missing or incorrect answer but evidence of some reasoning or application of mathematics to solve the problem	**STEP 3** A correct answer with some evidence of reasoning or application of mathematics to solve the problem, or an incorrect answer with substantial evidence of solid reasoning or application of mathematics to solve the problem	**STEP 4** A correct answer supported by substantial evidence of solid reasoning or application of mathematics to solve the problem
1	**a** S-ID.A.3	Student selects B or C.	N/A	N/A	Student selects A.
	b S-ID.A.2	Student focuses on the counts and not on the number of siblings.	Student calculates the correct location (407) but fails to convert that to a value. OR Student only finds the median of 0–6, ignoring the tally information.	Student provides a reasonable value (2) but does not clearly explain how it was found.	Student calculates the correct location (407) and then uses the tally information to find that number (2).
	c S-ID.A.3	Student does not make a clear choice or justification.	Student does not make a choice based on skewedness of the distribution.	Student makes a choice based on skewedness but fails to explain how the table provides information indicating that the female distribution is skewed to the right.	Student selects the mean to be larger based on the skewedness apparent in the distribution from the initial large counts, which decrease with number of friends.
	d S-ID.A.3	Student fails to compare the two distributions based on the table.	Student fails to use the frequency information in comparing the two distributions.	Student selects the females to have the higher mean based on the higher counts but does not consider the higher number of females (all frequencies are higher).	Student selects the females to have the higher mean based on the higher relative frequencies at the higher values (e.g., $\frac{86+56+37}{813} \approx 22\%$ vs. $\frac{42+40+33}{654} \approx 18\%$.

2	**S-ID.B.5**	Student does not address the association between the two variables.	Student focuses only on frequencies and not proportions.	Student appears to use appropriate conditional proportions but does not fully justify his or her approach.	Student uses appropriate conditional proportions (e.g., $\frac{189}{11034}$ vs. $\frac{104}{11037}$) to compare the two groups.
3	**S-ID.B.5**	Student does not complete table.	Student fills in some values without enough justification to allow completion of the table.	Student understands need to equalize the conditional distributions but is not able to perform the algebra to do so.	Student's approach equates the conditional proportions across the columns or equates conditional proportions across rows.
4	**a** **S-ID.B.6** **S-ID.C.8**	Student does not make a consistent choice or give an explanation.	Student chooses C because of the clear nonlinear association.	Student chooses B and gives an explanation consistent with looking for the strongest association.	Student chooses B and gives an explanation focusing on the lack of an apparent linear association.
	b **S-ID.B.6** **S-ID.C.8**	Student does not make a consistent choice or give an explanation.	Student chooses B based on a lack of association.	Student chooses A and justifies the choice based on a strong linear association.	Student chooses C based on the dots following a clear curve.
	c **S-ID.B.6** **S-ID.C.8**	Student gets one or none of the answers correct.	Student answers two of the three correctly (see Step 4 answer).	N/A	Student circles (1) predicted; (2) definitely; (3) larger.
	d **S-ID.B.6** **S-ID.C.8**	Student does not provide an estimate.	Student only uses Graph C with a January high of about 30.	Student attempts to integrate information across two or more graphs but fails to provide a reasonable estimate.	Student's prediction finds a January high temperature from Graph A and then uses that with Graph C to estimate a snow fall amount.
5	**a** **S-ID.C.7**	Student focuses on intercept.	Student reverses the role of the explanatory and response variables.	Student does not use context and/or does not interpret in terms of *predicted* time.	Student correctly addresses predicted change in time with a (one-unit) change in age (in context).
	b **S-ID.B.6**	Student does not define residual in terms of predicted vs. actual.	Student reverses positive and negative residuals.	Student compares predicted and actual value correctly but not in context.	Student places the actual value below the predicted value and relates it to context.

	c **S-ID. B.6**	Student does not interpret the model.	Student does not address the curved nature of the model.	Student discusses times decreasing and then increasing but fails to provide a justification or does not relate the justification to the graph.	Student interprets the decrease and then increase in context and discusses this in terms of the age of the runners.
	d **S-ID.B.6**	Student response does not address seeing how well the model fits the data.	Student does not focus on residuals.	Student focuses on residuals but does not explain how they would be used.	Student focuses on examination of residuals as a way to measure model fit and addresses whether there is a pattern to the residuals or the overall sizes of the residuals.
	e **S-ID.A.2** **S-ID.B.6**	Student selects B, C, or D.	N/A	N/A	A response that indicates if the sum is zero, the mean must be zero, but median may not necessarily equal zero.

Name _____ Date _____

1. A recent social survey asked 654 men and 813 women to indicate how many "close friends" they have to talk about important issues in their lives. Below are frequency tables of the responses.

Number of Close Friends	0	1	2	3	4	5	6	Total
Males	196	135	108	100	42	40	33	654
Females	201	146	155	132	86	56	37	813

a. The shape of the distribution of the number of close friends for the males is best characterized as

A. Skewed to the higher values (right or positively skewed).
B. Skewed to the lower values (left or negatively skewed).
C. Symmetric.

b. Calculate the median number of class friends for the females. Show your work.

$$\frac{(813+1)}{2} = 407$$

$$201 + 146 = 347 + 155 = 502$$

407^{th} observation falls in 2 columns

2 close friends

c. Do you expect the mean number of close friends for the females to be larger or smaller than the median you found in part (b), or do you expect them to be the same? Explain your choice.

Mean should be larger than median because of skewedness.

d. Do you expect the mean number of close friends for the males to be larger or smaller than the mean number of close friends for the females, or do you expect them to be the same? Explain your choice.

From 2 on, females have higher counts than males in every column. This shows that females tend to have more friends. So, the male average is probably smaller than the female average.

2. The physician's health study examined whether physicians who took aspirin were less likely to have heart attacks than those who took a placebo (fake) treatment. The table below shows their findings.

	Placebo	Aspirin	Total
Heart attack	189	104	293
No heart attack	10,845	10,933	21,778
Total	11,034	11,037	22,071

Based on the data in the table, what conclusions can be drawn about the association between taking aspirin and whether or not a heart attack occurred? Justify your conclusion using the given data.

$\frac{118}{1{,}034} = 0.017 \qquad \frac{104}{11{,}037} = 0.0094$

The placebo group had higher proportion of heart attacks, although both numbers are pretty small.

3. Suppose 500 high school students are asked the following two questions:

   - What is the highest degree you plan to obtain? (check one)
     ☐ High school degree ☐ College (Bachelor's degree)
     ☐ Graduate school (e.g., Master's degree or higher)
   - How many credit cards do you currently own? (check one)
     ☐ None ☐ One ☐ More than one

   Consider the data shown in the following frequency table.

	No Credit Cards	One Credit Card	More than One Credit Card	Total
High school	?	y	6	59
College	120	240	40	394
Graduate school				47
Total	x	297	z	500

Fill in the missing value in the cell in the table that is marked with a "?" so that the data would be consistent with no association between education aspiration and current number of credit cards for these students. Explain how you determined this value.

$$\frac{?}{x} = \frac{y}{297} = \frac{6}{z} = \frac{59}{500}$$

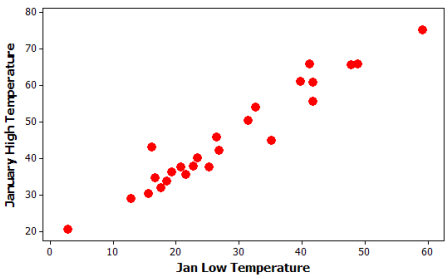
$$z = \frac{500 \times 6}{59} = 50.8 \approx 51$$

$$500 - 51 - 297 = 152$$

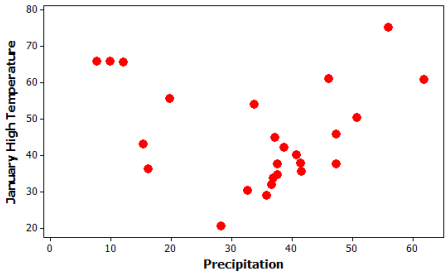
$$\frac{?}{152} = \frac{59}{500}$$

$$\text{So, } ? = 17.936 \approx 18$$

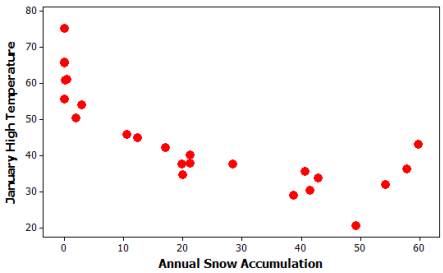
4. Weather data were recorded for a sample of 25 American cities in one year. Variables measured included January high temperature (in degrees Fahrenheit), January low temperature (in degrees Fahrenheit), annual precipitation (in inches), and annual snow accumulation. The relationships for three pairs of variables are shown in the graphs below (Jan. Low Temperature—Graph A; Precipitation—Graph B; Annual Snow Accumulation—Graph C).



Graph A



Graph B



Graph C

a. Which pair of variables will have a correlation coefficient closest to 0?

A. Jan. high temperature and Jan. low temperature
B. Jan. high temperature and Precipitation
C. Jan. high temperature and Snow accumulation

Explain your choice: *There is not much of a linear association, but lots of scatter.*

b. Which of the above scatterplots would be best described as a strong nonlinear relationship? Explain your choice.

Graph C has a strong nonlinear relationship because it has a curved pattern, and the dots follow the pattern pretty closely.

c. Suppose we fit a least squares regression line to Graph A. Circle one word choice for each blank that best completes this sentence based on the equation:

If I compare a city with a January low temperature of 30°F and a city with a higher January low temperature, then the __(1)__ January high temperature of the second city will __(2)__ be __(3)__.

(1) actual, **predicted** From the equation

(2) probably, **definitely**

(3) smaller, **larger**, the same, equally likely to be higher or lower

d. For the city with a January low temperature of 30°F, what do you predict for the annual snow accumulation? Explain how you are estimating this based on the three graphs above.

The annual snow accumulation will be about 10 inches because January's low of 30 corresponds to a January high of about 50 in Graph A, which matches with 10 inches in Graph C.

5. Suppose times (in minutes) to run one mile were recorded for a sample of 100 runners, ages 16–66 years, and the following least squares regression line was found.

Predicted time in minutes to run one mile $= 5.35 + 0.25 \times (age)$

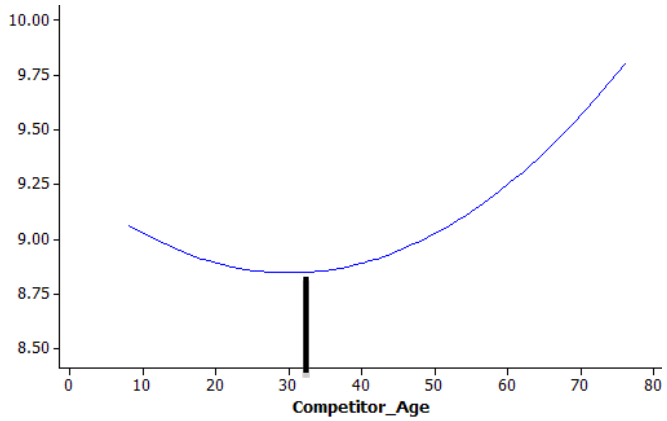
a. Provide an interpretation in context for this slope coefficient.

For every year older that a runner is, we predict the time to run one mile to increase by 0.25 seconds.

b. Explain what it would mean in the context of this study for a runner to have a negative residual.

The runner was even faster (lower time) than we could have predicted for that age.

c. Suppose, instead, that someone suggests using the following curve to predict time to run one mile. Explain what this model implies about the relationship between running time and age, and why that relationship might make sense in this context.



Runners get faster (lower times) until around age 30, when they start to slow down.

d. Based on the results for these 100 runners, explain how you could decide whether the first model or the second model provides a better fit to the data.

Look at the residuals and see which model (straight line or curve) provides a better match to the data.

e. The sum of the residuals is always equal to zero for the least squares regression line. Which of the following must also always be equal to zero?

A. The mean of the residuals
B. The median of the residuals
C. Both the mean and the median of the residuals
D. Neither the mean nor the median of the residuals