Statistics 10-3

Pearson's Product Moment Correlation Coefficient

Statistics 10-3

Homework

Discussion p541

Ex 10-3 p 541 4, 5, 9, 10, 11, 12, 13, 21, 25, 26

CW questions a, b, and c

CW2 questions a, b, c, and d

Regression Activity 1 - 7

Objective

Compute the correlation coefficient, r, for two quantitative, continuous, normally distributed variables.

Correlation

This most common statistic for measuring correlation is the Pearson product moment correlation coefficient or **Pearson's r**. The correlation coefficient is denoted **r** for a sample and **ρ** (rho) for the population.

The correlation coefficient can range from -1 to 1.

$$-1 \leq \rho \leq 1$$

A correlation coefficient of 1 denotes a perfect positive (direct) relationship between variables. A correlation coefficient of -1 denotes a perfect negative (inverse) relationship between the variables.

Correlation

The closer to ±1 the correlation coefficient the stronger the relationship between the variables.

When we say the relationship is strong, we mean that the variables tend to **cluster to a line**. This suggests the variables have a linear relationship and the closer the data values fit a linear equation the stronger the relationship.

It is extremely important not to confuse correlation with causation. Two variables may correlate strongly but that in **no way** implies that changes in one variable **cause** changes in the other variable.

Correlation ≠ Causation

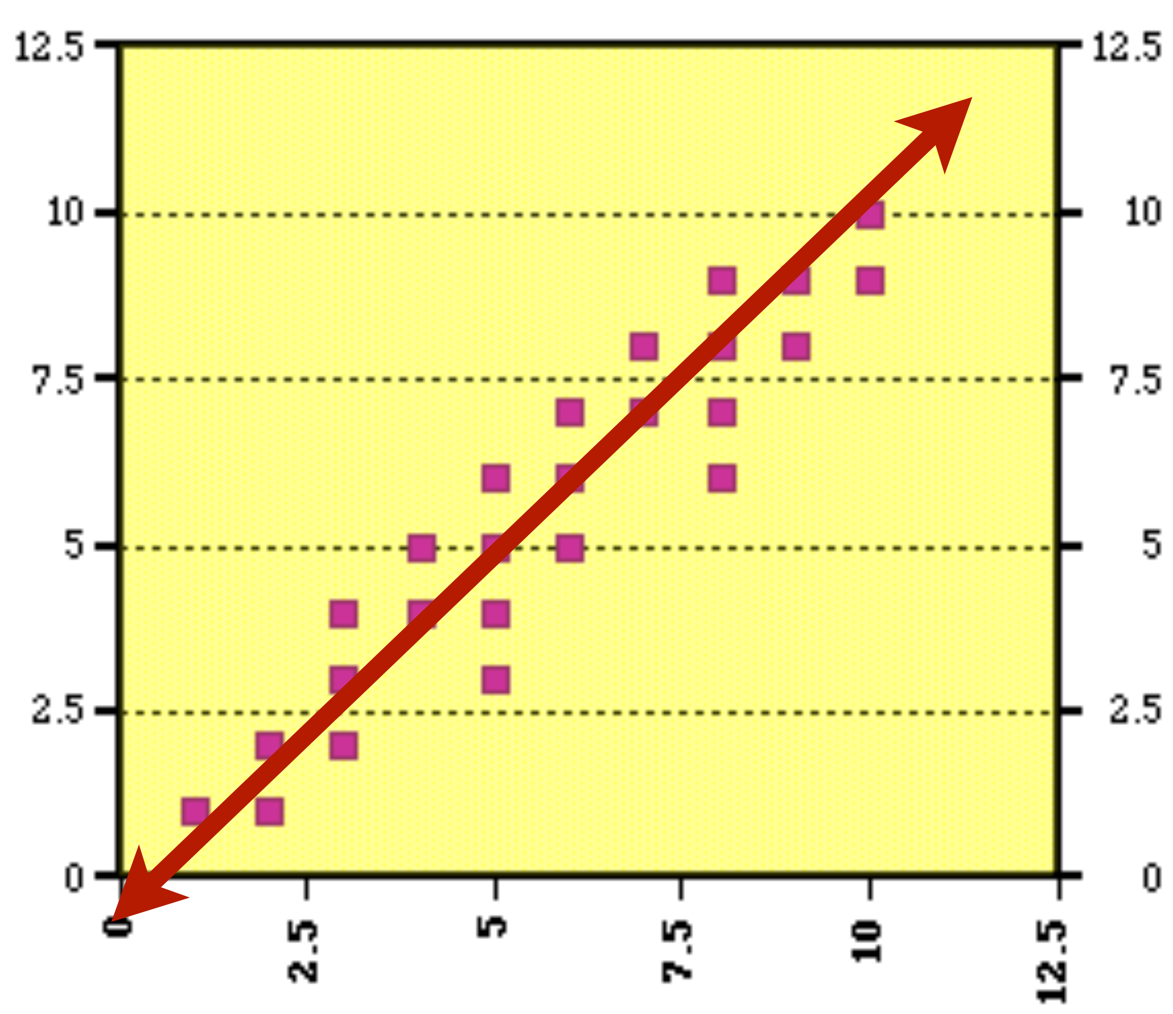
The number of churches in a city is strongly correlated with the amount of crime in that city.

Students taking a 4th year math class in high school will have better academic results in college.

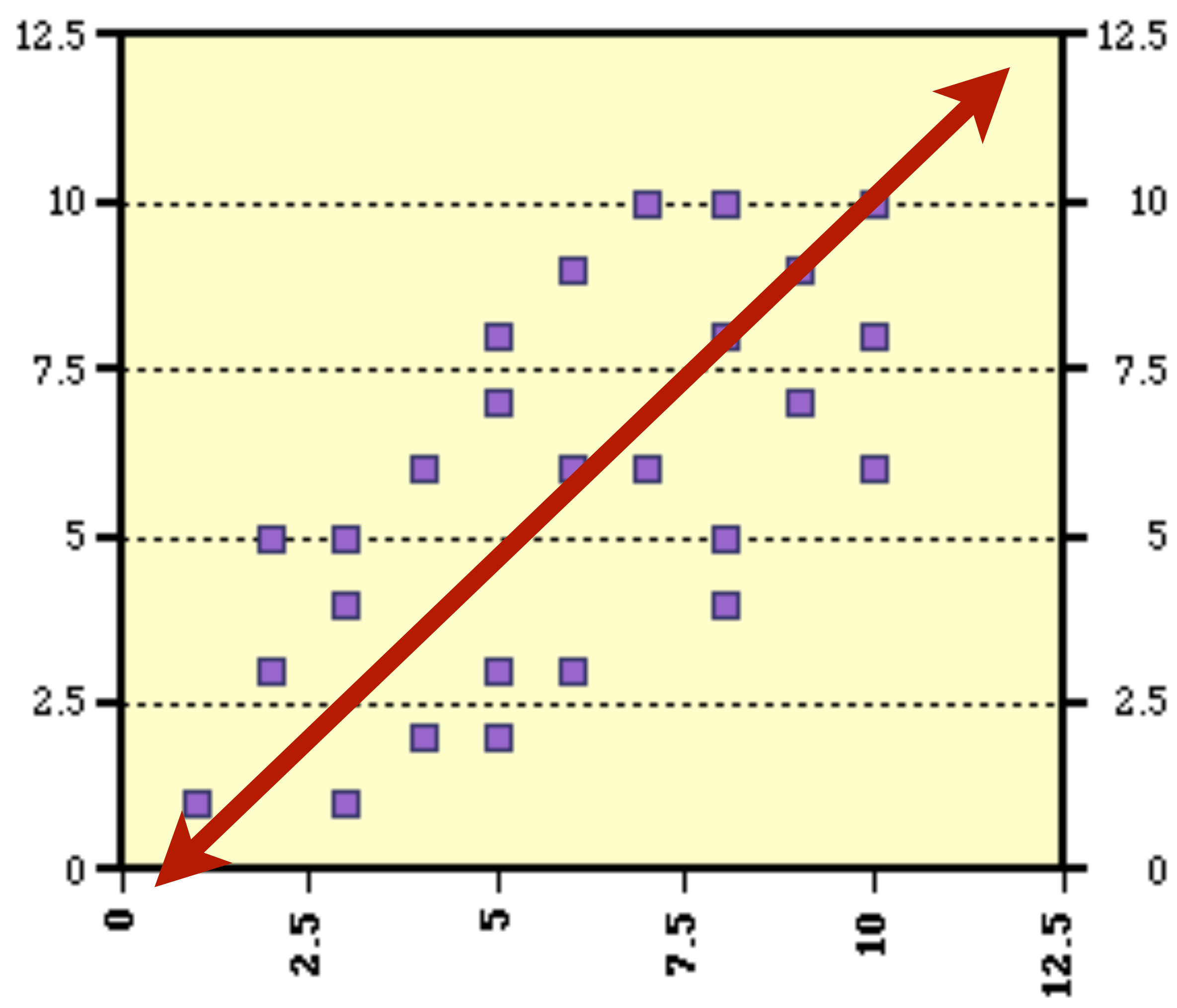
There are many reasons why a correlation does not imply causation. There might be another variable affecting both variables of study, it might just be coincidence, the relationship may be far more complex than simple cause-and-effect, and the relationship may be causal but which causes which?

How Well Do the Data Fit a Line?

High Positive Correlation



Low Positive Correlation



Note that in the above examples, the data on the left tend closer to a line. In this case, both lines suggest the same model, but the data on the left is a closer fit, suggesting a better predictive value for the model.

Pearson's Product Moment Correlation Coefficient.

Also called **Pearson's r**, between two variables is defined as the **covariance** of the two variables divided by the product of their standard deviations.

For a population the formula for finding the correlation coefficient is:

$$\rho = \frac{\mathrm{cov}(x,y)}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$$

where E is the expected value operator (in this case, mean), and cov means covariance.

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

Pearson's r

$$\rho = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y} = \frac{E[(x-\mu_x)(y-\mu_y)]}{\sigma_x \sigma_y}$$

We've already met the variance: it's the mean value of all the squared differences from the mean.

The covariance is similar: it's the mean value of all the pairs of differences from the mean (deviations) for X multiplied by the differences from the mean (deviations) for Y.

If X and Y aren't closely related to each other, they don't co-vary, so the covariance is small, and the correlation is small. If X and Y are closely related, cov(XY) turns out to be almost the same as $\sigma_x \cdot \sigma_y$, so the correlation is near 1.

Pearson's r

The previous formula defines the **population correlation coefficient**, represented by the Greek letter ρ (rho). Substituting estimates of the covariances and variances based on a **sample** gives the **sample correlation coefficient, r**:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}_i)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}_i)}{(n-1)s_x s_y}$$

This is NOT how you will calculate r. This is to help you understand what r actually is.

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y_i})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y_i})}{(n-1)s_x s_y}$$

To calculate this value we create several lists of data similar to the lists used finding standard deviation.

x_i	y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$(X_i - \bar{X})(Y_i - \bar{Y})$
⋮	⋮	⋮	⋮	⋮	⋮	⋮
$\sum_{i=1}^{n} X_i$	$\sum_{i=1}^{n} Y_i$			SS_{xx}	SS_{yy}	SS_{xy}

I really do not want to do all that work. Do you?

Pearson's r and Standard Scores

An equivalent expression gives the correlation coefficient as the mean of the products of the standard scores. Based on a sample of paired data (Xi, Yi), the sample Pearson correlation coefficient is

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_x}\right)\left(\frac{Y_i - \bar{Y}}{s_y}\right) \qquad r = \frac{1}{n-1}\sum Z_x Z_y$$

Sometimes the formula is given with N instead of n-1. If you are using the data from the population use N.

Calculational Formula

A commonly used calculation formula is shown below. The formula looks a bit complicated, but taken step by step as shown in the following example, it is actually relatively simple. I am sure you know how I feel about this formula, but if you do not have a TI-84 or equivalent, this is what you use.

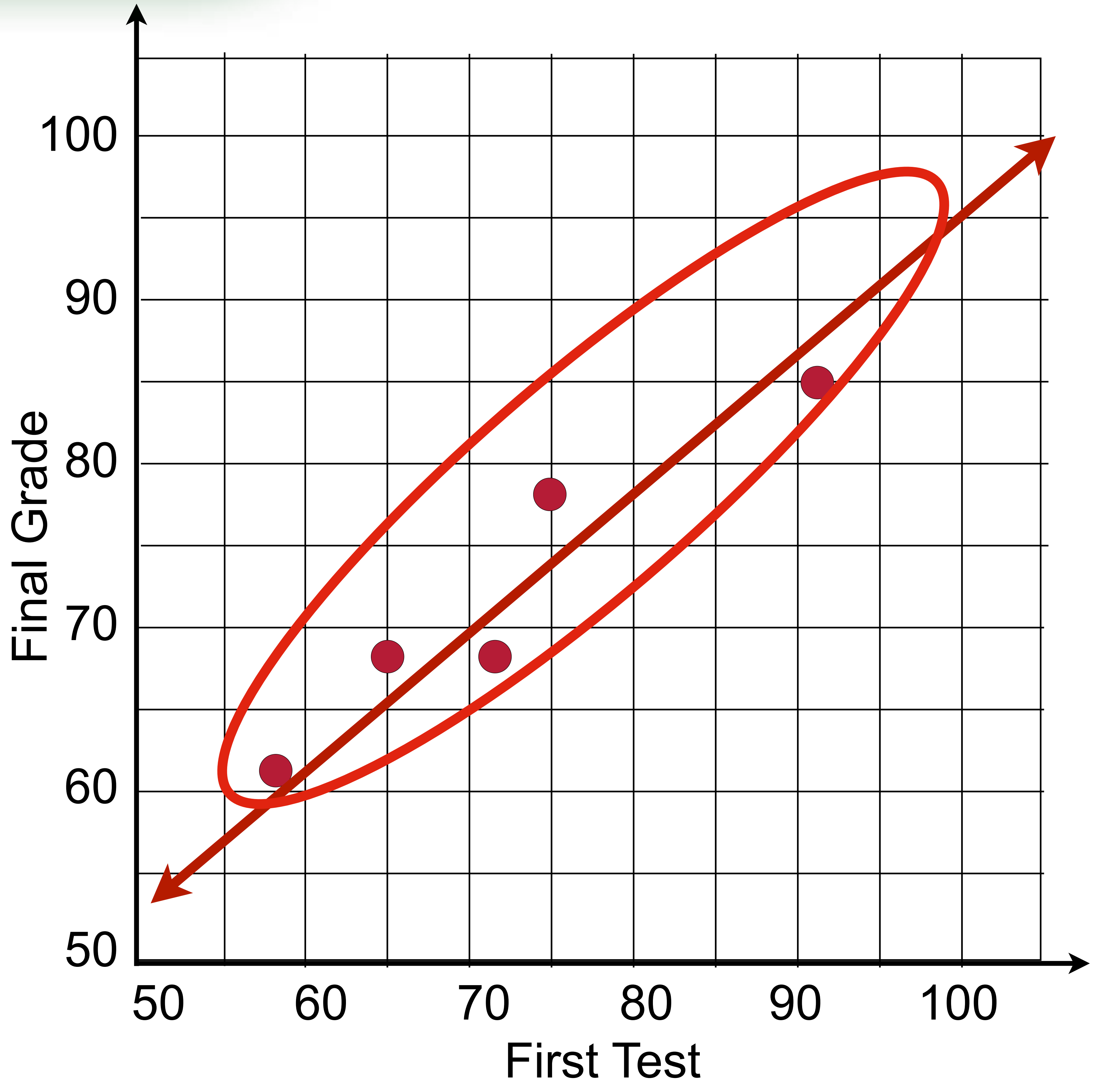
$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{\left(\sum X\right)^2}{N}\right)\left(\sum Y^2 - \frac{\left(\sum Y\right)^2}{N}\right)}} = \frac{N\sum XY - \sum X \sum Y}{\sqrt{\left(N\sum X^2 - \left(\sum X\right)^2\right)\left(N\sum Y^2 - \left(\sum Y\right)^2\right)}}$$

The second form is as shown in your book.

DO NOT USE THIS FORMULA, USE YOUR CALCULATOR.

Example

Grade First Test	Final Grade
73	70
86	80
93	96
92	85
72	68
65	68
58	62
75	78



Note: that the points come close to forming a line.

Example

Grade First Test	Final Grade	Z_x	Z_y	Z_xZ_y
73	70	-0.2978	-0.5298	0.1577
86	80	0.7341	0.3720	0.2731
93	96	4.2896	1.8147	2.3403
92	85	1.2102	0.8228	0.9958
72	68	-0.377	-0.7101	0.2677
65	68	-0.9325	-0.7101	0.6622
58	62	-1.488	-1.251	1.8617
75	78	-0.1389	0.1919	-0.8266
614	607			6.5317

Note the similarity between the z scores for each variable. The size and sign are similar. Note also that the product, with one exception, is positive. These indicate the variables tend to behave alike from one point to the next.

$$r = \frac{1}{n-1}\sum Z_x Z_y \qquad r = \frac{1}{8-1}6.5317$$

$$r = .9331$$

Interpreting Pearson's r

When interpreting Pearson's r, discuss **shape**, **strength**, **direction**, and **unusual points.**

In this example the correlation coefficient (.9331) and graph suggest a **strong**, **linear**, **positive** relationship between grade on first exam and final grade in class. There are no unusual points.

That suggests that as grade on first exam increases, the final grade in the class also **tends** to increase.

To complete our interpretation, we need to account for r^2.

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

Coefficient of Determination

r^2 (called the **coefficient of determination**) can be interpreted as the proportion of variance in Y that is contained in (or explained by) X (the model).

$1 - r^2$ is the **coefficient of non-determination**, the proportion of variance **not** explained by the model.

r^2 tells us what percentage of the variability in the response variable is accounted for (explained by) by the changes in the predictor variable (or the model).

If $r^2 = .63$, then 63% of the variability in the response variable is explained by the model.

Interpreting Pearson's r

Now we can expand on our interpretation of **r** to include **r^2**, and give a complete response for our correlation analysis.

In this example the graph and correlation coefficient (r = .9331) suggest a **strong, linear, positive** relationship between grade on first exam and final grade in class. As grade on the first test increases, the final grade tends to increase.

87.07% (**r^2**) of the variability in final grade is explained by changes in the first exam grade.

Significance of Pearson's r

We can always find a correlation coefficient. Given two equal sized sets of numbers we can create ordered pair that when put into the formula will find a correlation coefficient. The question then becomes; "is that coefficient significant or meaningful?".

That is not an easy question to answer. The correlation between smoking and lung cancer is usually reported between 0.5 and 0.7. Is that significant? Given the gravity of the consequences from smoking, we would probably say yes.

So significance is definitely context sensitive.

Significance of Pearson's r

To test for **statistical significance** we follow the same steps as previous hypothesis testing but there are some assumptions that must be met:

1. The variables are random, quantitative, and continuous.
2. The variables do have a linear relationship.
3. The variables are **bivariate normal (homoscedastic)**. i.e. at every value of the normally distributed independent variable, the dependent variable is normally distributed.

Testing for Significance

To test the significance of Pearson's r:

1. State the hypotheses
2. Determine the statistic and find the critical value.
3. Compute the test statistic and p-value
4. Decide using the criteria (α) determined apriori.
5. Tell the reader what you concluded.

The hypotheses are about the population parameter, ρ. Since the value of ρ is between -1 and 1, with 0 meaning no relationship between the variables:

$$H_0: \rho = 0; \quad H_a: \rho \neq 0$$

Testing for Significance

H_0: $\rho = 0$; H_a: $\rho \neq 0$

One statistic used is the **two-tailed** t-statistic, calculated:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

n-2 degrees of freedom, and n≥6.

The reason for n-2 degrees of freedom lies in the least squares analysis of the regression model and residual error which are beyond the scope of this class. Simply consider that there are two variables involved in regression.

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Testing Significance

From our example where r = .933 and n = 8:

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

$$t = .933\sqrt{\frac{8-2}{1-.933^2}} = .933(6.806) \quad = 6.35$$

Grade First Test	Final Grade
73	70
86	80
93	96
92	85
72	68
65	68
58	62
75	78

We find the probability $p(t \geq 6.35)$ as we have done previously. But we do not need to calculate the probability for a t-score of 6.35 to know that we reject the null.

We find the critical value of t (t*) using invt(α/2, n-2)

Another method to determine significance is to look in a significance table like Table I on page 778

df (= N-2) (N= number of pairs)	Level of significance for one-tailed test			
	.05	.025	.01	.005
	Level of significance for two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708
11	.476	.553	.634	.684
12	.458	.532	.612	.661
13	.441	.514	.592	.641
14	.426	.497	.574	.628
15	.412	.482	.558	.606
16	.400	.468	.542	.590
17	.389	.456	.528	.575
18	.378	.444	.516	.561
19	.369	.433	.503	.549
20	.360	.423	.492	.537
21	.352	.413	.482	.526
22	.344	.404	.472	.515
23	.337	.396	.462	.505
24	.330	.388	.453	.495
25	.323	.381	.445	.487
26	.317	.374	.437	.479
27	.311	.367	.430	.471
28	.306	.361	.423	.463
29	.301	.355	.416	.456
30	.296	.349	.409	.449
35	.275	.325	.381	.418
40	.257	.304	.358	.393
45	.243	.288	.338	.372
50	.231	.273	.322	.354
60	.211	.250	.295	.325
70	.195	.232	.274	.302
80	.183	.217	.256	.284
90	.173	.205	.242	.267
100	.164	.195	.230	.254

We find critical values of r by locating the intersection of the $\alpha/2$-level and d.f.. If our calculated value of r is greater than the value in the table, we have significance. For negative values we look for values more negative.

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

Testing Significance

For our example, if we wished to test at the .05 level we compare our calculated value of .933 to the interval bounded by the value given in the table.

If |r| > the value in the table we consider the correlation coefficient (r) significant.

α = .05, two tail test

d.f. = 8 - 2 = 6

.933 > .707 thus we reject the null of ρ = 0, and contend our correlation is statistically significant.

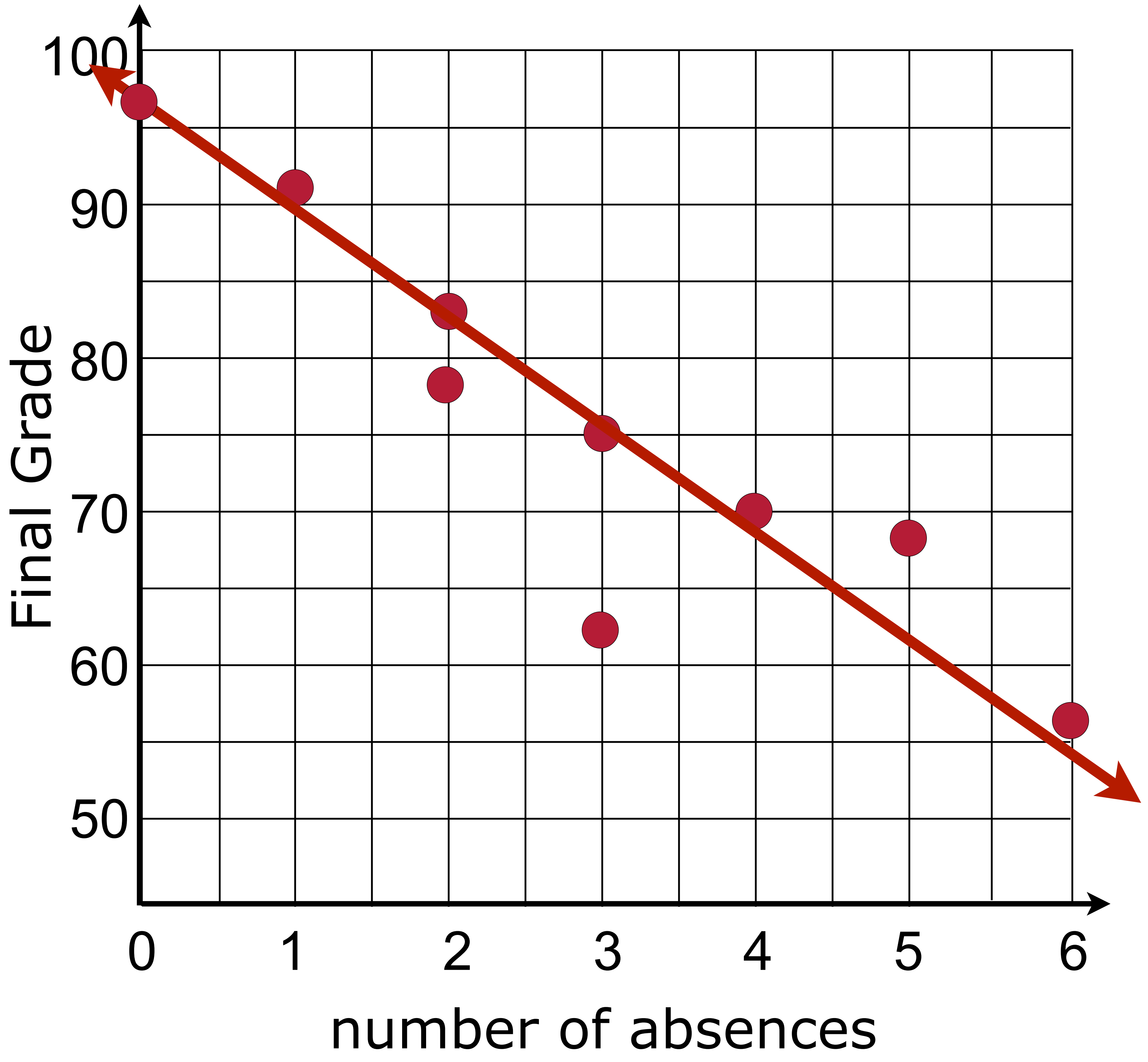
df (= N-2) (N= number of pairs)	Level of significance for one-tailed test			
	.05	.025	.01	.005
	Level of significance for two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	[illegible]	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708

.707

Another Example

Suppose we are interested in the relationship between number of absences a student has and the final grade the student earns.

No. of Absences	Final Grade
0	96
1	91
2	78
2	83
3	75
3	62
4	70
5	68
6	56



The scatter plot shows a **strong**, **negative**, **linear** relationship, with one potentially unusual result at (3, 62).

As number of absences increases, final grade **tends** to decrease.

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Another Example

No. of Absences	Final Grade	$x - \bar{x}$	$y - \bar{y}$	$(x-\bar{x})(y-\bar{y})$
0	96	-2.8889	20.5556	-59.38
1	91	-1.8889	15.5556	-29.38
2	78	-0.8889	2.5556	-2.272
2	83	-0.8889	7.5556	-6.716
3	75	0.1111	-0.4444	-0.0494
3	62	0.1111	-13.4444	-1.494
4	70	1.1111	-5.4444	-6.048
5	68	2.1111	-7.4444	-15.72
6	56	3.1111	-13.4444	-60.49
x = 2.8889 s_x=1.9003	y = 75.4444 s_y=13.0969			-181.5556

Once again note the pattern of the deviations for x and y. When x is larger than it's mean, y tends to be smaller than it's own mean and vice versa.

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}_i)}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}_i)}{(n-1)s_x s_y}$$

$$r = \frac{-181.5556}{8 \cdot 1.900 \cdot 13.0969} = -.9120$$

This is NOT how you will calculate r. This is to help you understand what r actually is.

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

TI-84

Now I suppose you would like to use the calculator to do all the work for you. To find the correlation coefficient (r) on the TI you have to prepare the calculator to report r.

If you want the TI-84 to calculate the Pearson correlation coefficient r, you must turn "Diagnostics" ON:

catalog

2nd 0 ∨ DiagnosticOn ENTER ENTER

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Another Example

No. of Absences	Final Grade
0	96
1	91
2	78
2	83
3	75
3	62
4	70
5	68
6	56

Stat → 1:Edit → Select List "L1" → Enter first value "0" → Enter

Enter 2nd value "1" → Enter → Repeat to end of list

Move to List "L2" → Enter first value "96" → Enter

Repeat to end of list → 2nd → Quit

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

To have the TI-84 plot the points created by the two lists:

2nd y= (STAT PLOT) Enter ON TYPE" Dotplot XList 2nd 1 YList 2nd 2 Mark Zoom 9

No. of Absences	Final Grade
0	96
1	91
2	78
2	83
3	75
3	62
4	70
5	68
6	56

Calculating r

To have the TI-84 calculate Pierson's r:

STAT ➢ CALC ⋎ 4:LinReg(ax+b) Enter
STAT ➢ CALC ⋎ 8:LinReg(a+bx) Enter

Xlist: L_1 2nd 1
Ylist: L_2 2nd 2
FreqList:
Store RegEQ:
Calculate Enter

y=ax+b
a=-6.284615385
b=93.6
r^2=.8315029586
r=.-.9118678405

No. of Absences	Final Grade
0	96
1	91
2	78
2	83
3	75
3	62
4	70
5	68
6	56

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

Result

No. of Absences	Final Grade
0	96
1	91
2	78
2	83
3	75
3	62
4	70
5	68
6	56

LinReg
y = ax+b
a = -6.284615385
b = 93.6
r^2 = .8315029586
r = -.9118678405

-.9119 indicates a very strong negative linear relationship, with a possible outlier at (3, 62).
As number of absences increases, final grade tends to decrease. 83% of the variability in final grade is accounted for by change in number of absences.

Round to 4 decimal places

Hypothesis Test

Now we do the hypothesis test.

H_0: ρ = 0; H_1: ρ ≠ 0

$$t = -.9119\sqrt{\frac{9-2}{1-.9119^2}} = -5.8775$$

Obviously significant. $t^* = \text{invt}(.025, 7) = -2.3646$

p(t < -5.8775) = tcdf(-10^99, -5.8775, 7) = .0003 x 2 = .0006

Remember, with tcdf and two tail test, p-value x 2.

There is a t-test, but we are not there yet.

-.912 < -.666, .0006 < .05, thus we **reject the null** of ρ = 0, and contend **our correlation is statistically significant**.

df (= N-2) (N= number of pairs)	Level of significance for one-tailed test			
	.05	.025	.01	.005
	Level of significance for two-tailed test			
	.10	.05	.02	.01
1	.988	.997	.9995	.9999
2	.900	.950	.980	.990
3	.805	.878	.934	.959
4	.729	.811	.882	.917
5	.669	.754	.833	.874
6	.622	.707	.789	.834
7	.582	.666	.750	.798
8	.549	.632	.716	.765
9	.521	.602	.685	.735
10	.497	.576	.658	.708

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

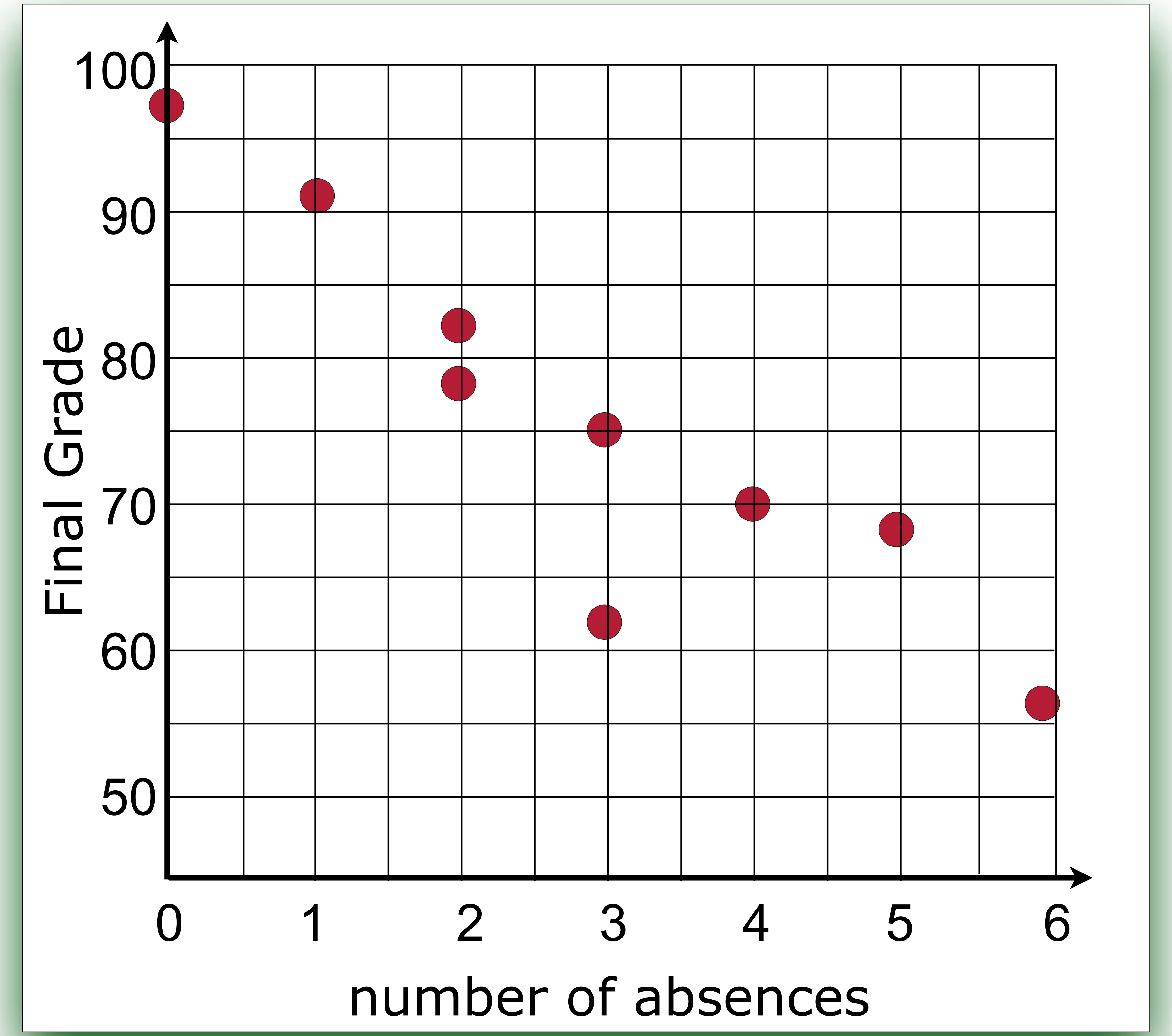
Statistics 10-3

In Conclusion

In our example the correlation coefficient (r = -.9119) and graph suggest a strong, linear, negative relationship between number of absences and final grade in class. There may be an outlier at 3 hours, with 62 final grade. As number of absences increase, final grade tends to decrease.

r^2 = .8315, indicating 83% of the variability in final grade is explained by variability in number of absences.

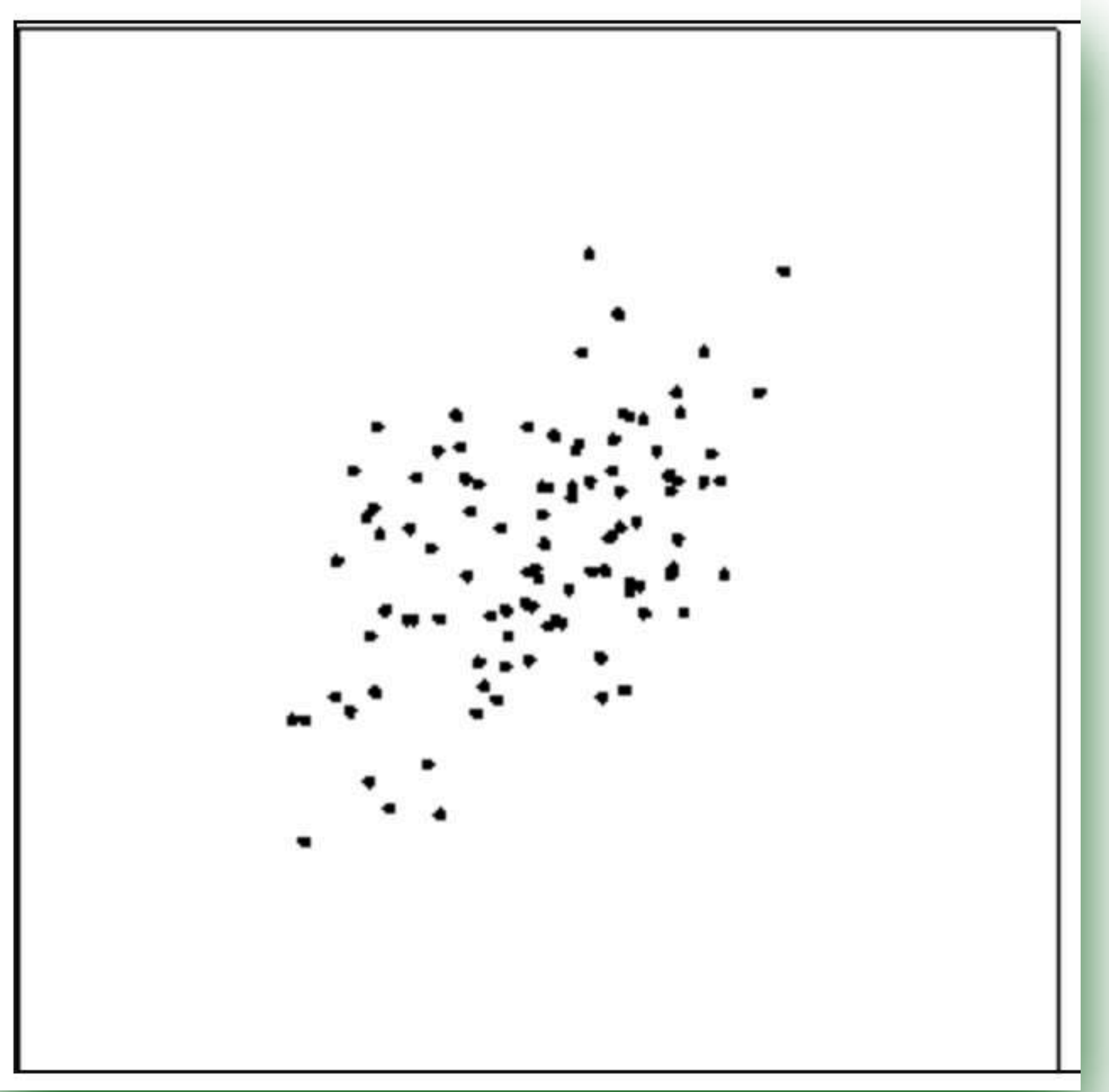
-.912 < -.666, .0006 < .05, thus we **reject the null** of $\rho = 0$, and contend **our correlation is statistically significant**.



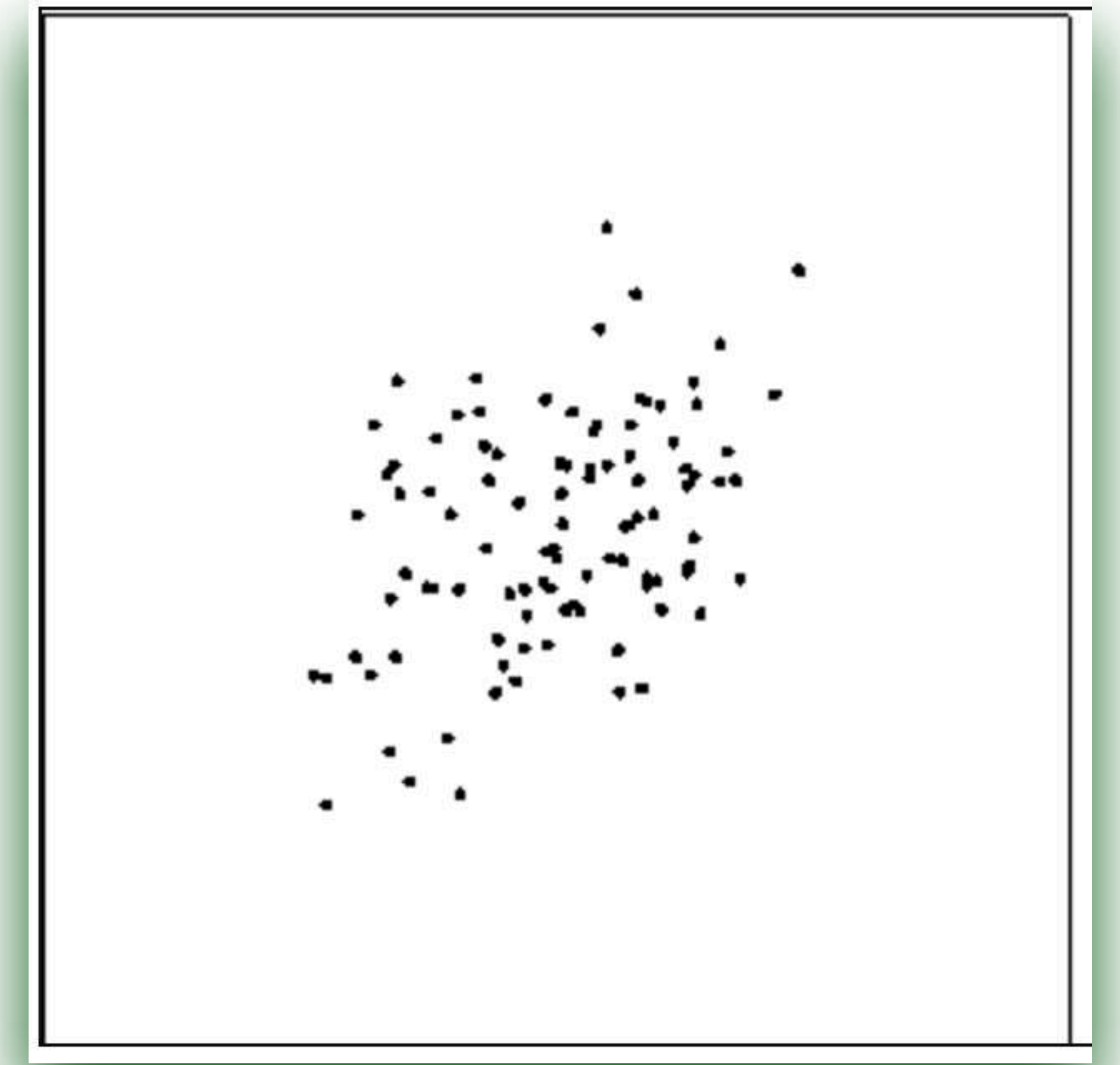
Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

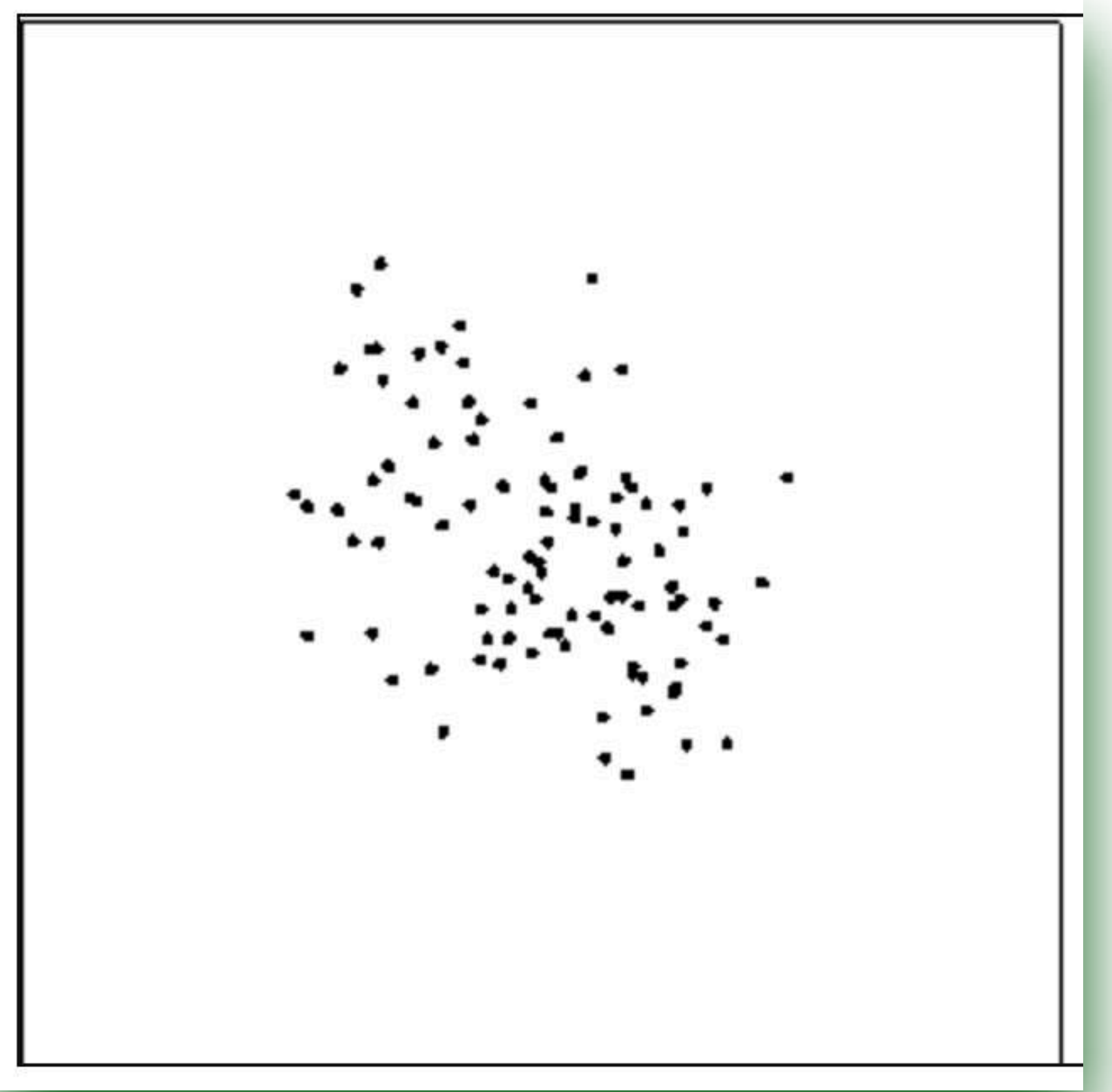
Estimate r



r = +0.5



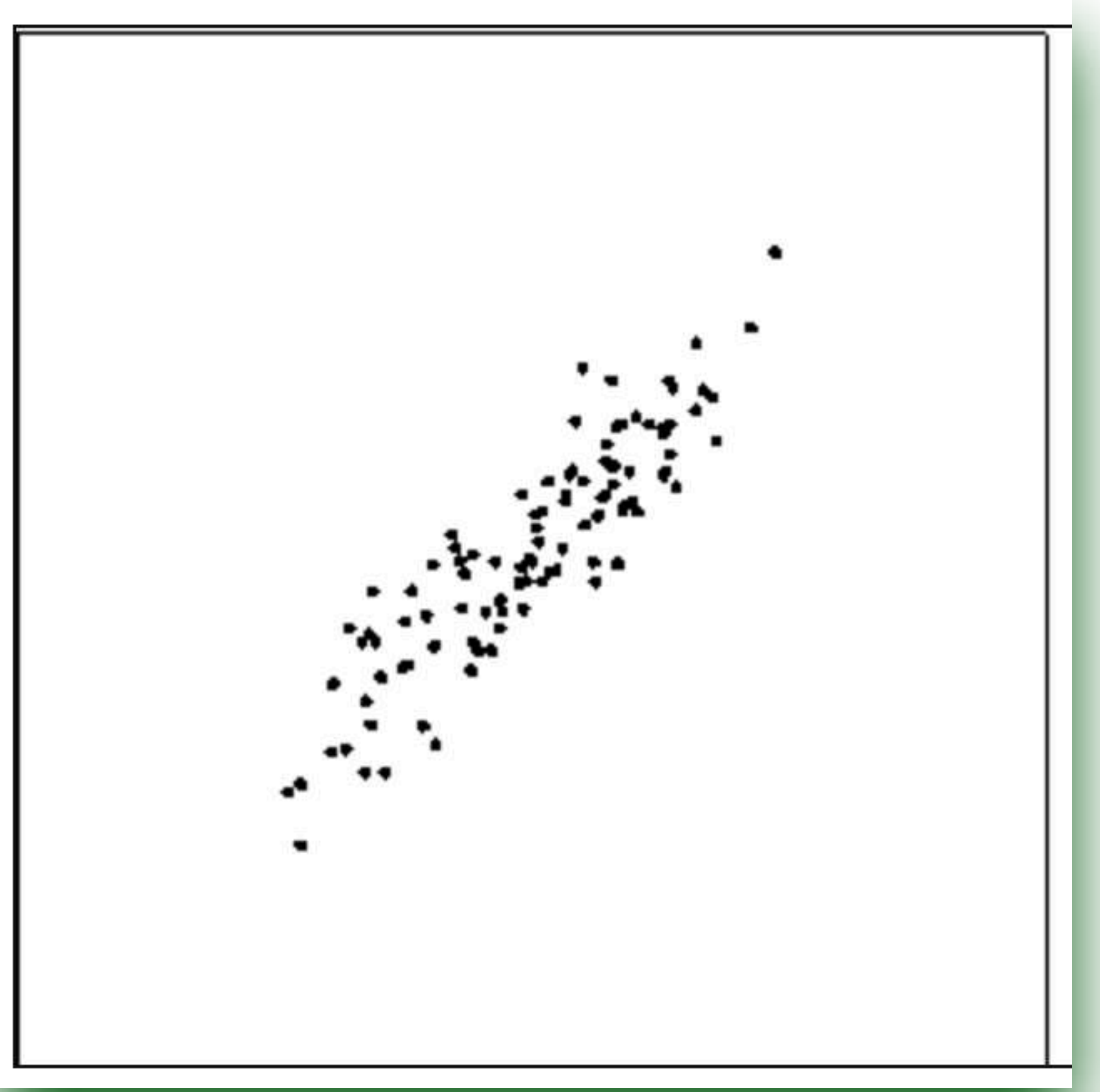
r = +0.4



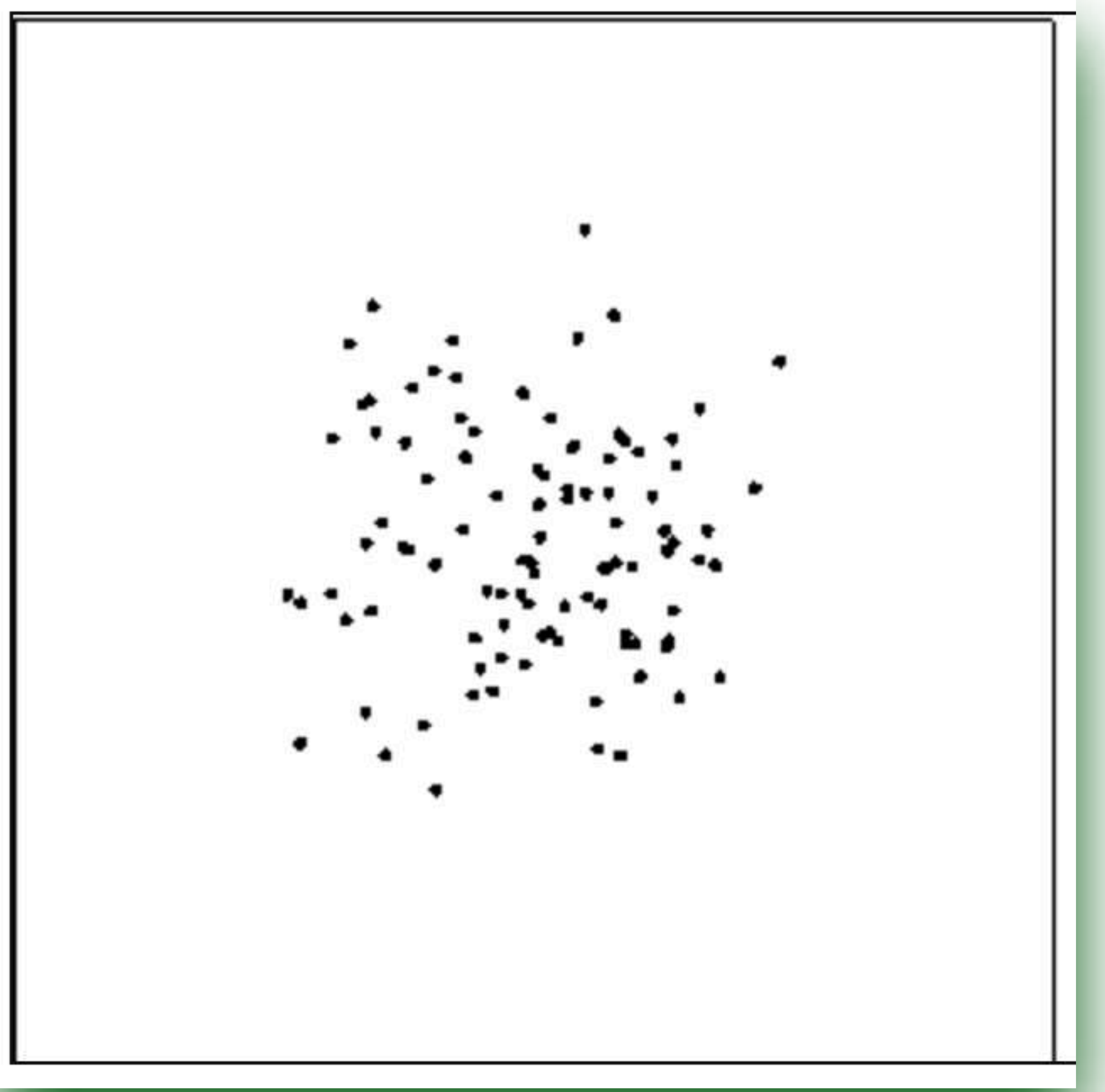
r = -0.4

Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

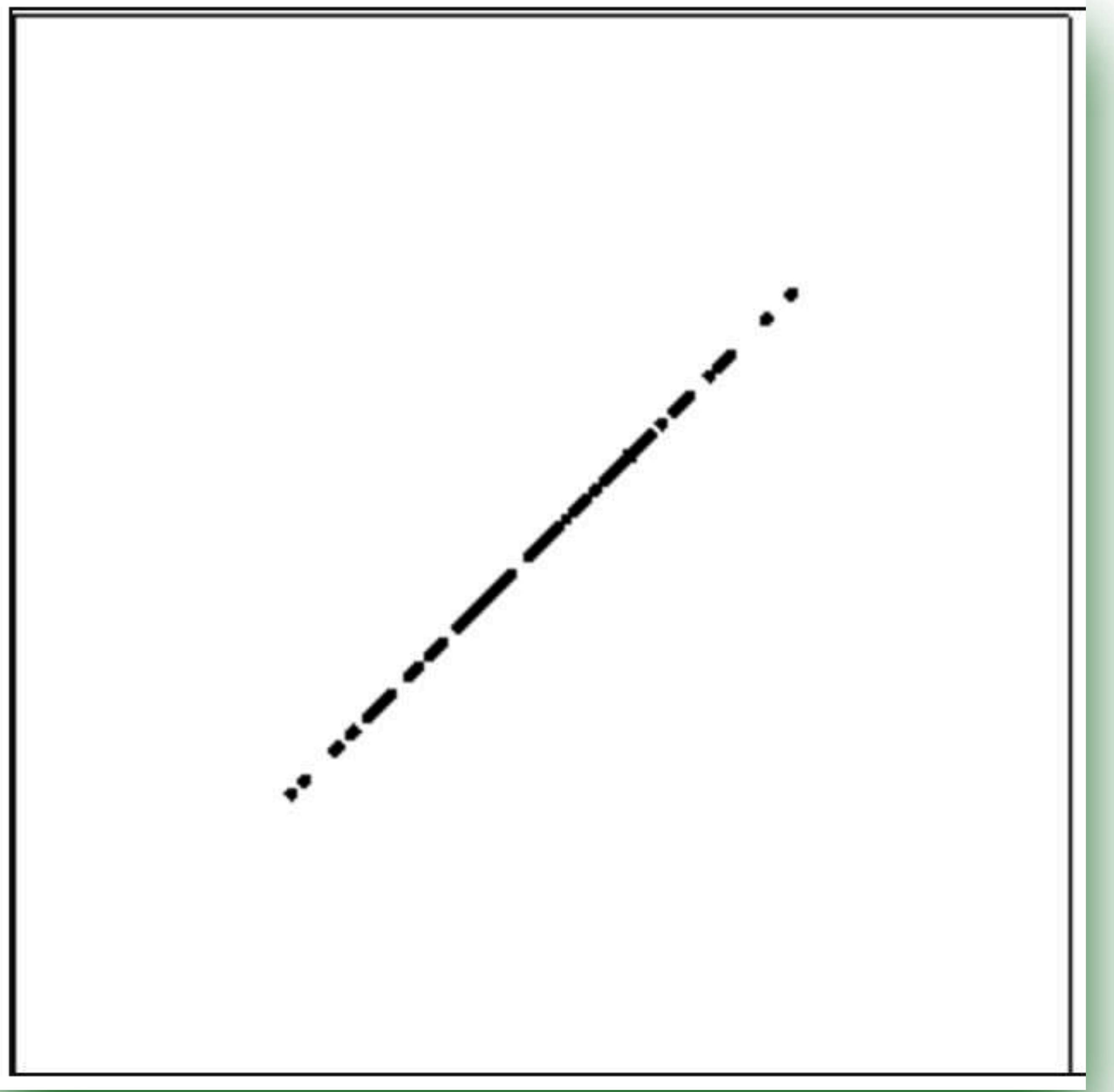
Estimate r



r = +.9



r = 0

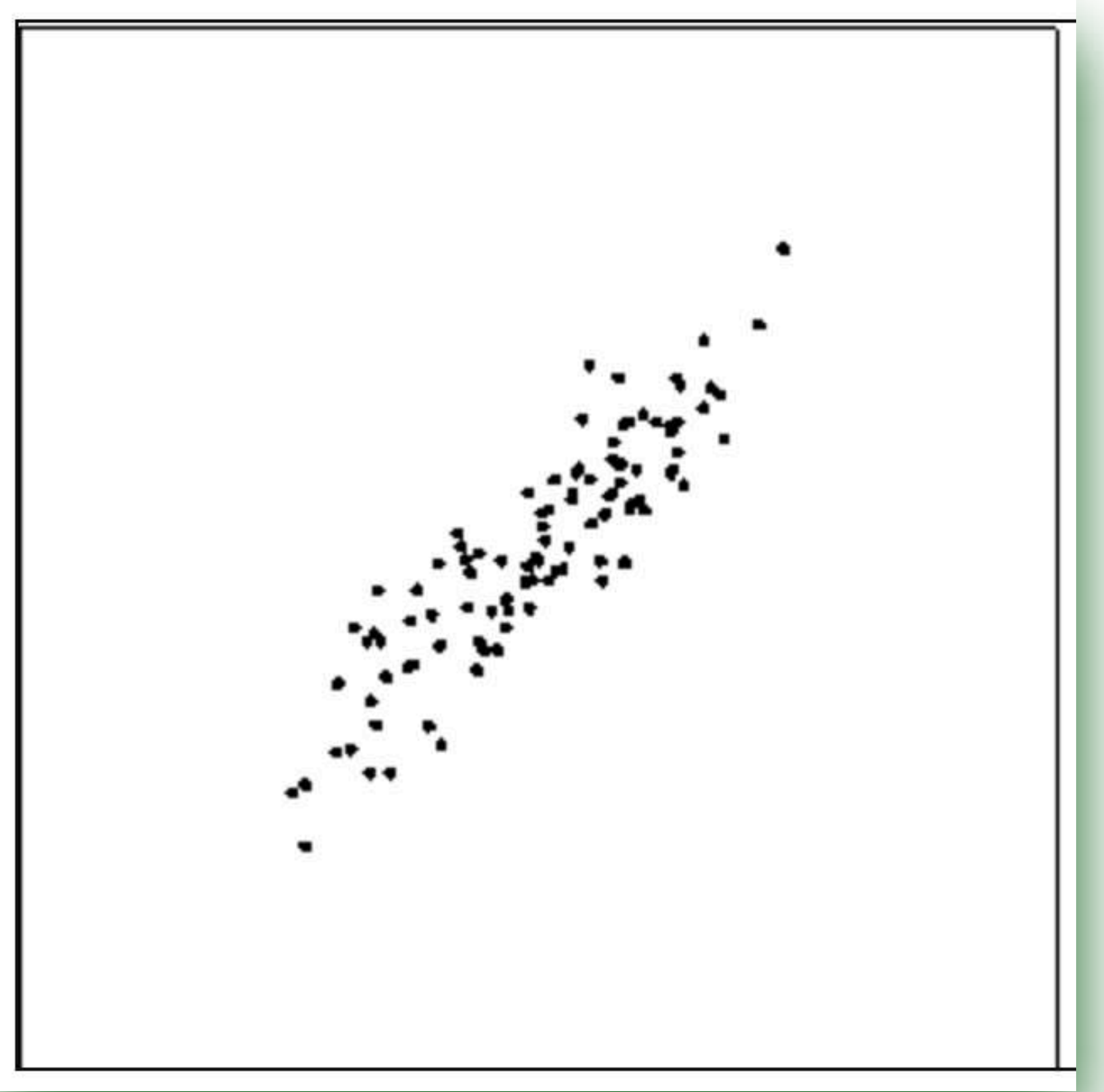


r = 1

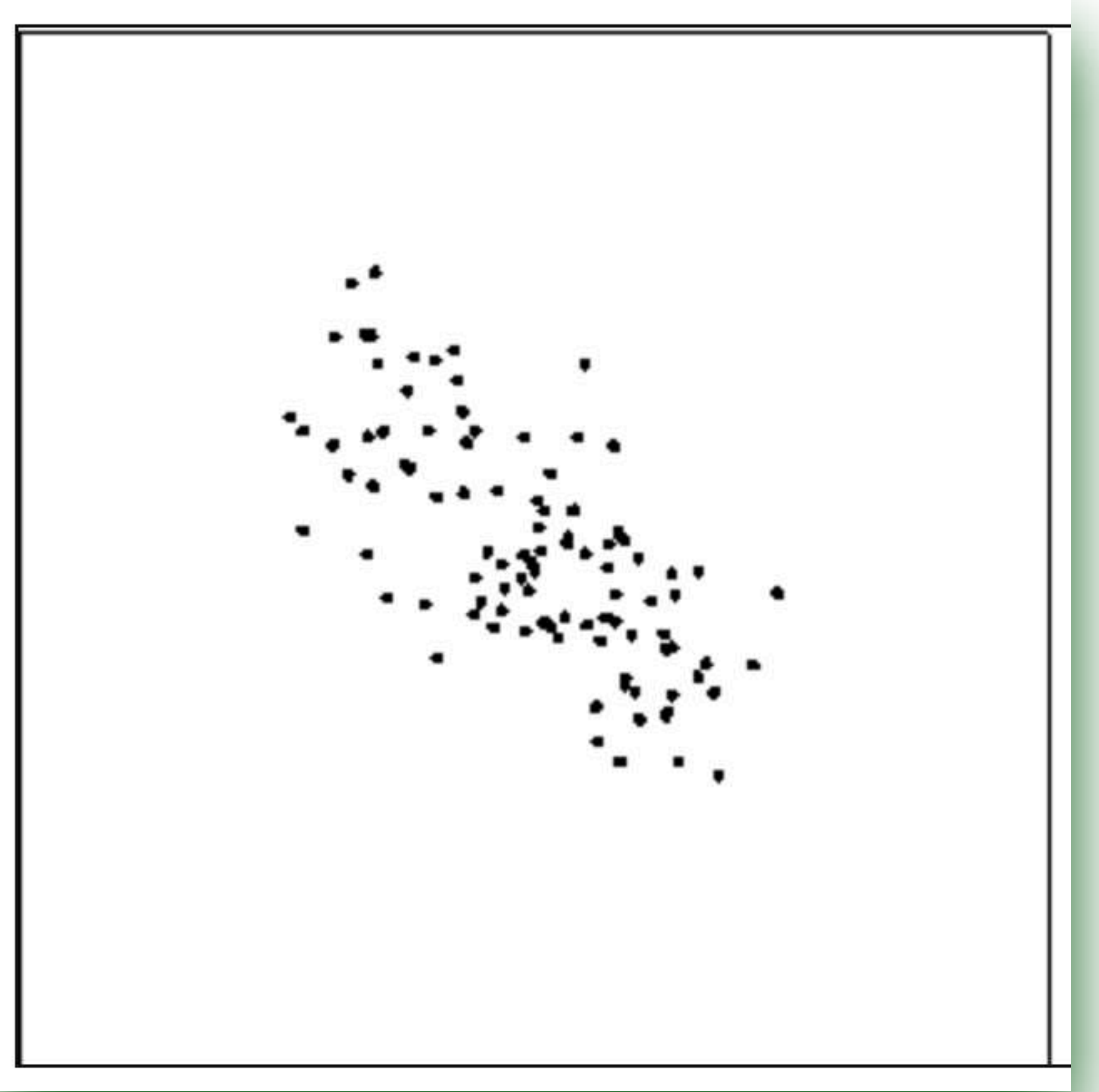
Objective: Students will will compute the correlation coefficient, r, for two quantitative variables.

Statistics 10-3

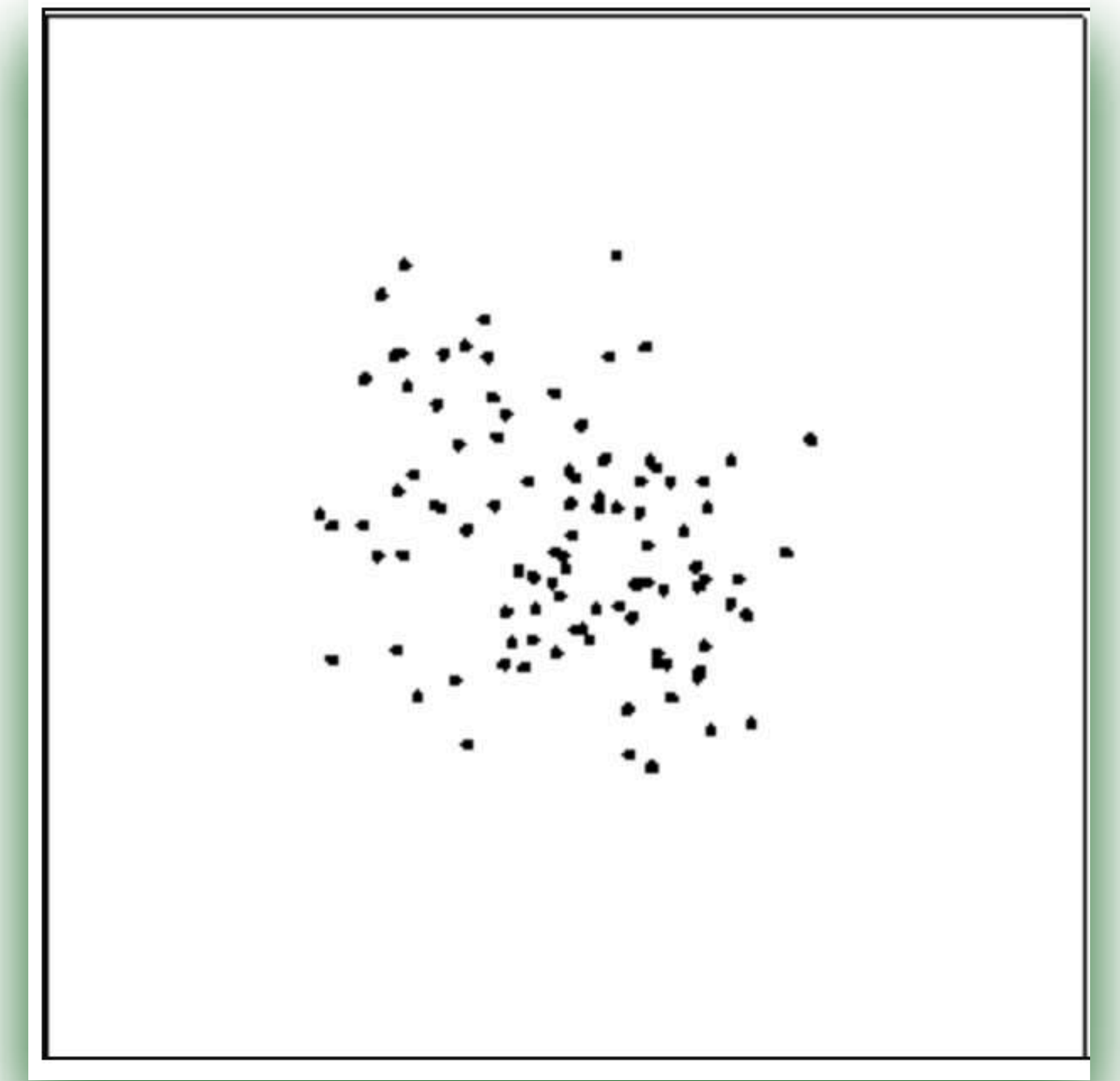
Estimate r



r = +0.8



r = -0.7



r = -0.3