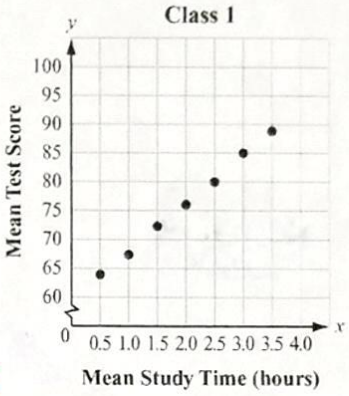
NOTES: Scatter Plots and Correlation

A **scatter plot** is often used to present bivariate **quantitative** data. Each variable is represented on an axis and the axes are labeled accordingly.

A scatter plot displays data as points on a grid using the associated numbers as coordinates or ordered pairs (x, y). The way the points are arranged by themselves in a scatter plot may or may not suggest a relationship between the two variables. For instance, by reading the below, do you think there is a relationship between the hours spent studying and exam grades?



graph

If y tends to increase as x increases, then the data have **positive** correlation. ↑ ↑ BOTH GO THE SAME DIRECTION

If y tends to decrease as x increases, then the data have **negative** correlation. ↑ ↓ THE VARIABLES GO OPPOSITE DIRECTIONS

Correlation Coefficient

Direction	
Positive Correlation	Negative Correlation
	

Strength		
Weak Correlation	No Correlation	~~Negative~~ STRONG Correlation
		

In order to determine the direction and strength of the model, we use something called a **correlation coefficient**. It is represented by the letter ***r***.

Direction and correlation coefficients

- r is always between -1 and 1
- A positive r is a positive correlation, a negative r is a negative correlation

Strength and correlation coefficients

- A r of 1 is a perfect positive correlation (an r of -1 is a perfect negative correlation)
- The closer the value to 1 (or -1) the stronger the correlation
- Generally 0.8 – 1 represent strong correlations
- 0.7-0.79 represent weak correlation
- Anything below 0.7 generally has no correlation

r = 0.93	r = -0.95	r = -0.75	r = 0. 61	r = 1
strong positive	strong negative	weak negative	no correlation	perfect positive

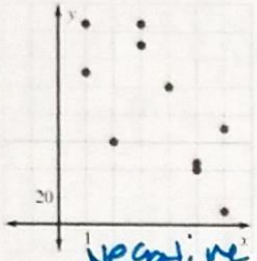
Practice Problems:

For each scatter plot, tell whether the data have a

a) positive correlation, a negative correlation, or no correlation.

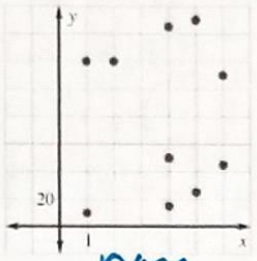
b) If the correlation is positive or negative, determine if it is a strong or weak correlation

1.



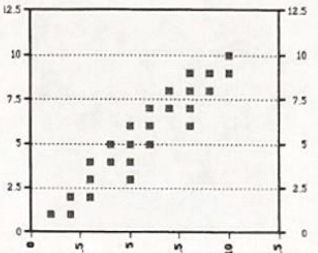
negative weak

2.



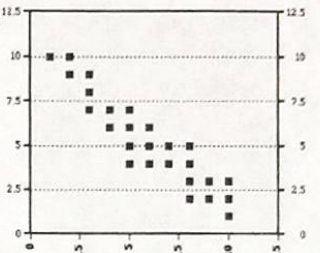
none

3



strong positive

4



strong negative

5. Positive, negative, or no correlation?

a. Amount of exercise and percent of body fat ↑↓ negative

b. A person's age and the number of medical conditions they have ↑↑ positive

c. Temperature and number of ice cream cones sold ↑↑ positive

d. The number of students at Sprayberry and the number of dogs in Atlanta ↑? no correlation

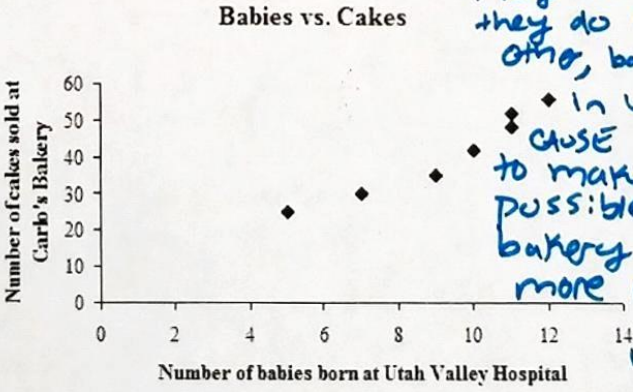
e. Age of a tadpole and the length of its tail ↑↓ negative

Correlation vs. Causation

When a scatter plot shows a correlation between two variables, even if it's a strong one, there is *not necessarily* a *cause-and-effect relationship*. Both variables could be related to some third variable that actually causes the apparent correlation. Also, an apparent correlation simply could be the result of chance.

Example 1: During the month of June the number of new babies born at the Utah Valley Hospital was recorded for a week. Over the same time period, the number of cakes sold at Carlo's Bakery in Hoboken, New Jersey was also recorded. What can be said about the correlation? Is there causation? Why or why not?

Number of babies born	Number of cakes sold
5	25
7	30
9	35
10	42
11	48
11	52
12	56



They are related but they do not cause each other, babies being born in Utah does not cause a NJ bakery to make cakes. It's possible that the bakery is making more cakes because there are more weddings in June

Example 2: An American medical researcher wants to see if there is a link between a person's socio-economic status (how much money they have) and certain types of cancer. His research seems to indicate that there is a link (rich people seem to suffer from more cancers than poor people do). His Causation Statement: *Being rich will make you more likely to get cancer.* What can be said about the correlation? Is there causation, why or why not? BEING RICH DOES NOT CAUSE CANCER. A THIRD FACTOR COULD EXPLAIN THIS. PEOPLE WHO HAVE MORE MONEY ARE MORE LIKELY TO GO TO THE DOCTOR REGULARLY SO THEY HAVE A GREATER CHANCE OF FINDING CANCER AS OPPOSED TO PEOPLE WHO CANNOT GO TO THE DOCTOR AND DON'T KNOW THEY HAVE CANCER UNTIL IT IS TOO LATE.

NOTES: Scatter plots and line of best fit

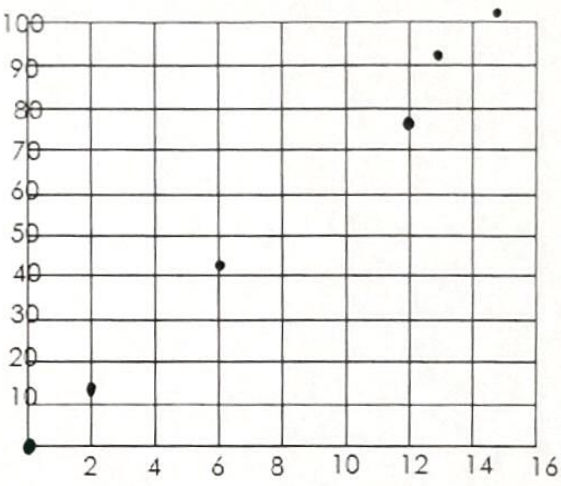
Kim works as waitress. Below is a table of how many tables she works and what she makes in tips

Tables	0	2	6	10	12	13	15
Amount in tips	0	13	42	65	78	92	102

1) Do you think the relationship is a correlation or a causation? Why?
CAUSATION. THE MORE TABLES SHE SERVES HAS A DIRECT RELATIONSHIP ON THE ~~NUMBER~~ AMOUNT OF TIPS SHE GETS

2) Do you think there is a positive or negative correlation coefficient? Why?
positive the more she works the more she will make

3) Graph the function, what kind of graph is this? LINEAR (Scatter plot)



4) Can you predict how much money Kim will make in tips if she served 22 tables?
~~Maybe~~

a is the same as m

Line of best fit: the linear equation $y=ax+b$ that matches a line to a scatter plot as close as possible.

Calculator Steps:

1) Go to [data]. If needed clear out the previous data by pressing [data] and second time and pressing enter on whichever list you want to clear
2) Enter all the x values into L1 and all the y values into L2
3) Go to [stat-reg/distr]. You can do this by pressing [2nd] [data]
4) Select 2-Var stats (because we now have 2 variables)
5) Scroll down to find the variables you need

a	b	r	$\bar{x}, \bar{y}$
m (slope) rate of change	y-intercept (starting value)	correlation coefficient	the mean (of either x or y)

How y changes for every x

5) What is the line of best fit for Kim's waitressing variables?

$y = 6.80x - .32$

6) What is the correlation coefficient?

$r = .998$

7) Based on your correlation coefficient is this a strong or a weak model? Would it make a good predictor? Why or why not?
Strong! yes because the correlation coefficient is very close to 1

8) Using the line of best fit model, how much money would you predict that Kim would make if she served 22 tables?

$y = 6.80(22) - .32$ $\$149.28$

An old myth says 1 dog year equals 7 human years, but that comparison is not accurate. An approximate comparison of ages is shown in the table:

Age of Cat or Dog *(in years)*	3 months (.25 years)	6 months (.5 years)	1	2	4	6	8	10	14	18	20	21
Approximate Equivalent Human age *(in years)*	5	10	15	24	32	40	48	56	72	90	94	101

1. Use your calculator to determine the regression formula

   $y = 4.32x + 11.20$

2. What is the correlation coefficient?

   $r = .995$

3. Using your equation, determine the equivalent age of the animal who is:

   x-variable, plug in for x

   a. Dog: 12 years old

   $y = 4.32(12) + 11.20$

   63.04 years!

   b. Cat: 16 years old

   $y = 4.32(16) + 11.20$

   80.32 years!

4. Using your equation, how old would a cat or dog have to be to have the maturity of a human who is:

   Human years are the y-variable. plug in for y and solve

   a. 65 years old

   $65 = 4.32x + 11.20$

   $-11.20 \quad -11.20$

   $\frac{53.8}{4.32} = \frac{4.32x}{4.32}$

   12.5 years

   b. 78 years old

   $78 = 4.32x + 11.20$

   $-11.20 \quad -11.20$

   $\frac{66.8}{4.32} = \frac{4.32x}{4.32}$

   $x = 15.5$ years

5. The documented world's oldest dog was named Bluey from Australia who died when he was 29 years and 6 months old. Find Bluey's approximate equivalent human age.

   $y = 4.32(29.5) + 11.20$

   $y = 138.64$ years old!