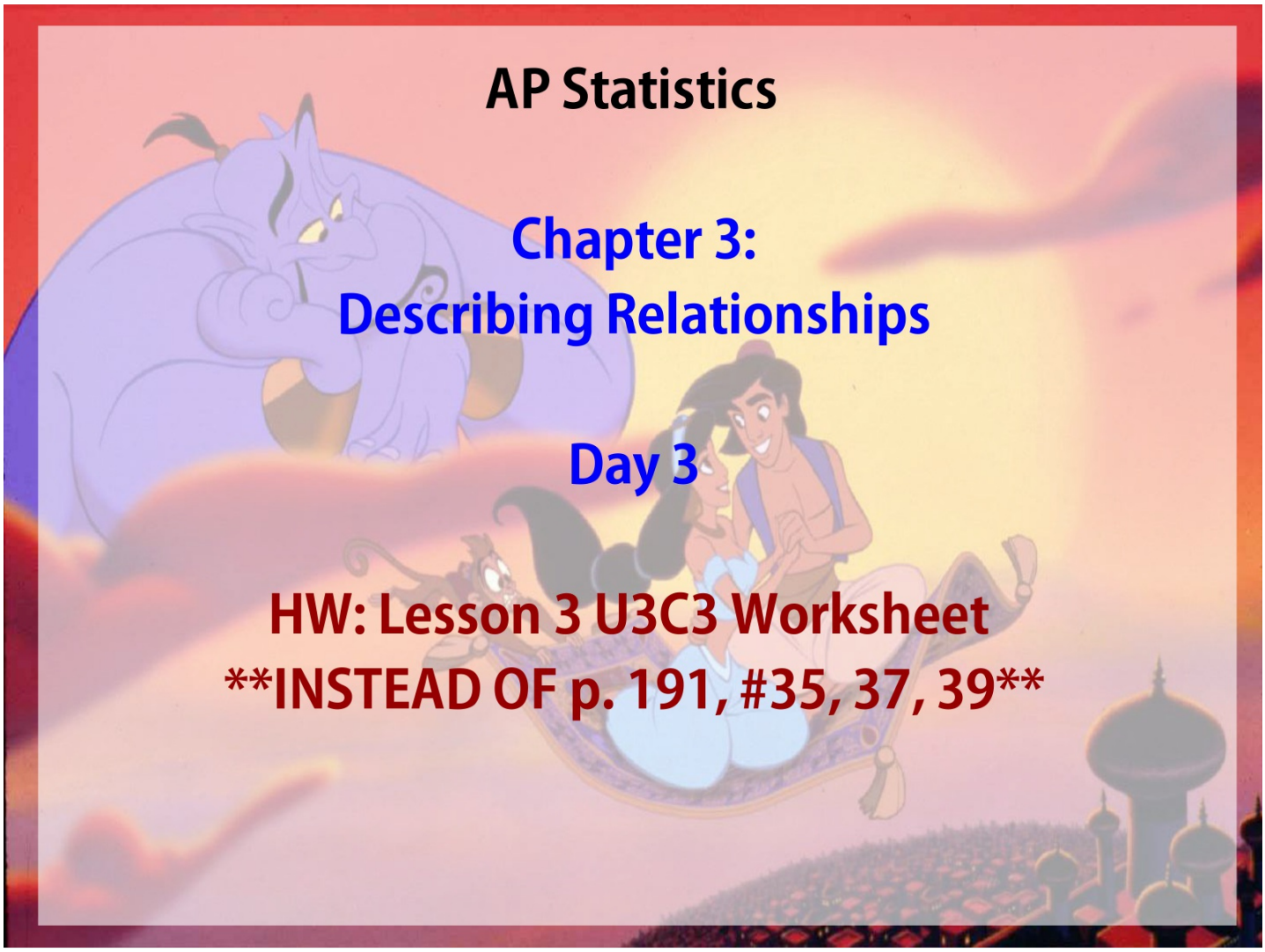
Correlation Coefficient Activity

At your table, take 10 minutes to finish up the activity from last class.

In your group, make sure to answer the two questions about the correlation coefficient and the association between the two variables.

Be prepared to share your answers with the whole class.



AP Statistics

Chapter 3:
Describing Relationships

Day 3

HW: Lesson 3 U3C3 Worksheet
****INSTEAD OF p. 191, #35, 37, 39****

p. 160-163, #15, 17, 18, 21, 25, 27-32

15.
(a) $r = 0.9$
(b) $r = 0$
(c) $r = 0.7$
(d) $r = -0.3$
(e) $r = -0.9$

17.
(a) Gender is a categorical variable and the correlation coefficient r measures the strength of linear association for two quantitative variables.

(b) The largest possible value of the correlation coefficient r is 1.

(c) The correlation coefficient r has no units.

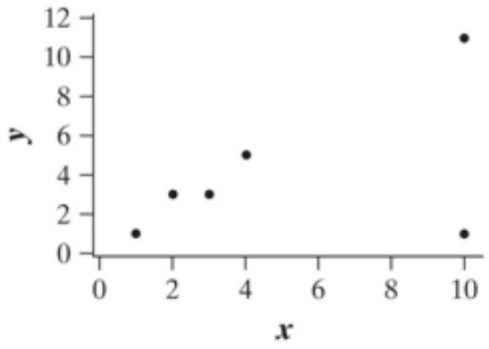
18. The paper's report is wrong because the correlation $(r = 0.0)$ is interpreted incorrectly. The author incorrectly suggests that a correlation of zero indicates a negative association between research productivity and teaching rating. The psychologist meant that there is no linear association between research productivity and teaching rating. In other words, knowledge of a professor's research productivity will not help you predict her teaching rating.

21.

(a) There is a strong positive linear association. High-calorie hot dogs tend to be high in salt, and low-calorie hot dogs tend to have lower sodium.

(b) It would tend to decrease the correlation. An outlier generally in line with the bulk of the data will tend to increase the correlation.

25. The scatterplot is shown below. The one unusual point (10, 1) is responsible for reducing the correlation. Outliers tend to have fairly strong effects on correlation; the effect is very strong here because there are only six observations.



27. a 28. e 29. d 30. b 31. c 32. d

Least-Square Regression Line

A least-squares regression line is a model for predicting the value of a dependent variable *Y*, based on the value of an independent variable *X*.

This method calculates the best-fit line by minimizing the sum of the squares of the differences between the observed values and the predicted values from the line (residuals).

$$\sum (ob - pred.)^2$$

Least-Square Regression Line

$$\hat{y} = b_0 + b_1 x$$

- The symbol $\hat{y}$ is read "y-hat" and is the predicted value of the dependent variable.
- b_0 is the y-intercept of the line and b_1 is the slope.
- *x* is the independent variable.
- The line of best fit will always pass through the point $(\bar{x}, \bar{y})$, the mean of the *x*-values and *y*-values.

Notation Alert!

The formulas provided on the AP Statistics Exam use b_0 for the intercept and b_1 for the slope.

Graphing calculators and some textbooks use a for the intercept and b for the slope.

You must be able to use these interchangeably.

$\hat{y} = a + bx$

$\hat{y} = b_0 + b_1 x$

Least-Square Regression Line

The slope and intercepts are calculated by the following formulas:

$$b_1 = r\frac{s_y}{s_x}$$ SD

$$b_0 = \bar{y} - b_1\bar{x}$$

- Where $\bar{x}$ and $\bar{y}$ are the means of the explanatory and response variables.
- s_x and s_y are the standard deviations.

Example

A random sample of 10 office assistants hired within the last six months was selected from a large company. Each assistant's experience (in months) at the time of hire and annual starting salary (in thousands of dollars) were recorded in the table below.

Experience (in months)	Starting Salary (in $1,000 dollars)
5	28
12	34
2	24
0	19
2	24
10	32
5	25
1	20
10	29
5	23

Example

Compute the slope for the line of best fit. Interpret it.

$b_1 = r\frac{s_y}{s_x}$

$b_1 = .935\left(\frac{4.894}{4.185}\right) = 1.093$

For every additional 1 month of experience, the salary should increase by 1.093 thousand dollars.

Compute the *y*-intercept for the line of best fit. Interpret it.

$b_0 = \bar{y} - b_1\bar{x}$

$b_0 = 25.8 - 1.093(5.2) = 20.116$

A person with no experience should earn a salary of 20.116 thousand dollars.

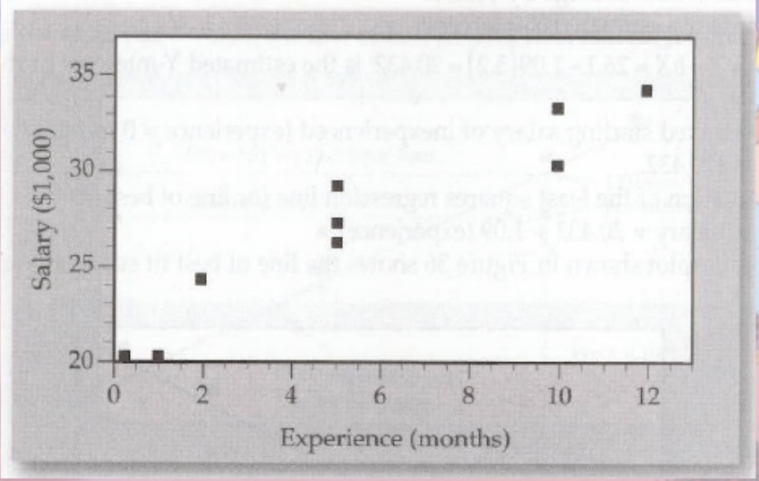
Experience (in months)	Starting Salary (in $1,000 dollars)
5	28
12	34
2	24
0	19
2	24
10	32
5	25
1	20
10	29
5	23

Example

Find the equation of the least square regression line to estimate starting salary using experience.

$\hat{y} = b_0 + b_1 x$ $y = 20.116 + 1.093x$

Plot the line of best fit in a scatterplot using your calculator. $(\bar{x}, \bar{y})$ $(5.2, 25.8)$



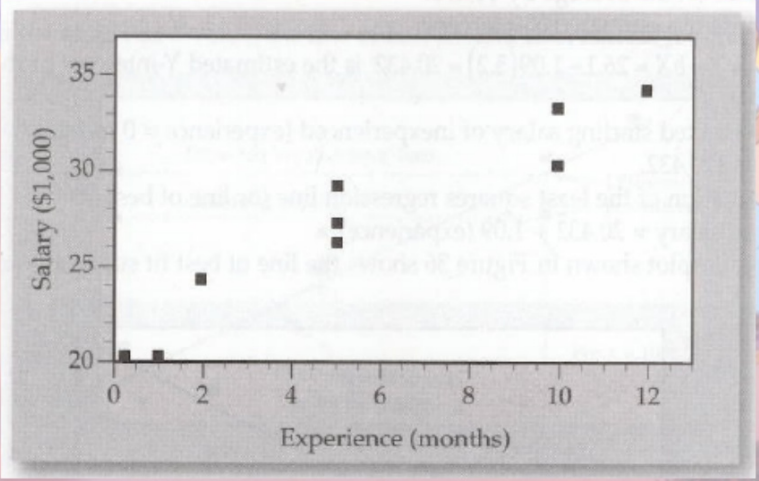
Experience (in months)	Starting Salary (in $1,000 dollars)
5	28
12	34
2	24
0	19
2	24
10	32
5	25
1	20
10	29
5	23

Example

Predict the starting salary for an office assistant with six months of prior experience.

$\hat{y} = 20.116 + 1.093(6)$

$\hat{y} = 26.674$



Experience (in months)	Starting Salary (in $1,000 dollars)
5	28
12	34
2	24
0	19
2	24
10	32
5	25
1	20
10	29
5	23

Caution!!!

We would not want to use this model to predict a starting salary for an office assistant whose experience was outside the range used to find the LSRL.

If the most experienced office assistant has worked for 12 months, we would not want to use the equation to predict the the starting salary of an office assistant who had 24 months experience.

Experience (in months)	Starting Salary (in $1,000 dollars)
5	28
12	34
2	24
0	19
2	24
10	32
5	25
1	20
10	29
5	23

Note

Using this model to predict a salary that is above or below the range is called extrapolation. ✳

Responses predicted this way could be unreasonable.

Salaries may not necessarily increase at the same rate for people who have many years of experience.

Be careful of extrapolation in any model!