Section 9.1

In this handout, we examine bivariate data. For bivariate data, each data value carries information about two variables.

Definition: A **scatter diagram** is a graph in which data pairs (x, y) are plotted as individual points on a grid with horizontal axis x and vertical axis y. We call x the **explanatory variable** and y the **response variable**.

Exercise 1. Suppose you are interested in seeing the relationship between the average tip a person gives and the worth of his/her car. We find this information for seven people:

Average Tip (%)	20.6	14.2	18.9	26.7	23.5	17.2	24.9
Worth of Car (in thousands of dollars)	19.2	13.5	17.8	26.2	22.0	16.7	24.5

(a) Display the data in a scatter diagram.

(b) From the scatter plot, does there appear to a relationship between the average tip a person gives and the worth of his/her car? If so, what is it?

it seems like

Definition: Two variables are **positively associated** when above average values of one tend to accompany above-average values of the other and below-average values also tend to occur together.

Definition: Two variables are **negatively associated** when above average values of one accompany below-average of the other, and vice versa.

(c) Are the two variables positively associated or negatively associated or is there no association?

The two variables are

(d) Does this mean that if you are more generous when it comes to tipping, your car will increase in value or vice versa?

increasing your tips your car worth will increase!

Important Point: Association does not imply causation!

Definition: A **lurking variable** is a variable that is neither an explanatory nor a response variable. Yet, a lurking variable may be responsible for changes in both x and y.

(e) What is the lurking variable in this problem?

Exercise 2. We suspect that the number of hours a student spends online impacts his/her test score. Here is the number of hours 9 students spent online during the weekend and the scores of each student who took a test the following Monday:

Hours spent online, x	0	1	2	3	5	5	6	7	10
Test score, y	96	85	82	74	68	76	58	65	50

(a) Display the data in a scatter plot.

(b) From the scatter plot, does there appear to be a relationship between the hours spent online and the test score? If so, what is it?

it appears that hours spent online corresponds to test scores.

Definition: A **response variable** measures an outcome of a study.
Definition: An **explanatory variable** explains or causes changes in the response variables.

(c) What is the response variable in this problem?

The response variable is

(d) What is the explanatory variable in this problem?

The explanatory variable is

(e) Are the two variables positively associated or negatively associated or is there no association?

The two variables are

(f) Does there appear to be a linear relationship between the two variables?

From the scatter plot, there between the two variables.

You can draw a on the scatter plot which would be to all the data points.

Definition: The **sample correlation coefficient r** is a numerical measurement that assesses the strength of a *linear* relationship between two variables x and y. Here is the computational formula for r:

$$r = \frac{n\Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n\Sigma x^2 - (\Sigma x)^2}\sqrt{n\Sigma y^2 - (\Sigma y)^2}}$$

1. r is a unitless measurement between -1 and 1. In symbols, -1 $\leq r \leq$ 1. If $r = 1$, there is a perfect positive linear correlation. If r = -1, there is perfect negative linear correlation. If $r = 0$, there is no linear correlation. The closer r is to 1 or -1, the better a line describes the relationship between two variables x and y.
2. Positive values of r imply that as x increases, y tends to increase. Negative values of r imply that as x increases, y tends to decrease.
3. The value of r is the same regardless of which variable is the explanatory variable and which

is the response variable. In other words, the value of r is the same for the pairs (x, y) and the corresponding pairs (y, x).
4. The value of r does not change when either variable is converted to different units.

(g) Calculate r.

(h) What does the r-value in (g) indicate about the relationship between hours spent online and test score?

Since -0.949 , the value in (g) indicates a
relationship between hours spent online and test score.

Remark: Keep in mind that this is a sample and we can't make any **definite** conclusions about the population based on this sample. The sample correlation coefficient is -0.949, but that does not mean that the correlation coefficient for the entire population is -0.949.

(i) Assuming that the sample data is very representative of the population, does this mean that if you spend fewer hours online, your test score will increase?

the correlation coefficient measures the strength of the linear relationship between the two variables, It's possible that students who spend a lot of hours online are

Remark: Again, association doesn't imply causation!

Exercise 3. A group of 9 children are given donuts and asked to shoot 20 free throws. Here are the number of donuts eaten and the number of free throws made for each child:

Donuts, x	1	2	3	4	5	6	7	8	9
Free throws made, y	0	1	4	8	16	10	4	2	0

(a) Display the data in a scatter plot.

(b) Compute r.

(c) What does the r-value indicate about the relationship?

Since r is , this means that there is a relationship between donuts eaten and free throws made.

(d) The scatter plot seems to indicate some relationship between donuts eaten and free throws made. Does this contradict part (c)? Why or why not?

This does r measures the strength of a
The scatter plot seems to indicate a

Remark: Keep in mind that r measures the strength of a **linear** relationship. Just because the relationship is not linear doesn't mean that there is no relationship.

Class Exercise 1. Here is the ages (in years) of 10 men and their systolic blood pressure.

Age, x	16	25	39	45	49	64	70	29	57	22
Systolic blood pressure, y	109	122	143	132	199	185	199	130	175	118

(a) Make a scatter plot of the data.
(b) Calculate r using the formula.
(c) What does your answer to part (b) indicate about the relationship between age and systolic blood pressure?

Remark: There is an easier way to find r using the calculator.

Class Exercise 2. Calculate r for the following sets of points:
(a) {(15.6,5.2), (26.8, 6.1), (37.8,8.7), (36.4,8.5), (35.5,8.8), (18.6,4.9), (15.3,4.5), (7.9,2.5), (0.0,1.1)} **Answer:** r = **0.989**

(b) {(4.0,3.3), (4.7,8.3), (6.3,4.5), (8.2,9.3), (12.0,10.7), (15.9,16.4), (17.4,15.4), (18.1,17.6), (20.2,21.0), (23.9,21.7)} **Answer:** r = **0.967**

(c) {(94, 0.473), (96, 0.753), (95, 0.929), (95,0.939), (94,0.832), (95,0.983), (94,1.049), (104,1.178), (104,1.176), (106,1.292), (108, 1.403), (110, 1.499), (113,1.529), (118,1.749), (115,1.746), (121,1.897), (127,2.040), (131,2.231)} **Answer:** r = **0.968**

Class Exercise 3. The following data are based on information from *Domestic Affairs*. Let x be the average number of employees in a group health insurance plan, and let y be the average administrative cost as a percentage of claims. (#14)

x	3	7	15	35	75
y	40	35	30	25	18

(a) Make a scatter diagram and draw the line you think best fits the data.
(b) Would you say the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma = 135$, $\Sigma\ x^2 = 7133$, $\Sigma\ y = 148$, $\Sigma\ y^2 = 4674$, and $\Sigma\ xy = 3040$. Compute r. As x increases from 3 to 75, does the value of r imply that y should tend to increase or decrease? Explain. **Answer:** $r \approx$ **-0.945, decrease**

Class Exercise 4. Is the magnitude of an earthquake related to the depth below the surface at which the quake occurs? Let x be the magnitude of an earthquake (on the Richter scale), and let y be the depth (in kilometers) of the quake below the surface at the epicenter. The following is based on information taken from the National Earthquake Information Service of the U.S. Geological Survey. Additional data may be found by visiting the web site for the service. (#16)

x	2.9	4.2	3.3	4.5	2.6	3.2	3.4
y	5.0	10.0	11.2	10.0	7.9	3.9	5.5

(a) Make a scatter diagram and draw the line that you think best fits the data.
(b) Would you say that the correlation is low, moderate, or strong? positive or negative?
(c) Use a calculator to verify that $\Sigma\, x = 24.1$, $\Sigma\, x^2 = 85.75$, $\Sigma\, y = 53.5$, $\Sigma\, y^2 = 458.31$, and $\Sigma\, xy = 190.18$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain. **Answer: $r \approx$ 0.511, increase**

Class Exercise 5. Do larger universities tend to have more property crime? University crime statistics are affected by a variety of factors. The surrounding community, accessibility given to outside visitors, and many other factors influences crime rates. Let x be a variable that represents student enrollment (in thousands) on a university campus, and let y be a variable that represents the number of burglaries in a year on the university campus. A random sample of $n = 8$ universities in California gave the following information about enrollments and annual burglary incidents (Reference: *Crime in the United States*, Federal Bureau of Investigation). (#18)

x	2.9	4.2	3.3	4.5	2.6	3.2	3.4
y	5.0	10.0	11.2	10.0	7.9	3.9	5.5

(a) Make a scatter diagram and draw the line that you think best fits the data.
(b) Would you say that the correlation is low, moderate, or high? positive or negative?
(c) Using a calculator, verify that $\Sigma x = 152.8$, $\Sigma x^2 = 3350.98$, $\Sigma y = 246$, $\Sigma y^2 = 10030$, and $\Sigma xy = 5488.4$. Compute r. As x increases, does the value of r imply that y should tend to increase or decrease? Explain. **Answer: $r \approx$ 0.765, increase**

Homework

C Problems

Section 9.1: 13-17 ODD

B Problems

Section 9.1: 1, 3

A Problems

Section 9.1: 5-11 ODD