

1 Chapter 8 review exercises

1 (d) is a scatter diagram for the data. First, the point of averages—in this case, $(100, 100)$ —must be in the center of the cloud, ruling out (a) and (c). Second, the empirical rule tells us that 68% of the data should fall within 1 SD (e.g. 15 IQ points) of the mean along each axis, while 95% of the data on each axis should be within 2 SDs (e.g. 30) of the mean along each axis. We can rule out (b) because all of the data falls within about 15 IQ points of the point of averages. Hence, a scatter diagram that is centered at $(100, 100)$ and has 95% of the data between 70 and 130 IQ points along each axis is (d).

3 The correlation between the husbands' and their wives' heights would be $r = 1$, which is a perfect correlation where all the points lie on a line. To see this, note that an increase in the height of the husband leads to an increase in the height of the wife. We can easily calculate this relationship as $y = x - .08x$, or $y = .92x$, where x represents the height of the husband and y represents the height of the wife. Because all points lie on this same upward-sloping line, the correlation coefficient is 1.

9 Recall that

$$r = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y},$$

where

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

$$\sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}.$$

$$\text{cov}(x, y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).$$

(a) First we must calculate the averages for x and y by dividing the sums of x and y by n , which is 10 in this example. Then we calculate

the squared deviations of each observation from the mean for each variable (columns 5 and 6 below). Then we use the sum of the squared deviations in the equations for standard deviation listed above (i.e. note that the mean for sum of the squared deviation is the variance, and the standard deviation is a square root of the variance), and we use the sum of the products of the deviations (column 7 below) to calculate the covariance (e.g. the mean for the sum of the products of the deviations is the covariance).

(1) x	(2) y	(3) $x - \bar{x}$	(4) $y - \bar{y}$	(5) $(x - \bar{x})^2$	(6) $(y - \bar{y})^2$	(7) $(x - \bar{x})(y - \bar{y})$
1	5	-1	2	1	4	-2
1	3	-1	0	1	0	0
1	5	-1	2	1	4	-2
1	7	-1	4	1	16	-4
2	3	0	0	0	0	0
2	3	0	0	0	0	0
2	1	0	-2	0	4	0
3	1	1	-2	1	4	-2
3	1	1	-2	1	4	-2
4	1	2	-2	4	4	-4
$\sum_{i=1}^n x_i$ = 20	$\sum_{i=1}^n y_i$ = 30			$\sum_{i=1}^n (x_i - \bar{x})^2$ = 10	$\sum_{i=1}^n (y_i - \bar{y})^2$ = 40	$\sum_{i=1}^n (x - \bar{x})(y - \bar{y})$ = -16
\bar{x} = $\frac{20}{10} = 2$	\bar{y} = $\frac{30}{10} = 3$			σ_x^2 = $\frac{10}{10} = 1$	σ_y^2 = $\frac{40}{10} = 4$	$\text{cov}(x, y)$ = $\frac{-16}{10} = -1.6$
				$\sigma_x = \sqrt{\sigma_x^2} = 1$	$\sigma_y = \sqrt{\sigma_y^2} = 2$	

From $\sigma_x = \sqrt{\frac{10}{10}} = 1$, $\sigma_y = \sqrt{\frac{40}{10}} = 2$, and $\text{cov}(x, y) = -\frac{16}{10} = -1.6$, we can find that $r = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y} = \frac{-1.6}{1 \cdot 2} = -0.8$. This answer can also be computed by following the procedure in the book on pages 132-133.

(b) See steps in part (a).

(1)	(2)	(3)	(4)	(5)	(6)	(7)
x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	1	-1	-1	1	1	1
1	2	-1	0	1	0	0
1	1	-1	-1	1	1	1
1	3	-1	1	1	1	-1
2	1	0	-1	0	1	0
2	4	0	2	0	4	0
2	1	0	-1	0	1	0
3	2	1	0	1	0	0
3	2	1	0	1	0	0
4	3	2	1	4	1	2
$\sum_{i=1}^n x_i$ = 20	$\sum_{i=1}^n y_i$ = 20			$\sum_{i=1}^n (x_i - \bar{x})^2$ = 10	$\sum_{i=1}^n (y_i - \bar{y})^2$ = 10	$\sum_{i=1}^n (x - \bar{x})(y - \bar{y})$ = 3
\bar{x} = $\frac{20}{10} = 2$	\bar{y} = $\frac{20}{10} = 2$			σ_x^2 = $\frac{10}{10} = 1$	σ_y^2 = $\frac{10}{10} = 1$	$\text{cov}(x, y)$ = $\frac{3}{10} = 0.3$
				$\sigma_x = \sqrt{\sigma_x^2} = 1$	$\sigma_y = \sqrt{\sigma_y^2} = 1$	

From $\sigma_x = 1$, $\sigma_y = 1$, and $\text{cov}(x, y) = 0.3$, we can find that $r = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y} = \frac{0.3}{1 \cdot 1} = 0.3$.

(c)

(1)	(2)	(3)	(4)	(5)	(6)	(7)
x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-1	-2	1	4	2
1	2	-1	-2	1	4	2
1	2	-1	-2	1	4	2
1	2	-1	-2	1	4	2
2	4	0	0	0	0	0
2	4	0	0	0	0	0
2	4	0	0	0	0	0
3	6	1	2	1	4	2
3	6	1	2	1	4	2
4	8	2	4	4	16	8
$\sum_{i=1}^n x_i$ = 20	$\sum_{i=1}^n y_i$ = 40			$\sum_{i=1}^n (x_i - \bar{x})^2$ = 10	$\sum_{i=1}^n (y_i - \bar{y})^2$ = 40	$\sum_{i=1}^n (x - \bar{x})(y - \bar{y})$ = 20
\bar{x} = $\frac{20}{10} = 2$	\bar{y} = $\frac{40}{10} = 4$			σ_x^2 = $\frac{10}{10} = 1$	σ_y^2 = $\frac{40}{10} = 4$	$\text{cov}(x, y)$ = $\frac{20}{10} = 2$
				$\sigma_x = \sqrt{\sigma_x^2} = 1$	$\sigma_y = \sqrt{\sigma_y^2} = 2$	

From $\sigma_x = 1$, and $\sigma_y = 2$, and $\text{cov}(x, y) = 2$, we can find that $r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{2}{1 \cdot 2} = 1$.

- # 11 The correct answer is $r = -1$. Note that one more write answer means one less wrong answer because the sum of “number of right answers” and “number of wrong answers” is a fixed number, e.g. 10. That is, if we let x be the number of right answers and y be the number of wrong answers, then $x + y = 10$, which is equivalent to writing $y = 10 - x$. Since any increase in x must lead to a decrease in y by the same amount (i.e. the variables x and y move in opposite directions) and all observations fall on the line $y = 10 - x$, there is a perfect negative correlation between the two variables. Therefore, the correlation coefficient is $r = -1$.

2 Chapter 9 review exercises

- # 2 (a) False. With $r < 0$, the variables tend to move in opposite directions. So, on average, below average values of one variable are associated with *above* average values of the other.
- (b) False. What determines the sign of the correlation coefficient is whether the variables tend to move together in the same direction or

not, not whether one is usually less or greater than the other. Consider two variables x and y where $y = 0.5x$. Note that y is always less than x , while there is a perfectly positive correlation between x and y .

- # 4 The correlation coefficient for the combined data is somewhat higher. To see this, you can sketch a scatter diagram that represents the relationship between height and weight for men and women. (See the guidelines for drawing an oval about the cloud of points on page 125 in FPP.) Let height be on the x-axis and weight be on y-axis. (You can interchange the two variables.) First, draw the point of averages which locates the center of the cloud i.e., (average height, average weight) = (70, 144) for men, and (64, 120) for women. Second, sketch the spread of the data by using the information that the correlation coefficient between height and weight is about 0.6 for both men and women. The same correlation coefficients imply that the spread of the scatter diagram is almost identical for men and women. That is, scatter diagrams for men and women are very similarly shaped ovals but the points of averages are differently located, one at (70, 144) for men, and the other at (64, 120) for women.

We can see that the two scatter diagrams are in fact colinear. Since the SD_{weight} and SD_{height} are the same for men and women, the slope of the SD line, and therefore the slope of the scatter diagram for men and women is the same.

Since these two similarly shaped ovals (i.e. scatter diagrams for men and women) lie on the same SD line, the scatter diagram of the combined data will be a more narrow oval (i.e. data is now less dispersed around the SD line) than either one for men or for women. If the oval is more narrow, then, the correlation is somewhat higher.

- # 7 No. In this question, the correlation coefficient $r = 0.5$ is computed based on percentages across counties. Hence, $r = 0.5$ indicates some positive association between the percentage of native-born Americans and the percentage of the vote for Johnson “across counties” on average but it does not necessarily indicate association “across individuals”.

Recall that using the correlation for aggregate level units such as counties to infer the relationship for the individuals (e.g. the correlation between being a native-born American and her support for Johnson) is likely to lead to overstating the strength of an association across individuals. (See pages 148-149 for Ecological correlations.)

- # 8 The statement that “as you get older, you become less educated” is false. Note that the years of education that a person has received cannot fall as they get older. A sensible explanation for the pattern that we see is that on average in former times women attended school for less years than they did from the 1970s on. The fact that many more women now go on to study in college than used to be the case in the 1950s and 1960s explains the pattern we see in the data.
- # 10 (a) True. Since there is a negative correlation between the two variables (i.e. $r = -0.86$), it is true that test scores tend to be lower in the states where a higher percentage of the students take the test, on average. One plausible explanation for such a negative correlation is as follows: in states where a lower percentage of students take the test, the best students are likely to be the ones who take it. A higher percentage of the students taking the test is likely to mean a higher variation in the SAT scores with more students earning low scores, dragging down the average.
- (b) False. The data does not indicate any information about teaching performance. Although it may be true that the schools in Iowa are doing a better job at teaching math than the schools in Connecticut, we do not have information to conclude this. Also, there are other possible explanations that can explain the difference in the average score between the two states.
- # 12 (a) Note that for a given years of schooling completed for husbands, years of schooling completed for wives vary (i.e. there are multiple levels of wife’s education for a given level of husband’s education), making a vertical stripe. Likewise, for a given years of schooling completed for wives, years of schooling completed for husbands vary (i.e. there are multiple levels of husband’s education for a given level of wife’s education), making a horizontal stripe.
- (b) When there are many couples having the same combination of years of schooling, meaning that many data points are on top of each other, we see fewer dots than the number of couples. For example, when the number of combinations of schooling years between husband and wife for 530 couples is 104, we only see 104 data points.
- (c) (i) corresponds to C. “Wife’s educational level” is shown on the y-axis, and the horizontal grey stripe highlights all those combinations where the wife has had exactly 16 years of schooling.

(ii) is the one left over.

(iii) corresponds to B. The shaded area covers all combinations where the husband has had more than 16 years of schooling (indicated on the x-axis).

(iv) corresponds to A. The shaded area includes all those combinations with $(12, y)$, where $y < 12$.