CORRELATION

The association between two variables

Height & Weight of 20 Young Females

Arranged with Height from shortest to tallest

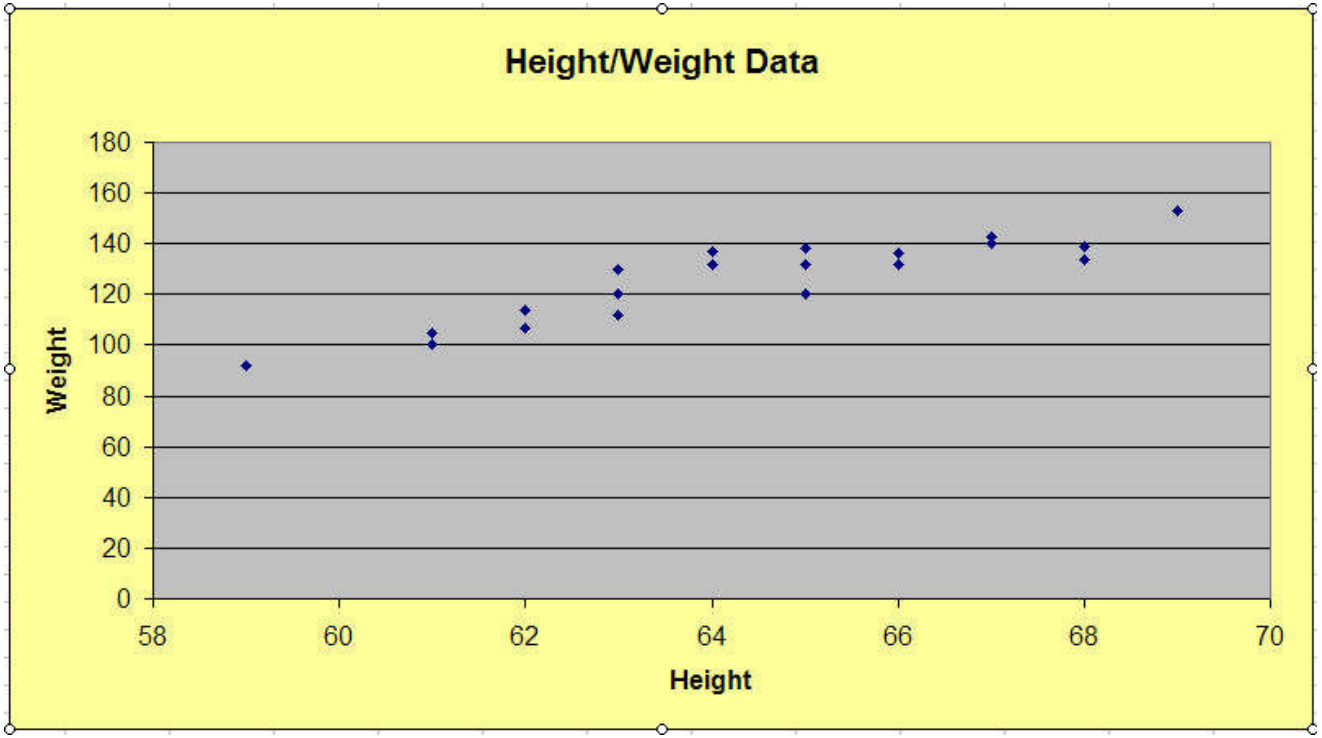
Case	Height (Inches)	Weight (Pounds)
1	59	92
2	61	105
3	61	100
4	62	107
5	62	114
6	63	112
7	63	120
8	63	130
9	64	132
10	64	137
11	65	132
12	65	138
13	65	120
14	66	136
15	66	132
16	67	140
17	67	143
18	68	139
19	68	134
20	69	153

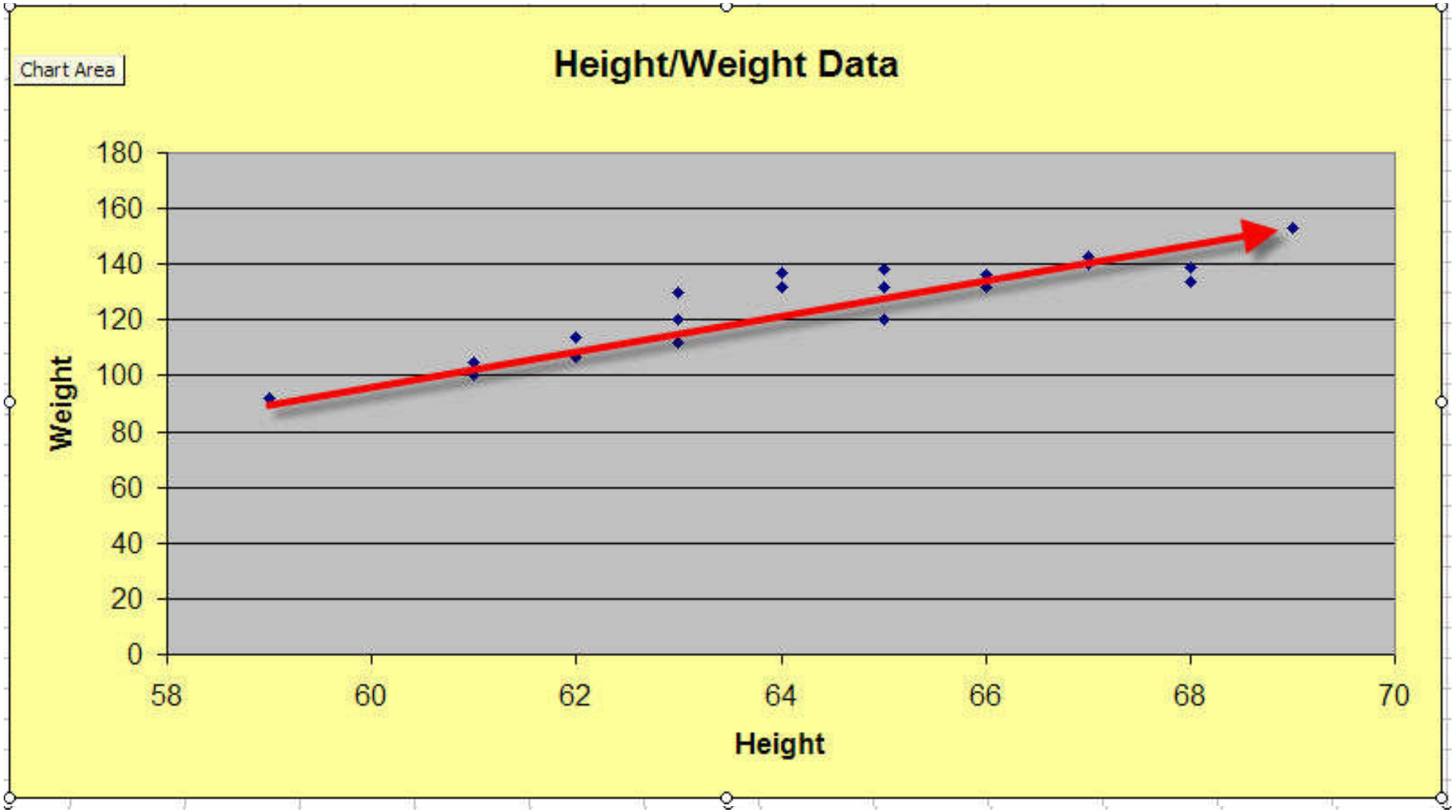
Does a pattern exist?

As the height increases what about the weight?

Scatter Plots give a “visual” representation of the data.

Consists of X & Y axis on the graph





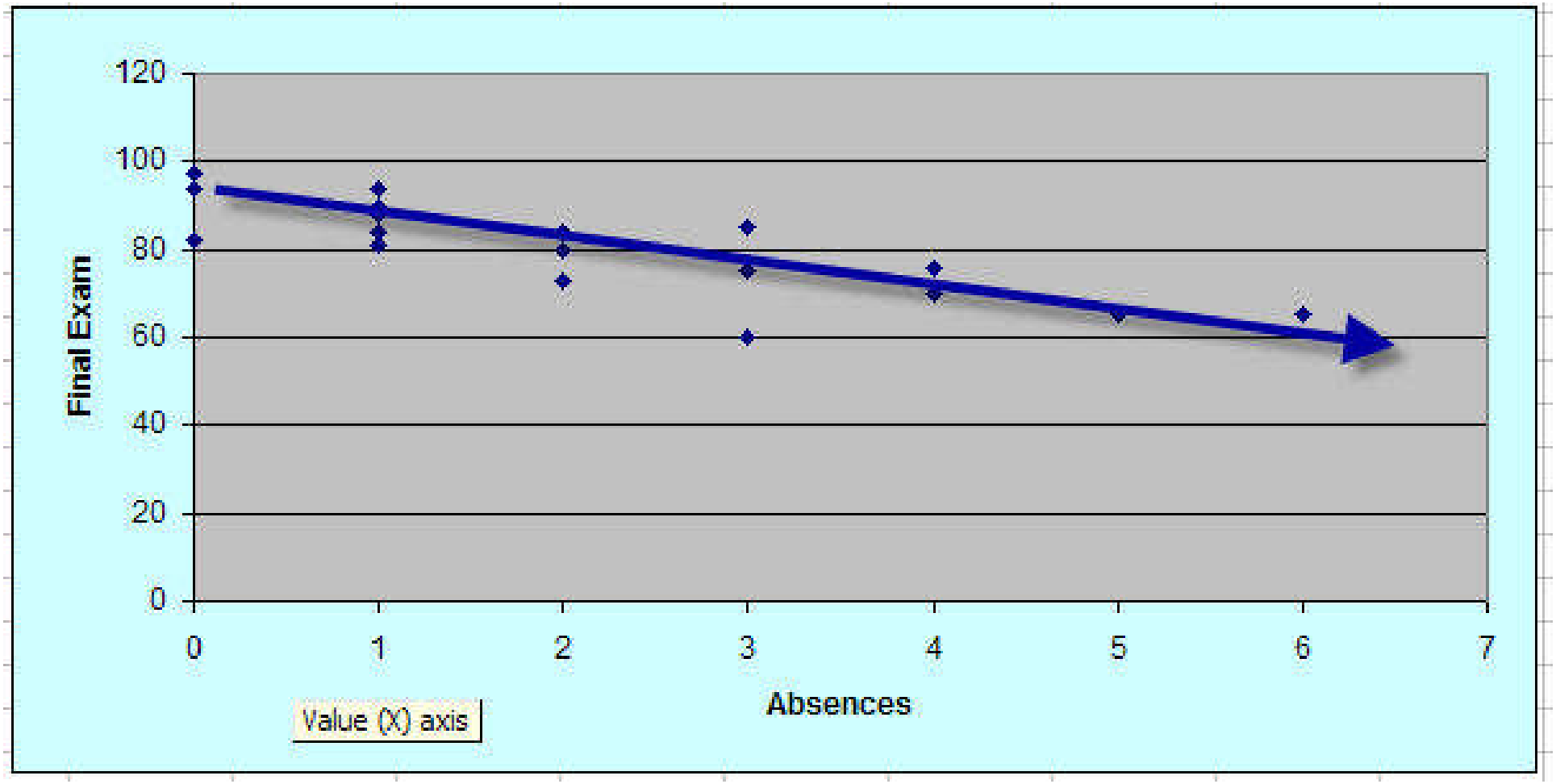
For this data, there seems to be an association between height and weight

(*as height increases, weight increases*)

Is there a relationship between the number of class absences and the final exam score?

Using a scatterplot helps us visualize any relationship.

Absences	Final Exam
0	82
0	94
0	97
1	81
1	84
1	88
1	90
1	94
2	73
2	80
2	84
3	60
3	75
3	85
4	70
4	76
5	65
6	65



Appears that as absences increase, scores decline

Scatter plot questions:

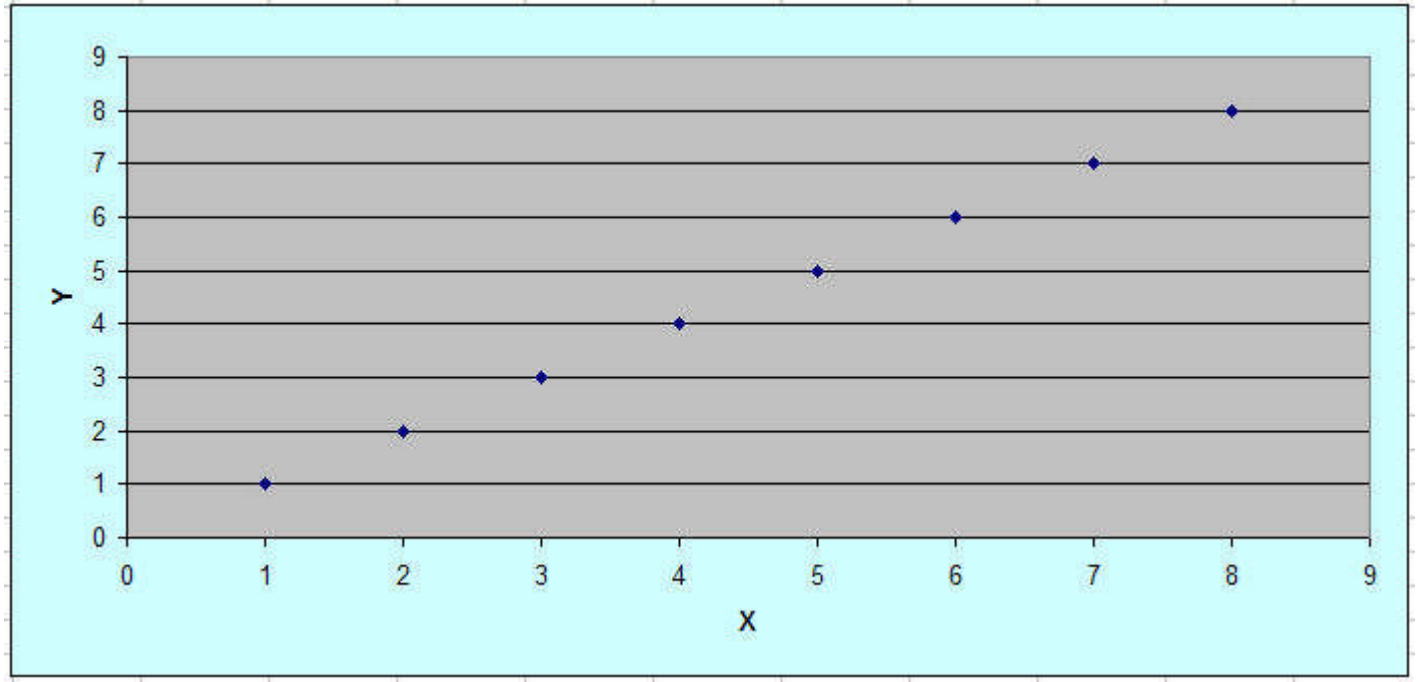
Is there a straight-line pattern or association?

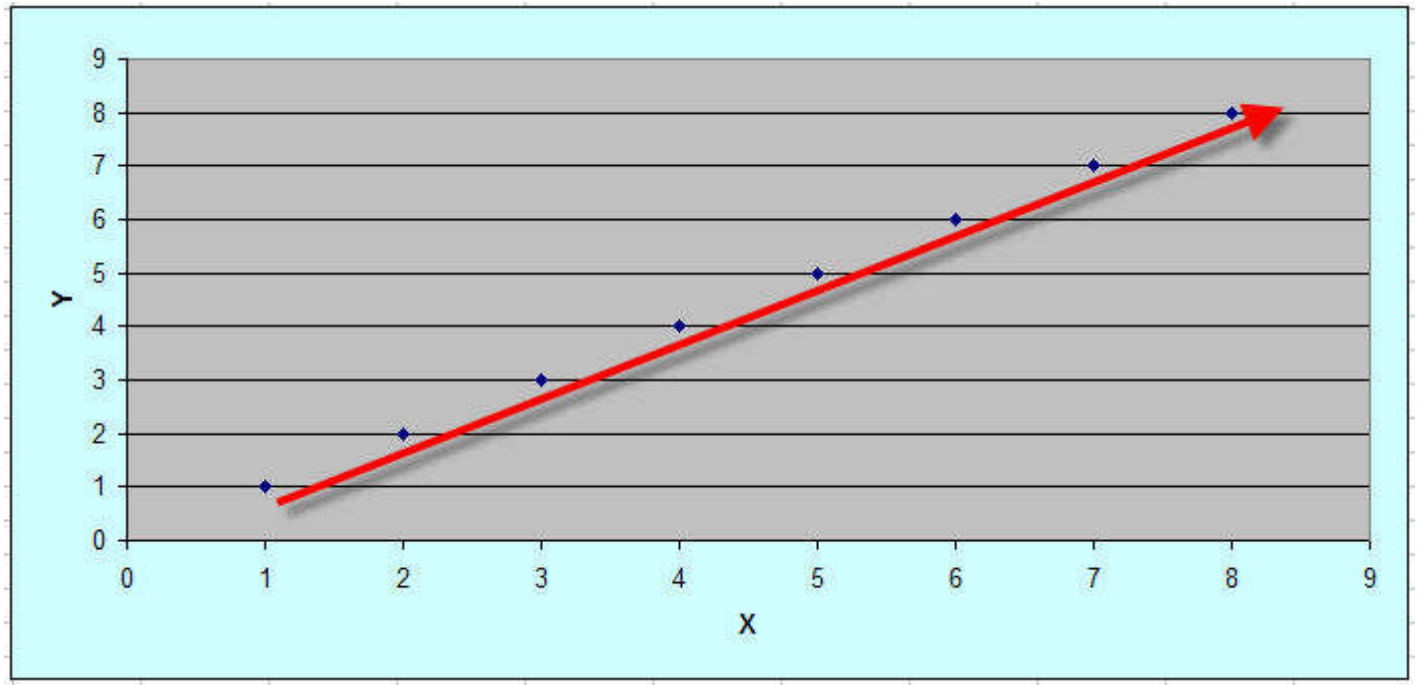
Does the pattern or association slope upward or downward?

Are the plotted values tightly clustered together in a pattern or widely separated?

Are there noticeable deviations from the pattern?

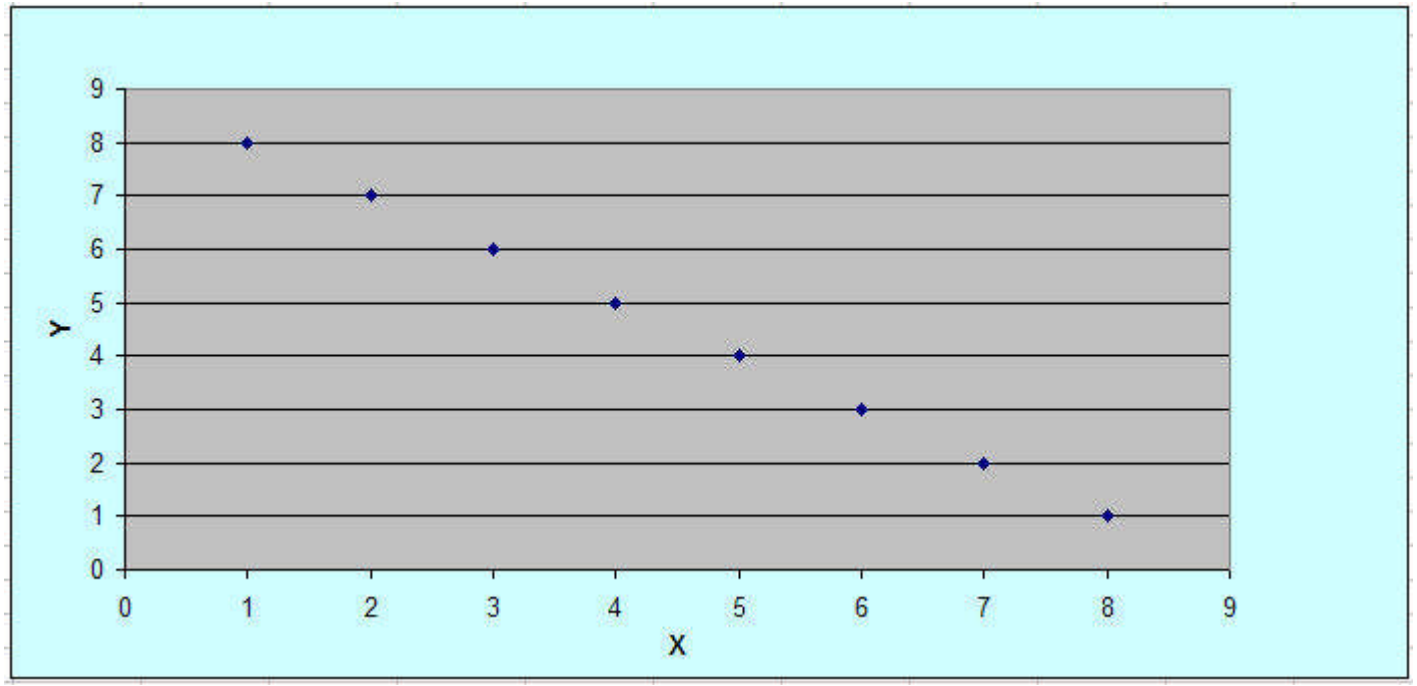
X	Y
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8

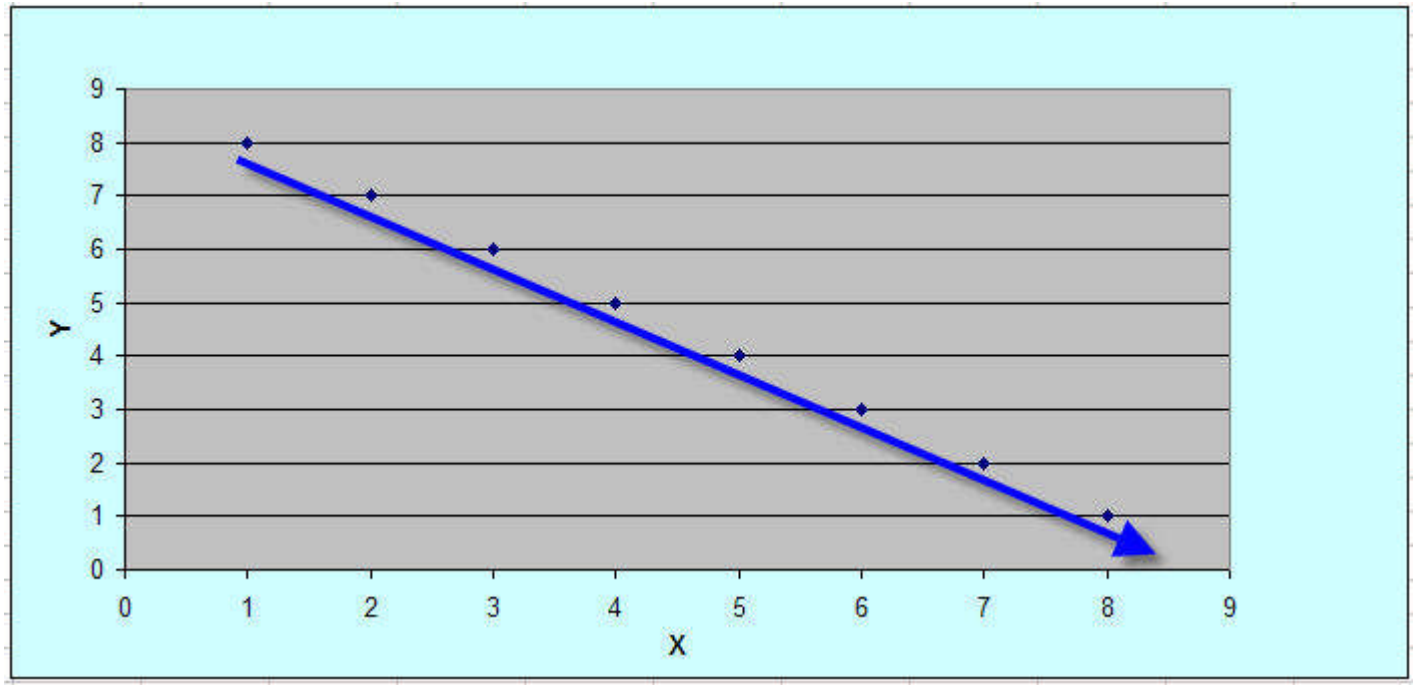




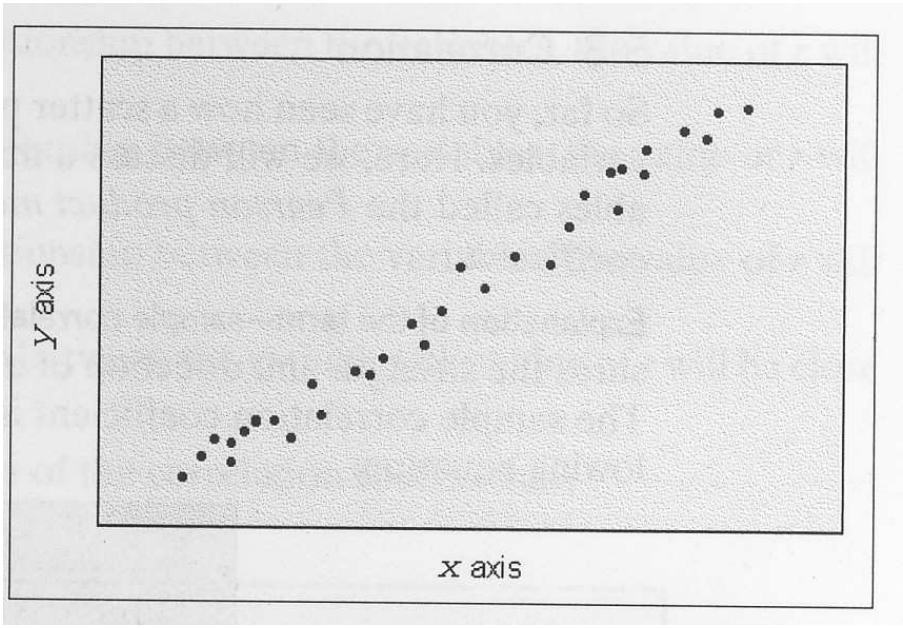
Perfect Positive Association

X	Y
1	8
2	7
3	6
4	5
5	4
6	3
7	2
8	1

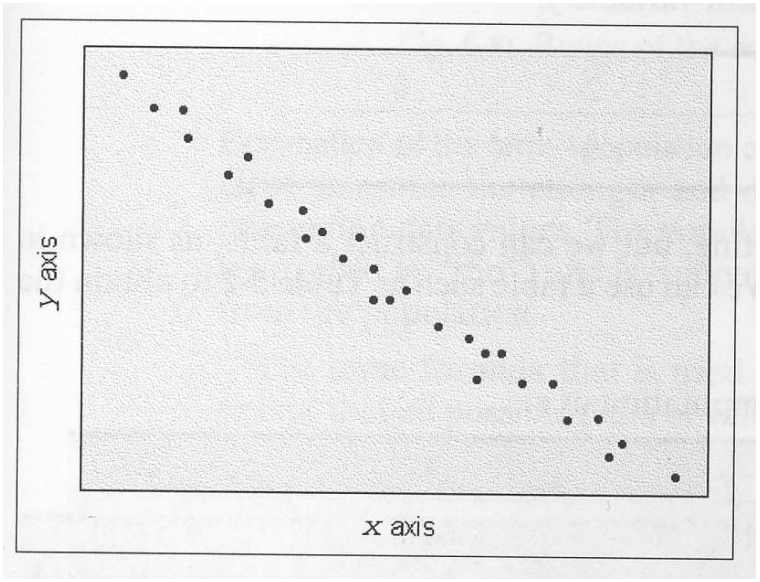




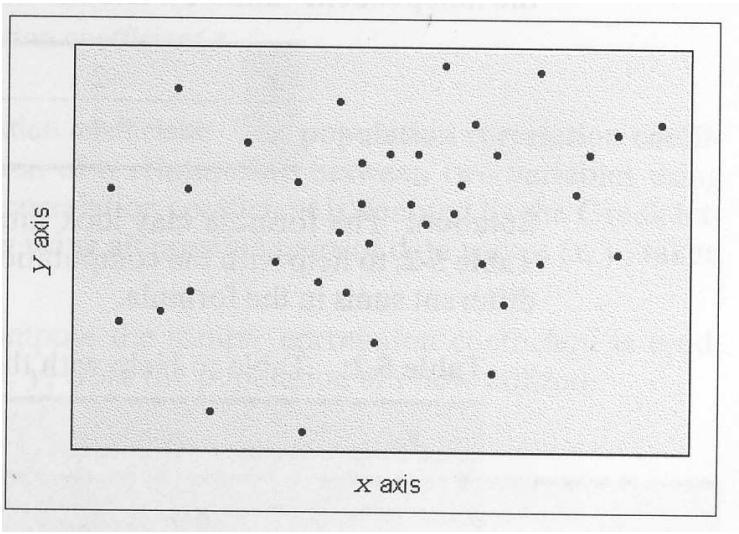
Perfect Negative Association



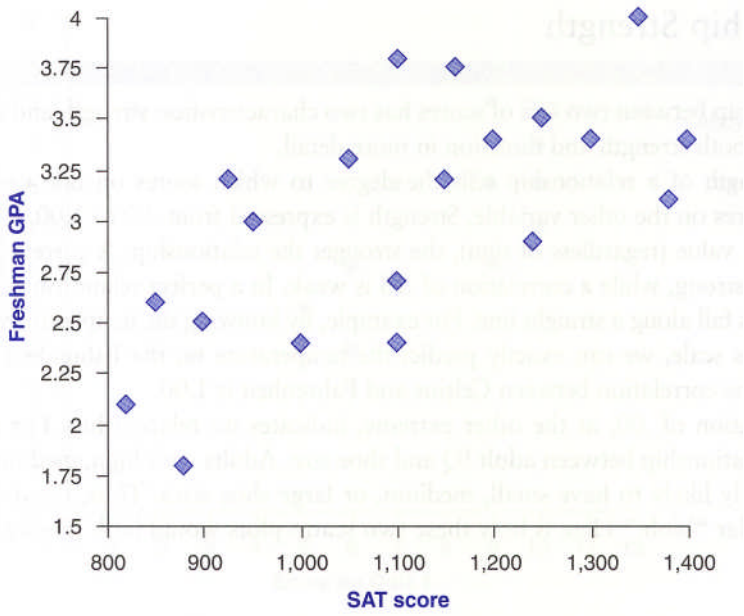
Very Strong Positive Association



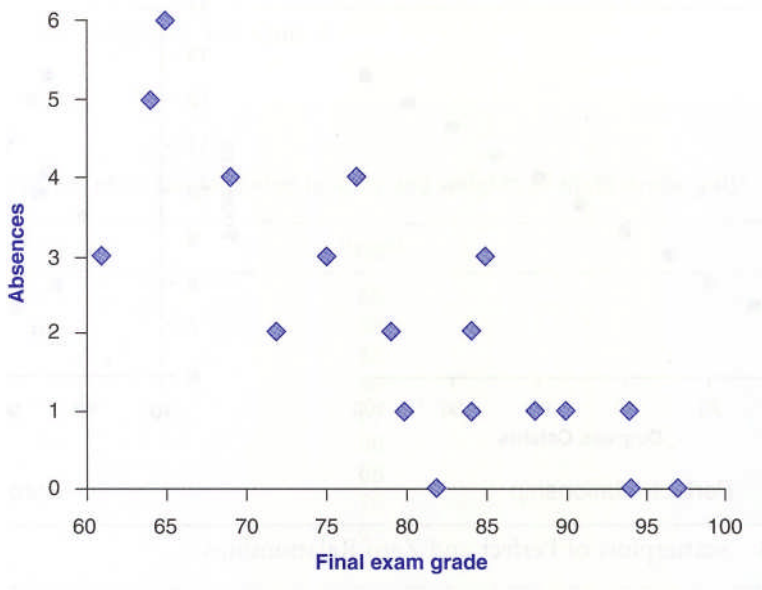
Very Strong Negative Association



No Association

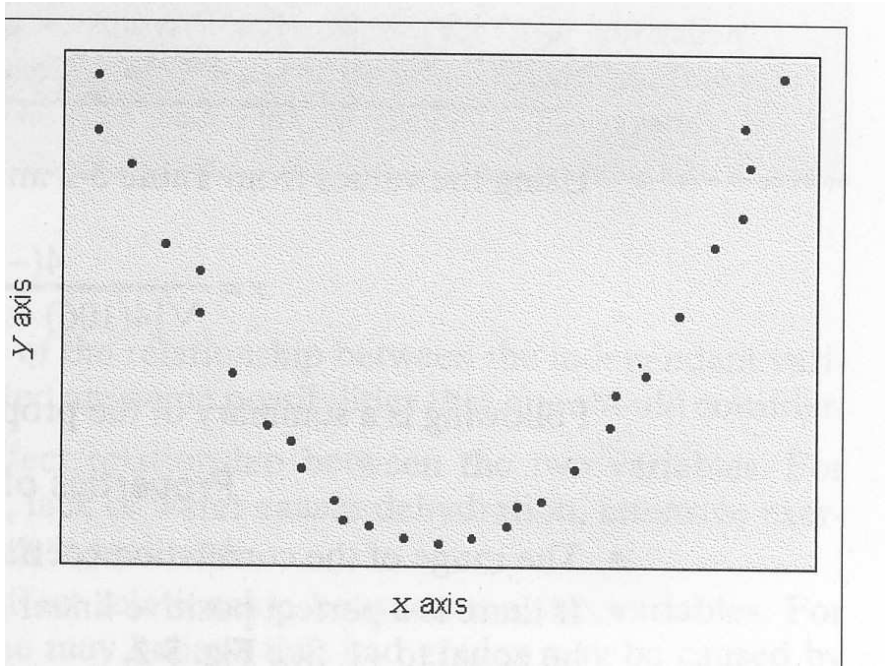


Moderate Association (Positive)



(Negative)

Side note: Non-linear association



Must look at all the data. Hand-plotted graph could be deceiving.

CORRELATION ANALYSIS

technique developed by Karl Pearson, hence

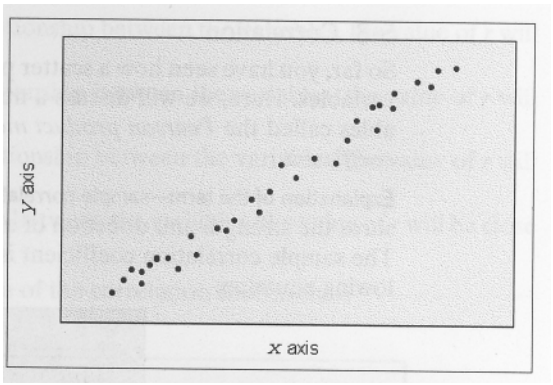
Pearson's r (or Pearson's rho, ρ)

Calculated value ranges from +1.0 to -1.0

Value depends upon

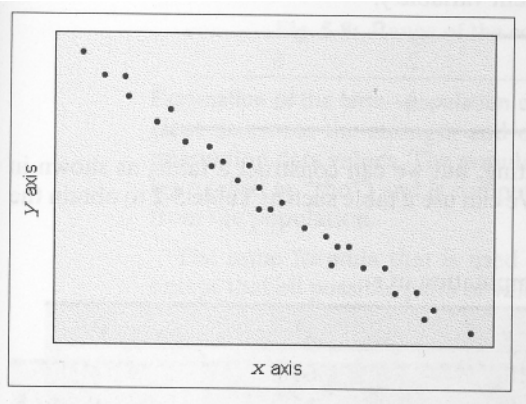
Strength of relationship

Direction

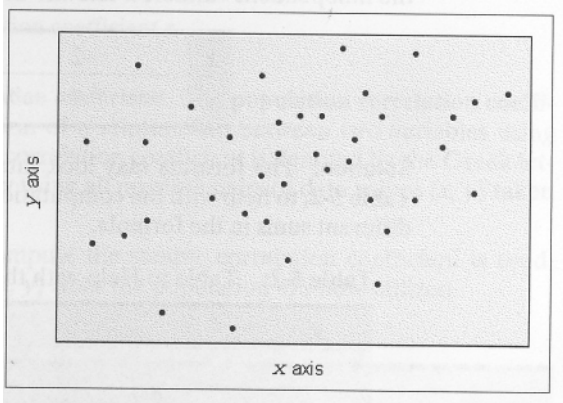


r = +0.94

Very strong, positive relationship



r = -0.94 *indicates very strong negative relationship*



r = 0.00 *indicates no relationship*

Formula for Pearson's r *(Raw score method)*

$$r = [\, n\Sigma XY - (\Sigma X)(\Sigma Y) \,] \div$$

$$\text{Square root of } ([\, n(\Sigma X^2) - (\Sigma X)^2 \,][\, n(\Sigma Y^2) - (\Sigma Y)^2])$$

where ***n*** is the number of cases

Case	X	Y	XY	X^2	Y^2
1	2	4	8	4	16
2	2	3	6	4	9
3	3	3	9	9	9
4	4	3	12	16	9
5	5	6	30	25	36
6	5	4	20	25	16
7	6	5	30	36	25
8	6	6	36	36	36
9	7	6	42	49	36
10	7	5	35	49	25
Σ=	**47**	**45**	**228**	**253**	**217**

$$r = [\, n\Sigma XY - (\Sigma X)(\Sigma Y) \,] \div$$

$$\text{Square root of } ([\, n(\Sigma X^2) - (\Sigma X)^2 \,][\, n(\Sigma Y^2) - (\Sigma Y)^2])$$

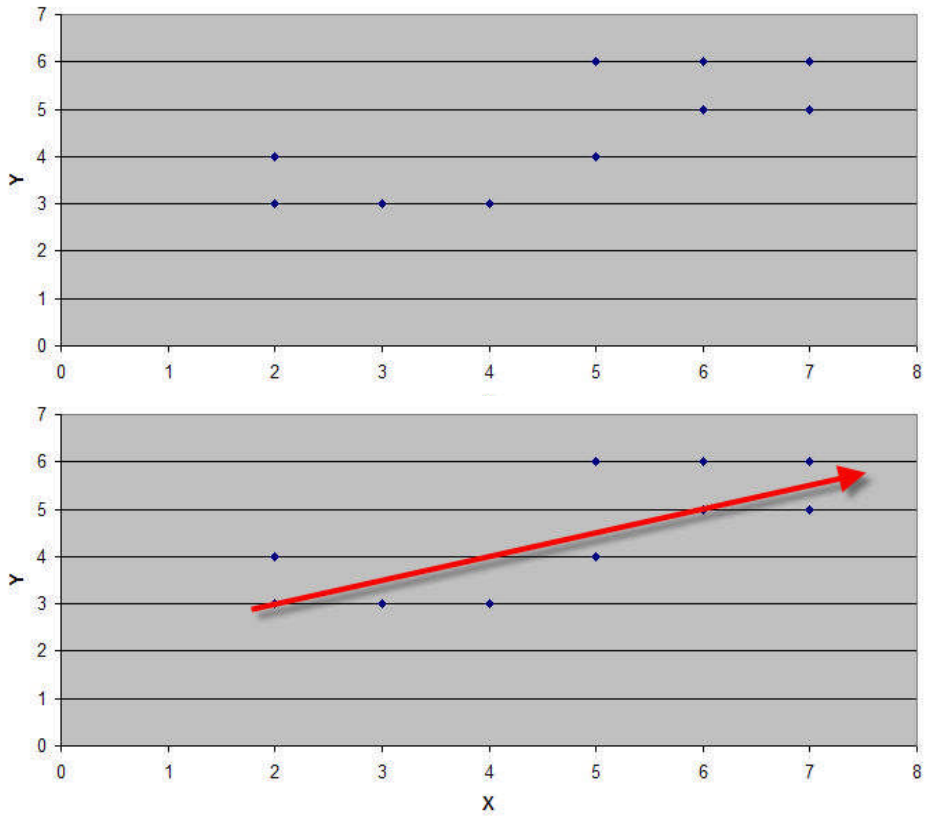
$$r = [10 \times 228 - 47 \times 45] \div$$

$$\text{SqRt}([10 \times 253 - 47^2] \times [10 \times 217 - 45^2]$$

$$r = 165 / 215.7$$

$$r = \mathbf{0.76}$$

Case	X	Y
1	2	4
2	2	3
3	3	3
4	4	3
5	5	6
6	5	4
7	6	5
8	6	6
9	7	6
10	7	5



r = 0.76

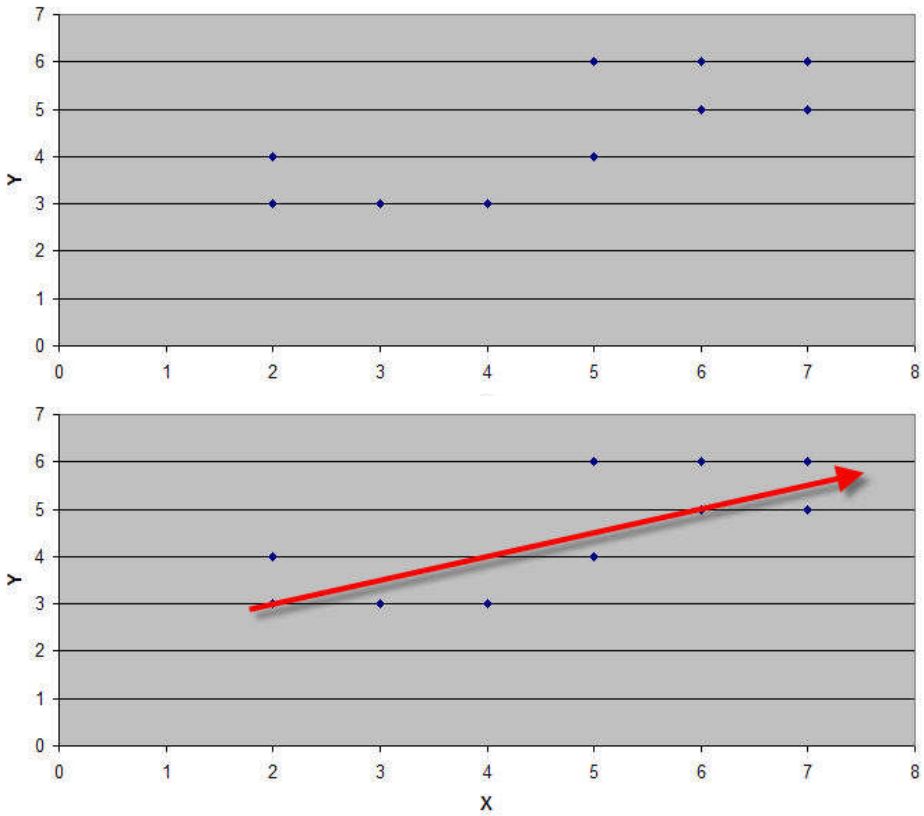
r is referred to as the *correlation coefficient*

Interpretation Scale:

.0 - .2	No relationship to very weak association
.2 - .4	Weak association
.4 - .6	Moderate association
.6 - .8	Strong association
.8 to 1.0	Very strong to perfect association

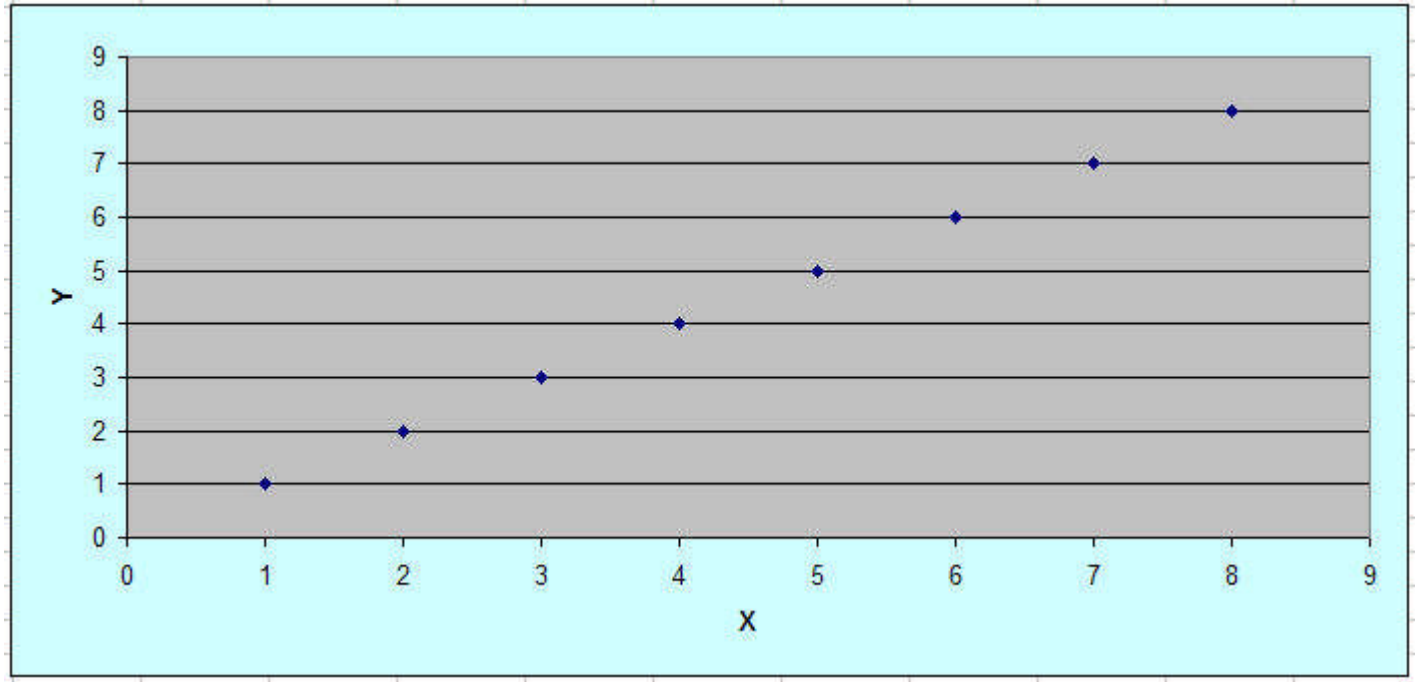
Case	X	Y
1	2	4
2	2	3
3	3	3
4	4	3
5	5	6
6	5	4
7	6	5
8	6	6
9	7	6
10	7	5

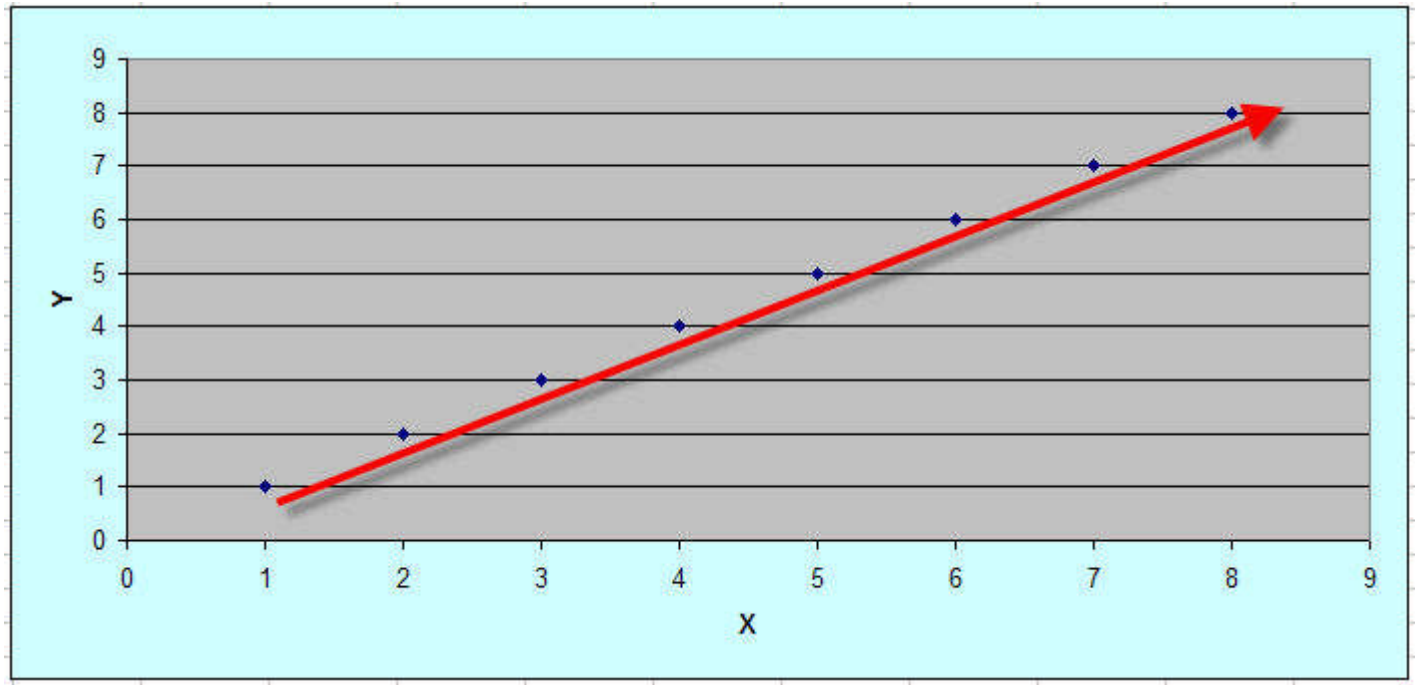
r = 0.76



Hence, a strong, positive association

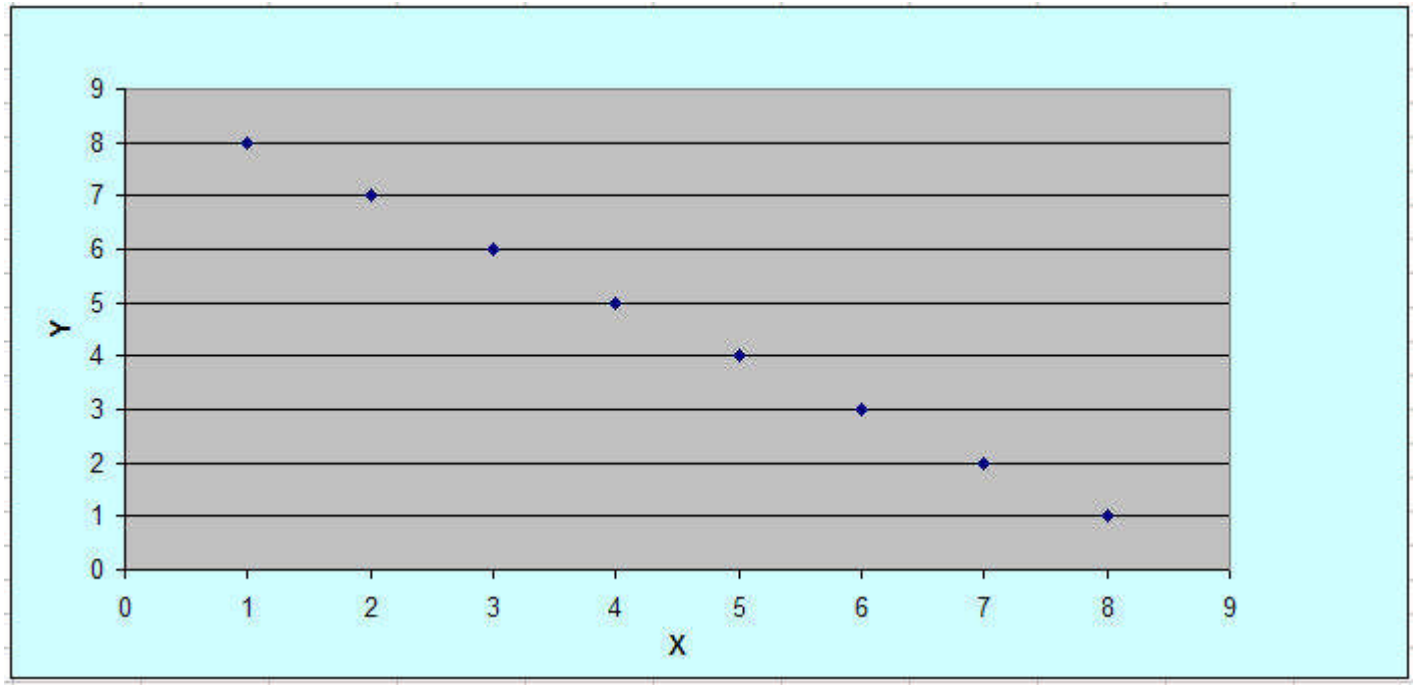
X	Y
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8

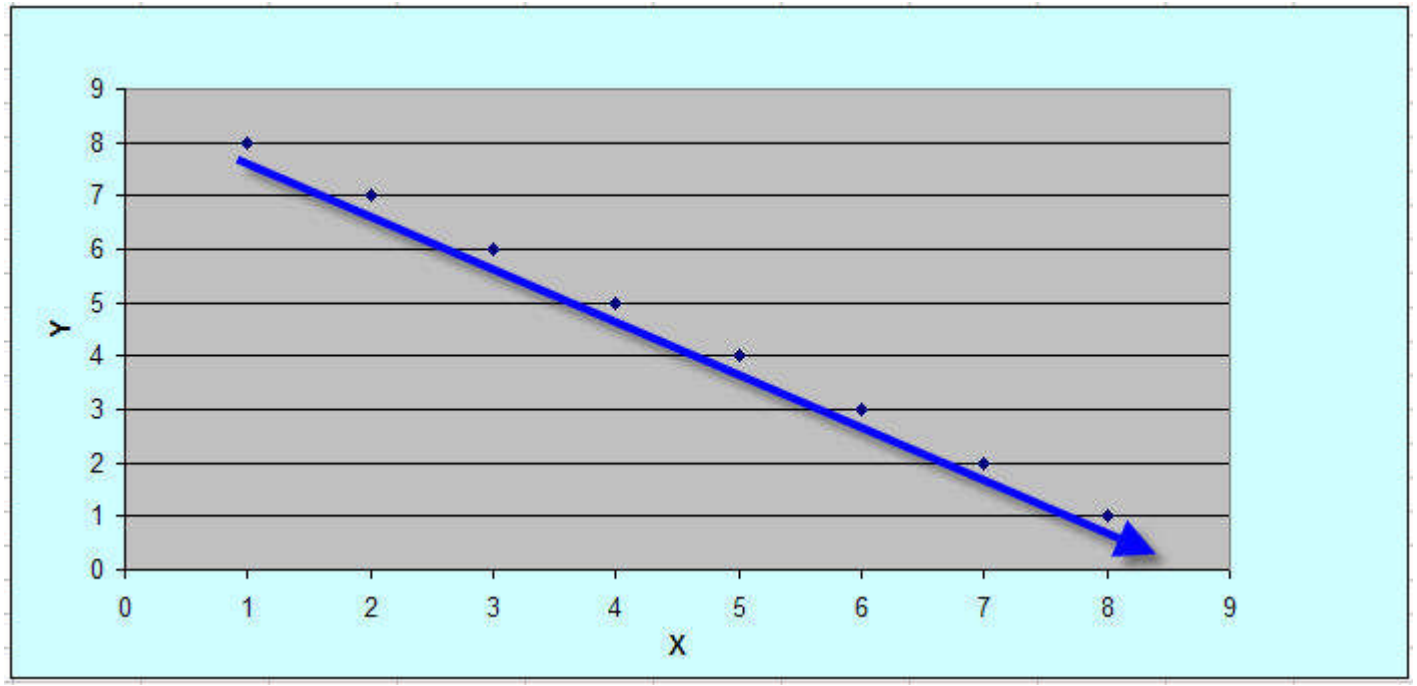




Perfect Positive Association

X	Y
1	8
2	7
3	6
4	5
5	4
6	3
7	2
8	1





Perfect Negative Association

CORRELATION ANALYSIS

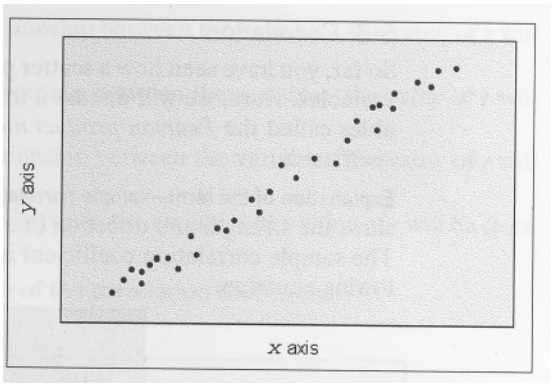
Pearson's r (or Pearson's rho, ρ)

Calculated value ranges from +1.0 to -1.0

Value depends upon

Strength of relationship

Direction



r = +0.94

Very strong, positive relationship

Formula for Pearson's r *(Raw score method)*

$$r = [\, n\Sigma XY - (\Sigma X)(\Sigma Y) \,] \div$$

$$\text{Square root of } ([\, n(\Sigma X^2) - (\Sigma X)^2 \,][\, n(\Sigma Y^2) - (\Sigma Y)^2])$$

where ***n*** is the number of cases

Case	X	Y	XY	X^2	Y^2
1	2	4	8	4	16
2	2	3	6	4	9
3	3	3	9	9	9
4	4	3	12	16	9
5	5	6	30	25	36
6	5	4	20	25	16
7	6	5	30	36	25
8	6	6	36	36	36
9	7	6	42	49	36
10	7	5	35	49	25
Σ=	**47**	**45**	**228**	**253**	**217**

Interpretation Scale:

r =	
.0 - .2	**Very weak** association
.2 - .4	**Weak** association
.4 - .6	**Moderate** association
.6 - .8	**Strong** association
.8 to 1.0	**Very strong** association

The value of ***r*** is referred to as the ***correlation coefficient***

r^2 is referred to as the ***coefficient of determination***

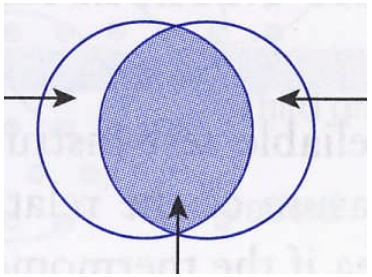
A measure of the explained variance

……the amount of variation in one variable that is attributable to variation in the other variable

Example: comparing level of education with number of memberships in voluntary organizations results in ***r*** = +0.7 (a strong association).

$r^2 = 0.49$ or 49%

Hence, 49% of the variation in the number of memberships is attributable to variation in level of education



51% does NOT overlap

49%

Hypothesis Testing with Correlation

Bob, a junior high and high school music teacher, wants to recruit the best students for his music program. He suspects there is a relationship between music grades and scores on the state tests in math and reading so he sets up a correlation study comparing music grades with math scores and music grades with reading scores.

H_0: There is no significant difference between music grades and math scores.

H_0: There is no significant difference between music grades and reading scores.

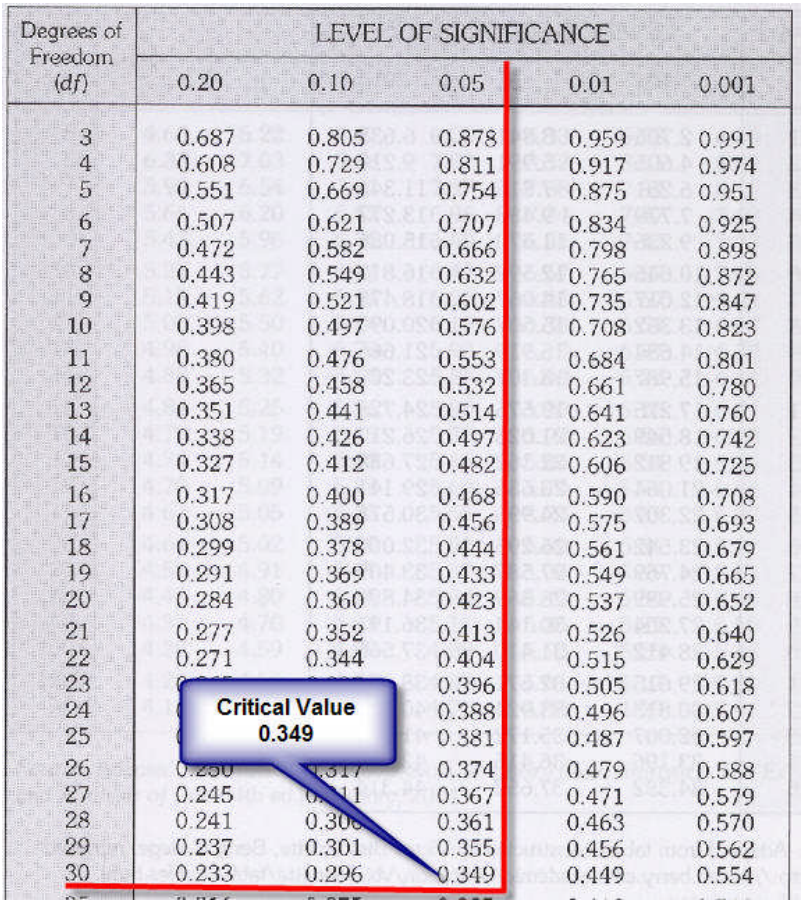
Step 2: Determine the critical value

α = 0.05 (level of significance)

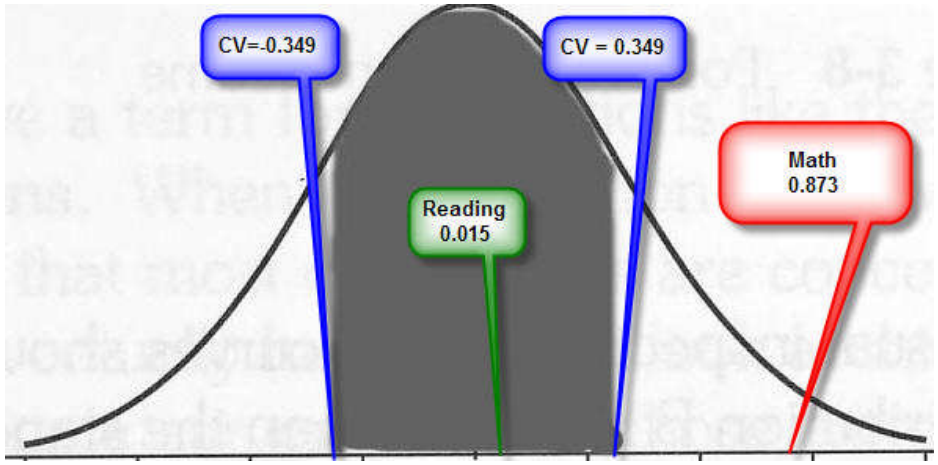
number of students = 32 (n)

degrees of freedom = 30 (n-2, since we have two variables)

Degrees of Freedom (*df*)	LEVEL OF SIGNIFICANCE				
	0.20	0.10	0.05	0.01	0.001
3	0.687	0.805	0.878	0.959	0.991
4	0.608	0.729	0.811	0.917	0.974
5	0.551	0.669	0.754	0.875	0.951
6	0.507	0.621	0.707	0.834	0.925
7	0.472	0.582	0.666	0.798	0.898
8	0.443	0.549	0.632	0.765	0.872
9	0.419	0.521	0.602	0.735	0.847
10	0.398	0.497	0.576	0.708	0.823
11	0.380	0.476	0.553	0.684	0.801
12	0.365	0.458	0.532	0.661	0.780
13	0.351	0.441	0.514	0.641	0.760
14	0.338	0.426	0.497	0.623	0.742
15	0.327	0.412	0.482	0.606	0.725
16	0.317	0.400	0.468	0.590	0.708
17	0.308	0.389	0.456	0.575	0.693
18	0.299	0.378	0.444	0.561	0.679
19	0.291	0.369	0.433	0.549	0.665
20	0.284	0.360	0.423	0.537	0.652
21	0.277	0.352	0.413	0.526	0.640
22	0.271	0.344	0.404	0.515	0.629
23			0.396	0.505	0.618
24			0.388	0.496	0.607
25			0.381	0.487	0.597
26			0.374	0.479	0.588
27	0.245		0.367	0.471	0.579
28	0.241	0.306	0.361	0.463	0.570
29	0.237	0.301	0.355	0.456	0.562
30	0.233	0.296	0.349	0.449	0.554



Step 3: Calculate Pearson's r



$r_{(music/math)} = 0.873$

$r_{(music/reading)} = 0.015$

critical value = 0.349

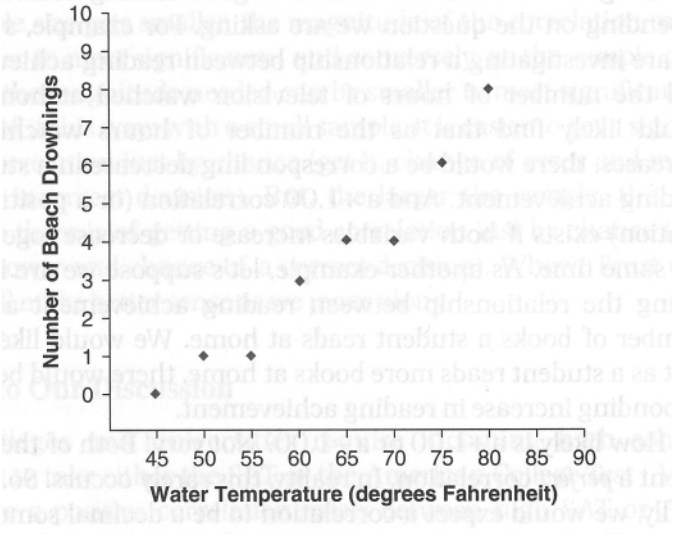
Hence, we reject the hypothesis that there is no significant difference between music grades and math scores, but accept the hypothesis that there is no significant difference between music grades and reading scores.

Bob can reasonably expect that students who score well on the state math test will do well in music.

CAUTION!

Correlation does not mean Causation

Researchers found a high correlation between water temperature and number of drownings



Researchers found a high correlation between the number of bottles of suntan lotion sold at a store and number of drownings

Researcher found a strong negative correlation between staff morale and frequent staff turnover

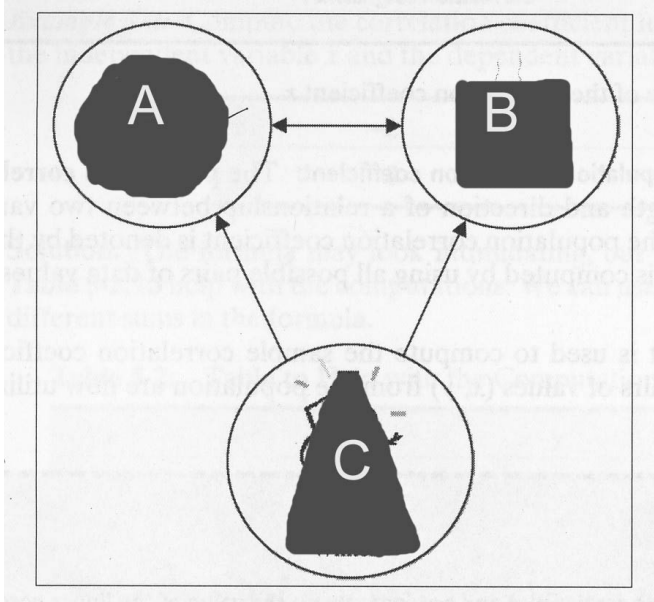
Does frequent staff turnover cause low morale?

Does low morale cause frequent staff turnover?

Or, is there a third factor that might cause both?

Violent students

Budget restrictions / low salaries



POSSIBILITIES with Correlation

- Direct cause and effect relationship – A causes B

  lack of water causes dehydration

- Reverse cause and effect relationship – B causes A

  Absences cause bad grades or bad grades cause absenses?

- Relationship may be due to chance – suntan lotion v. drownings

- Relationship may be due to confounding – interactions among several factors (Staff turnover)