We will look at how two variables are related to each other. We will focus on linear relationships.

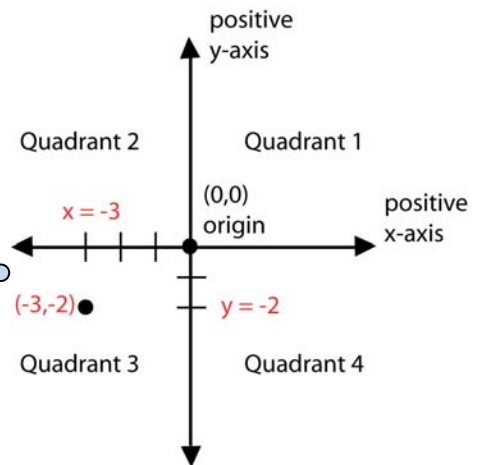General Education Statistics
Class Notes
Scatter Diagrams and Correlation (Section 4.1)

We will explore the relationship, if it exists, between two variables. Imagine how income and years of college education are related. What about the number of televisions in a household and the life expectancy of the people in that household? We will plot scatter diagrams to explore how these and other variables are related.

First, let's review the Cartesian plane.

Cartesian Plane: The Cartesian plane (or simply the xy-plane) is shown below. Familiarize yourself with its parts. Remember a point's coordinates are alphabetical, x then y or (x, y).

The x **coordinate** tells you how far left or right from center the point is. The y **coordinate** tells you how far up or down from center the point is.

positive
y-axis

Quadrant 2   Quadrant 1

$x = -3$ (0,0)   positive
                  origin   x-axis

(-3,-2)● $y = -2$

Quadrant 3   Quadrant 4

expl 1: Draw a quick xy-plane with five, evenly-spaced tick marks in each direction. Then plot the points given below.

(4, -5)
(0, 4)
(-2, -3)
(5, 2)

Optional Worksheet: Relationship between x and y:
This worksheet will review what an equation in x and y means. We explore its graph and x- and y-intercepts.

1

Definition: Linear relationship: A **linear relationship** is a relationship between two variables, often denoted by x and y, where the graph is a **straight line**.

The most commonly used equation that describes a linear relationship is $y = mx + b$. Here m is the slope of the line, b is the y-intercept, and (x, y) is a generic point on the line.

Can you make up an example?
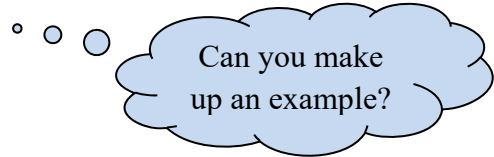
We will collect data in the form of ordered pairs. In other words, we will find the (quantitative) values of two characteristics for many individuals. For instance, we might ask many adults for their income and years of college education. We then make a scatter diagram of these points and look for a consistent trend among the points. This is the idea of **regression**.

To do this, we would first need to determine which is the response variable and which is the explanatory variable. Do you remember what those are?

Definition: The **response variable** is the variable whose value can be explained by the value of the **explanatory** or **predictor variable**.

expl 2: For the variables "current income" and "years of college education", which do you suppose is the response and which is the explanatory variable?
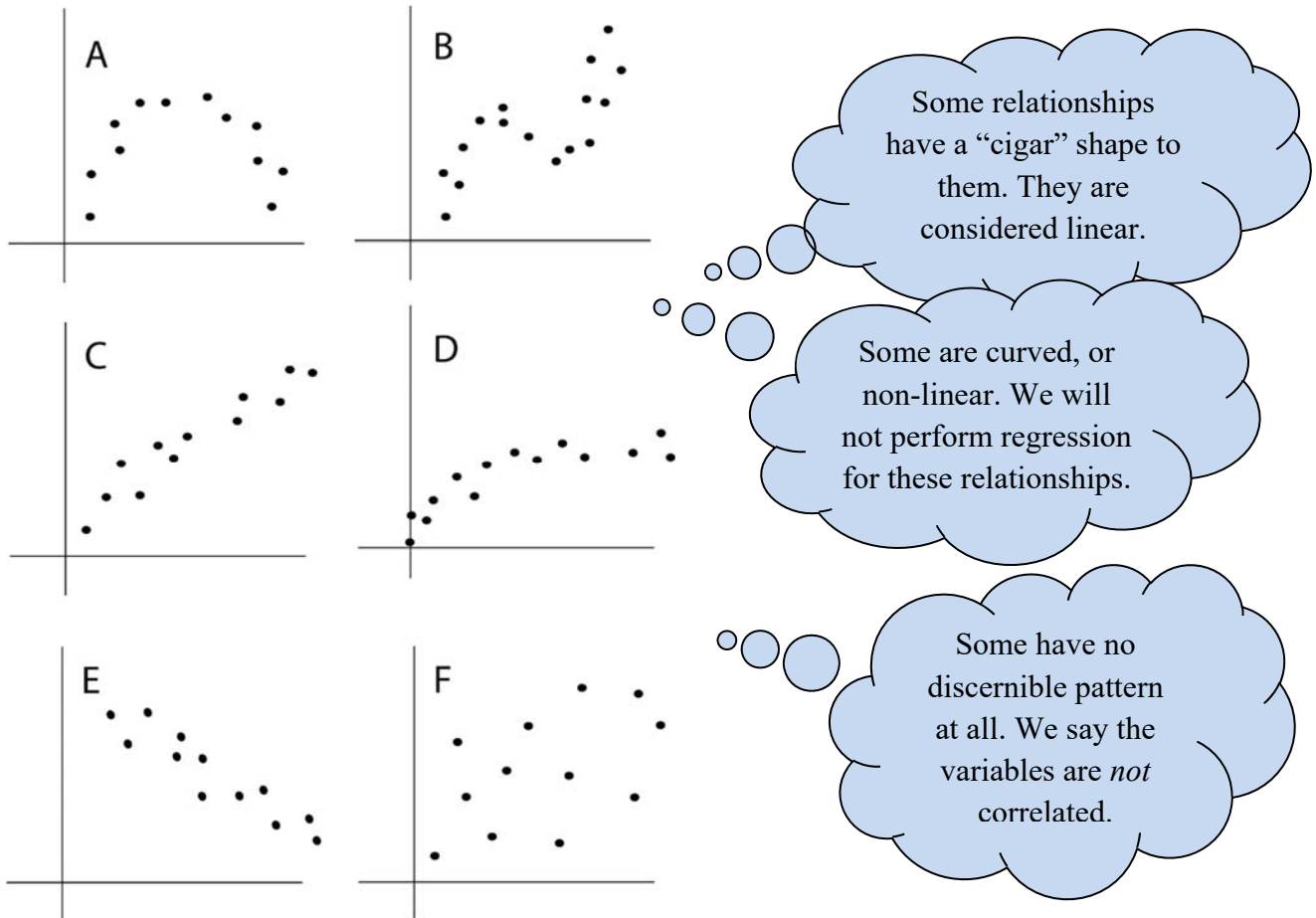
Definition: A **scatter diagram** (or **scatter plot**) is a graph that shows the relationship between two quantitative variables measured on the same individual. Each individual in the data set is represented by a point in the scatter diagram. The **explanatory variable** is plotted on the horizontal axis, and the **response variable** is plotted on the vertical axis. The points are *not* to be connected like you would in a line (or time-series) graph.

expl 3: Consider the variables "current income" and "years of college education". What do you suppose this relationship looks like?

Make up some data points and plot them to the right. Would you expect income to go up as years of education went up, or the opposite? Would you expect a few outliers? What does that mean?

Income

Years of college education
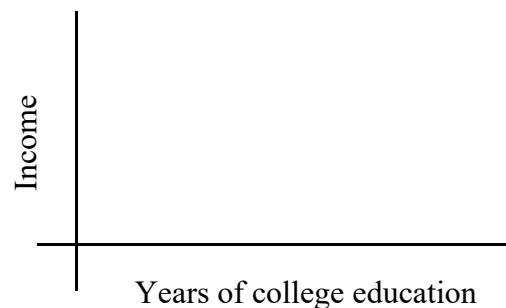
expl 4: Consider the following scatter plots. Which do you think show a linear relationship? On each graph, draw in the line or curve that mimics the pattern of points.

Some relationships have a "cigar" shape to them. They are considered linear.

Some are curved, or non-linear. We will not perform regression for these relationships.

Some have no discernible pattern at all. We say the variables are *not* correlated.

expl 5: What is the main difference between the graphs in parts c and e above? What would that imply about how the variables are related?

Definition: Two variables are **positively associated** if, whenever the value of one variable increases, the value of the other variable also increases.

Definition: Two variables are **negatively associated** if, whenever the value of one variable increases, the value of the other variable *decreases*.

It is good to look at a scatter plot when we start our investigation. However, it is *not* enough to say the graph *looks* linear. We use a more stringent criterion to determine if the two variables are linearly related.

Definition: Sample Linear Correlation Coefficient: We will calculate the following to help determine if the variables are linearly related. Below we see this correlation coefficient for a sample, but you can also use population parameters (μ and σ) if the data is from a population (but we call it ρ (rho) and you would divide by N). We will *not* be doing this.

$$r = \frac{\sum \left(\dfrac{x_i - \bar{x}}{s_x} \right)\left(\dfrac{y_i - \bar{y}}{s_y} \right)}{n - 1}$$

where $\bar{x}$ is the sample mean of the explanatory variable
s_x is the sample standard deviation of the explanatory variable
$\bar{y}$ is the sample mean of the response variable
s_y is the sample standard deviation of the response variable
n is the number of individuals in the sample

Once we calculate the value of r, we will compare it to a table of **Critical Values for Correlation Coefficient** on page A-2 (page 570) in our book.

Properties of the Linear Correlation Coefficient:

1. The linear correlation coefficient is always between -1 and 1, inclusive.
   That is, $-1 \le r \le 1$.

2. If $r = +1$, then a perfect positive linear relation exists between the two variables.
   If $r = -1$, then a perfect negative linear relation exists between the two variables.

3. The closer r is to $+1$, the stronger the evidence is of a *positive* association between the two variables.
   The closer r is to -1, the stronger the evidence is of a *negative* association between the two variables.

4. If r is close to 0, then little or no evidence exists of a *linear* relation between the two variables. This does *not* imply no relation, just no *linear* relation.

5. The linear correlation coefficient is a **unitless** measure of association. So the units of measure for x and y play no role in the interpretation of r.

6. The correlation coefficient is *not* resistant. An observation that does *not* follow the overall pattern of the data (an outlier) could affect the value of the linear correlation coefficient.
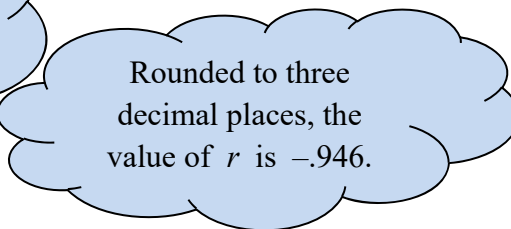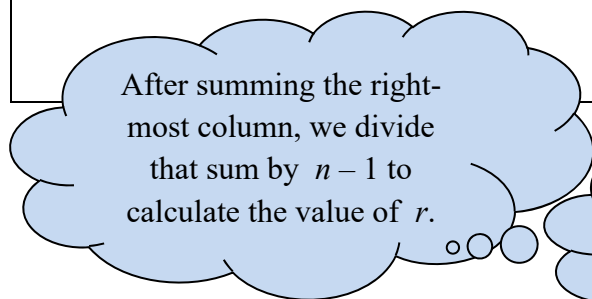
4

expl 6: Consider the following data. We will find the correlation coefficient. Much of the table is filled in for you. After this example, we will *not* do any of this by hand.

x	y
1	18
3	13
3	9
6	6
7	4

Draw a quick *xy*-plane and plot these five points. Does the pattern look linear?
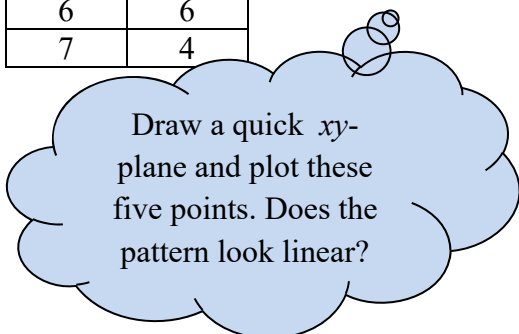
This table shows the many steps of the calculation. You will *not* be asked to do this by hand.

x	y	$\dfrac{x_i - \bar{x}}{s_x}$	$\dfrac{y_i - \bar{y}}{s_y}$	$\left(\dfrac{x_i - \bar{x}}{s_x}\right)\left(\dfrac{y_i - \bar{y}}{s_y}\right)$
1	18	-1.2247	1.4254	-1.7457
3	13	-0.4083	0.5345	-0.2182
3	9	-0.4083	-0.1782	0.0727
6	6	0.8165	-0.7127	-0.5819
7	4	1.2247	-1.0690	-1.3093

$$\sum \left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = -3.7824$$

After summing the right-most column, we divide that sum by $n-1$ to calculate the value of r.

Rounded to three decimal places, the value of r is $-.946$.

Considering the properties given on the previous page, do you think the relationship between x and y is a strong one? Why?

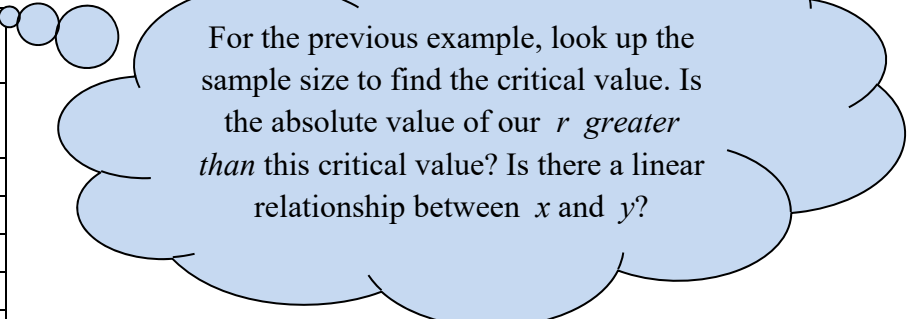**Testing for a Linear Relationship:**

While we can informally look at this value of r and make a judgement as to whether we *believe* it is close to 1 or –1 (and therefore shows a strong relationship), statisticians actually use a non-subjective measure.

We will find the absolute value of r. We compare this to the critical value given in the table on page A-2 in our book, using the sample size to look up the critical value. If the absolute value of r is *greater than* the critical value in the table, then we say there is a linear relationship between the two variables.

Otherwise, *no* linear relationship exists. Be careful here! This does *not* imply there is no relationship at all. There could be a *non-linear* relationship, such as the curved ones shown on page 3.

The table is reproduced in part here.

Critical Values for Correlation Coefficient	
Sample size, n	Critical Value
3	0.997
4	0.950
5	0.878
10	0.632
15	0.514
30	0.361

For the previous example, look up the sample size to find the critical value. Is the absolute value of our r *greater than* this critical value? Is there a linear relationship between x and y?

Instructions: Making Scatter Diagrams and Finding Correlation in StatCrunch:
If needed, enter the explanatory variable (x) in column "var1" and the response variable (y) in column "var2". Label the columns appropriately. To make a scatter diagram, select **Graph** > **Scatter Plot**. You'll need to tell it which column is the "X variable" and which is the "Y variable". After scrolling down, enter labels for the axes and a title. Click **Compute!** and it will make the plot.

Next, we'll calculate the correlation coefficient. With data in place, select **Stat** > **Summary Stats** > **Correlation**. Under "Select column(s)", Shift-select both "var1" and "var2" (or whatever you called the columns). Click **Compute!** and it will pop-up another window with the value of r.

Worksheet: Regression and Correlation:

Here we practice looking for linear relationships and explore the values of the correlation coefficient, *r.*

Correlation Versus Causation:

If two variables are correlated (meaning there is a relationship), can we say that one variable *actually causes* the change in the other? Let's consider a classic example.

expl 7: Consider the following graph that shows a sample of countries, detailing the number of televisions in a country and the life expectancy of the people.
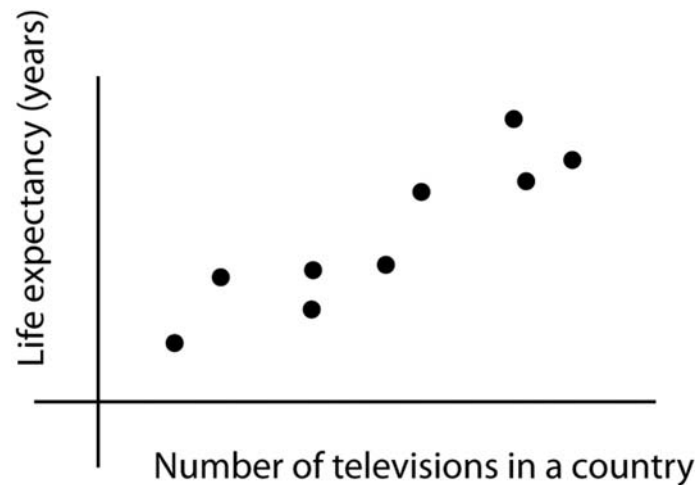
a.) Would you say there is a linear relationship between the two variables?

b.) What would you estimate the value of the correlation coefficient to be?

c.) Now, can we say that the number of televisions in a country *causes* the life expectancy to rise? Should we send our old TV's to poor countries to raise their life expectancies?

d.) Can you think of lurking variables that may have influenced our variables and really be responsible for the increase in life expectancy?

To claim causality, we have to perform a designed experiment. Observational data is *not* enough.