Chapter 6: Exploring Data: Relationships
Lesson Plan

For All Practical Purposes

Mathematical Literacy in Today's World, 9th ed.

- Displaying Relationships: Scatterplots
- Making Predictions: Regression Line
- Correlation
- Least-Squares Regression
- Interpreting Correlation and Regression

Displaying Relationships

- <u>Relationship Between Two Variables</u>
  - Examine data for two variables to see if there is a relationship between the variables. Does one influence the other?
  - Study both variables on the same individual.
  - If a relationship exists between variables, typically one variable influences or causes a change in another variable.
    - Explanatory variable explains, or causes, the change in another variable.
    - Response variable measures the outcome, or response to the change.

Response variable –
A variable that measures an outcome or result of a study (observed outcome).

Explanatory variable –
A variable that explains or causes change in the response variable.

Chapter 6: Exploring Data: Distributions
Displaying Relationships: Scatterplots

- Data to Be Used for a Scatterplot
  - A scatterplot is a graph that shows the relationship between two numerical variables, measured on the same individual.
    - Explanatory variable, ***x***, is plotted on the horizontal axis ***(x)***.
    - Response variable, ***y***, is plotted on the vertical axis ***(y)***.
    - Each pair of related variables ***(x, y)*** is plotted on the graph.

Example: A study is done to see how the number of beers that a student drinks predicts his/her blood alcohol content (BAC). *Results of 16 students:*

Explanatory variable, ***x*** = beers drunk

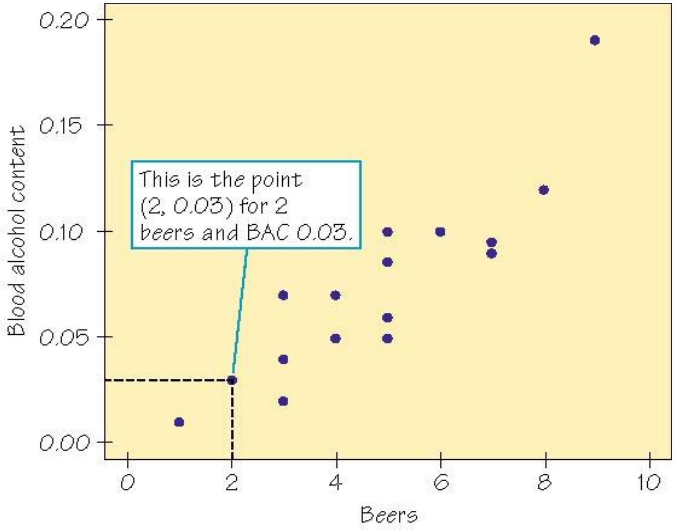
Response variable, ***y*** = BAC level

Student	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
Beers	5	2	9	8	3	7	3	5
BAC	0.10	0.03	0.19	0.12	0.04	0.095	0.07	0.06
Student	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**
Beers	3	5	4	6	5	7	1	4
BAC	0.02	0.05	0.07	0.10	0.85	0.09	0.01	0.05

Chapter 6: Exploring Data: Distributions
Displaying Relationships: Scatterplots

- Scatterplot
  - Example continued: The scatterplot of the blood alcohol content, **BAC**, (***y***, response variable) against the number of beers a young adult drinks (***x***, explanatory variable).
  - The data from the previous table are plotted as points on the graph ***(x, y)***.

BAC vs. number of beers consumed



Examining This Scatterplot…

1. What is the overall pattern (form, direction, and strength)?
   - Form – Roughly a straight-line pattern.
   - Direction – Positive association (both increase).
   - Strength – Moderately strong (mostly on line).
2. Any striking deviations (outliers)? Not here.

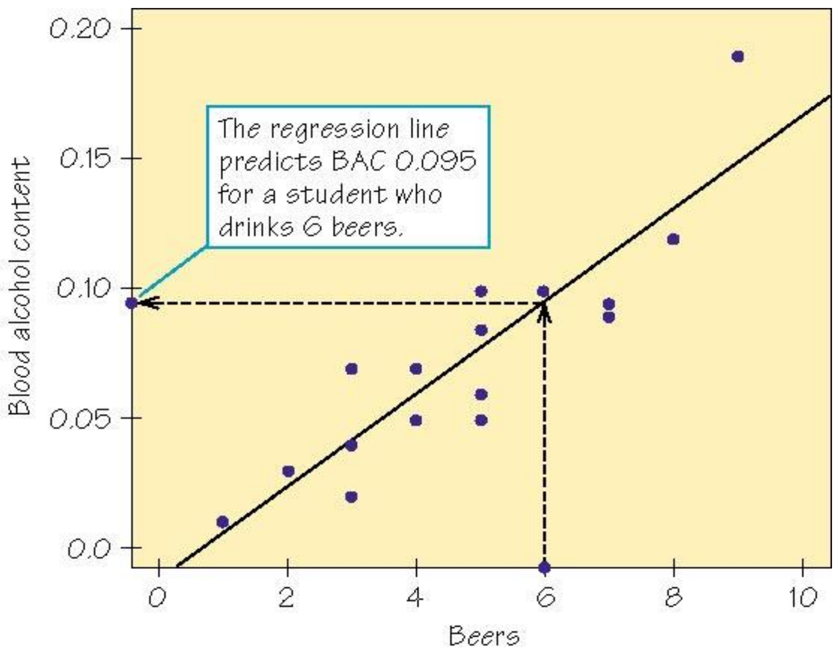
Outliers – A deviation in a distribution of a data point falling outside the overall pattern.

Regression Lines

Regression Line

- A straight line that describes how a response variable ***y*** changes as an explanatory variable ***x*** changes.
- Regression lines are often used to predict the value of ***y*** for a given value of ***x.***

BAC vs. number of beers consumed



A regression line has been added to be able to predict blood alcohol content from the number of beers a student drinks.

Graphically, you can predict that if ***x*** = 6 beers, then ***y*** = 0.95 BAC.

(Legal limit for driving in many states is BAC = 0.08.)

Chapter 6: Exploring Data: Distributions

Regression Lines

- <u>Using the Equation of the Line for Predictions</u>
  - It is easier to use the equation of the line for predicting the value of ***y***, given the value of ***x***.

    Using the equation of the line for the previous example:

    predicted BAC = −0.0127 + (0.01796)(beers)

    y = −0.127 + 0.01796 ***(x)***

    For a young adult drinking 6 beers **(*x* = 6):**

    predicted BAC = −0.0127 + 0.01796 **(6)** = 0.095

- <u>Straight Lines</u>
  - A straight line for predicting ***y*** from ***x*** has an equation of the form:

$$\hat{y} = mx + b$$

  - In this equation, ***m*** is the slope, the amount by which ***y*** changes when ***x*** increases by 1 unit.
  - The number ***b*** is the *y*-intercept, the value of ***y*** when ***x*** = 0.

Correlation

- Correlation, *r*
  - Measures the direction and strength of the straight-line relationship between two numerical variables.
  - A correlation ***r*** is always a number between −1 and 1.
  - It has the same sign as the slope of a regression line.
    - ***r*** > 0 for positive association (increase in one variable causes an increase in the other).
    - ***r*** < 0 for negative association (increase in one variable causes a decrease in the other)

Correlation

- <u>Correlation, *r*</u>
  - Perfect correlation $r = 1$ or $r = -1$ occurs only when all points lie exactly on a straight line.
  - The correlation moves away from 1 or −1 (toward zero) as the straight-line relationship gets weaker.
  - Correlation $r = 0$ indicates no straight-line relationship.

Correlation

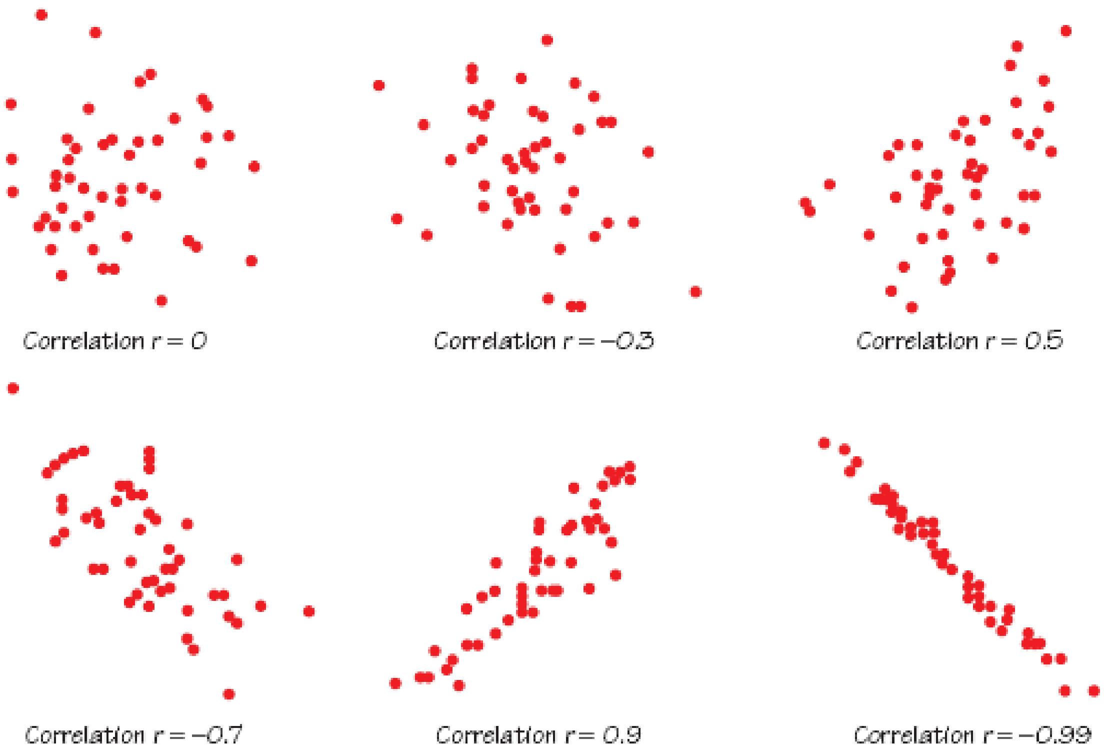
- <u>Correlation</u>
  - Correlation is strongly affected by a few outlying observations. (Also, the mean and standard deviation are affected by outliers.)

- <u>Equation of the Correlation</u>
  - To calculate the correlation, suppose you have data on variable ***x*** and ***y*** for ***n*** individuals.
  - From the data, you have the values calculated for the means and standard deviations for ***x*** and ***y***.
    - The means and standard deviations for the two variables are $\bar{x}$ and s_x for the ***x***-values, and $\bar{y}$ and s_y for the ***y***-values.
  - The correlation r between ***x*** and ***y*** is:

$$r = \frac{1}{n-1}\left[\frac{\left(x_1-\bar{x}\right)}{s_x}\frac{\left(y_1-y\right)}{s_y}+\frac{\left(x_2-\bar{x}\right)}{s_x}\frac{\left(y_2-y\right)}{s_y}+\cdots+\frac{\left(x_n-\bar{x}\right)}{s_x}\frac{\left(y_n-y\right)}{s_y}\right]$$

Correlation

- <u>Correlation</u>
  - The scatterplots below show examples of how the correlation ***r*** measures the direction and strength of a straight-line association.



Least-Squares Regression

- Least-Squares Regression Line
  - A line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.

- Equation of the Least-Squares Regression Line
  - From the data for an explanatory variable ***x*** and a response variable ***y*** for ***n*** individuals, we have calculated the means $\bar{x}, \bar{y}$, and standard deviations s_x, s_y, as well as their correlation ***r***.

The least-squares regression line is the line:

Predicted $\hat{y} = mx + b$

With slope … $m = r\frac{s_y}{s_x}$

And *y*-intercept … $b = \bar{y} - m\bar{x}$

This equation was used to calculate the line for predicting BAC for number of beers drunk.

Predicted
y = −0.0127 + 0.01796 ***x***

Interpreting Correlation and Regression

- A Few Cautions When Using Correlation and Regression
  - Both the correlation ***r*** and least-squares regression line can be strongly influenced by a few outlying points.
    - Always make a scatterplot before doing any calculations.
  - Often the relationship between two variables is strongly influenced by other variables.
    - Before conclusions are drawn based on correlation and regression, other possible effects of other variables should be considered.

Interpreting Correlation and Regression

- A Few Cautions When Using Correlation and Regression
  - A strong association between two variables is not enough to draw conclusions about cause and effect.
    - Sometimes an observed association really does reflect cause and effect (such as drinking beer causes increased BAC).
    - Sometimes a strong association is explained by other variables that influence both ***x*** and ***y****.*
    - *Remember, association does not imply causation.*