

1. Chapter 8 review exercises

- # 1 The scatter diagram that best matches the information given is (d). Note that the point of averages (in this case, (100,100)) must be in the center of the cloud. (a) and (c) therefore drop out as viable candidates. In (b), the point of averages is in the center of the circle, yet the circle has a very small radius (which is approximately ± 1 standard deviation (i.e., ± 15)). Thus, we choose (d) as best describing the data, as its center is the point of averages, and the radius of the circle is about ± 2 standard deviations for both “Wife” and “Husband,” so that most data will fall into it.
- # 3 In this case, the correlation would be perfect, i.e., $r = 1$. First note that we have a positive r , as for an increase in the height of the husband, that of the wife would also go up. To intuitively see why we get $r = 1$, think about the scatter diagram that we would obtain in this situation: All points would simply lie on the same line. To see why we don’t get $r = 0.92 (= 1 - 0.08)$, imagine that men always marry women who are exactly 8% taller. Would you say in this situation that $r = 1.08$? (Hopefully not, as by definition $-1 \leq r \leq 1$.)
- # 9 The complete work is shown here only for (a) and follows the approach described in FPP on pages 132-133. To obtain the results for (b) and (c), follow the same steps. (a) *Step 1*. We first have to convert the x-values into standard units. For this, we have to compute both the average and the standard deviation of x .

Average:

$$\begin{aligned}\mu &= \frac{1 + 1 + 1 + 1 + 2 + 2 + 2 + 3 + 3 + 4}{10} \\ &= \frac{20}{10} \\ &= 2\end{aligned}$$

Standard deviation:

Note that we have to use the deviations from the average in our calculation of σ (i.e., $-1 = 1 - 2$ instead of 1, $0 = 2 - 2$ instead of 2, $1 = 3 - 2$ instead of 3, and $2 = 4 - 2$ instead of 4):

$$\begin{aligned}
\sigma &= \sqrt{\frac{(1-2)^2 + (-1)^2 + (-1)^2 + (-1)^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 2^2}{\text{number of entries}}} \\
&= \sqrt{\frac{1+1+1+1+0+0+0+1+1+4}{10}} \\
&= \sqrt{\frac{10}{10}} \\
&= \sqrt{1} \\
&= 1
\end{aligned}$$

We can now convert the x-values into standard units, using the following formula:

$$x_{\text{standard}} = \frac{x-\mu}{\sigma}.$$

$$\text{Thus: } \frac{1-2}{1} = -1, \quad \frac{2-2}{1} = 0, \quad \frac{3-2}{1} = 1, \quad \frac{4-2}{1} = 2.$$

Step 2. The same procedure has to be repeated for the y-values.

Average:

$$\begin{aligned}
\mu &= \frac{5+3+5+7+3+3+1+1+1+1}{10} \\
&= \frac{30}{10} = 3
\end{aligned}$$

Standard deviation:

→ Don't forget to subtract the average from each y-value before plugging it into the formula!

$$\begin{aligned}
\sigma &= \sqrt{\frac{(5-3)^2 + 0^2 + 2^2 + 4^2 + 0^2 + 0^2 + (-2)^2 + (-2)^2 + (-2)^2 + (-2)^2}{\text{number of entries}}} \\
&= \sqrt{\frac{4+0+4+16+0+0+4+4+4+4}{10}} \\
&= \sqrt{\frac{40}{10}} \\
&= \sqrt{4} \\
&= 2.
\end{aligned}$$

Converting the y-values into standard units yields:

$$\frac{1-3}{2} = -1, \quad \frac{3-3}{2} = 0, \quad \frac{5-3}{2} = 1, \quad \frac{7-3}{2} = 2.$$

Step 3. Next, compute the product for each pair of x in standard units and y in standard units.

x	y	x in standard units	y in standard units	Product
1	5	-1	1	-1
1	3	-1	0	0
1	5	-1	1	-1
1	7	-1	2	-2
2	3	0	0	0
2	3	0	0	0
2	1	0	-1	0
3	1	1	-1	-1
3	1	1	-1	-1
4	1	2	-1	-2

Step 4. Take the average of the products:

$$\begin{aligned} r &= \frac{(-1) + 0 + (-1) + (-2) + 0 + 0 + 0 + (-1) + (-1) + (-2)}{10} \\ &= \frac{-8}{10} = -0.8 \end{aligned}$$

So, after quite a bit of arithmetic (and typesetting), we find that the correlation coefficient is -0.8. (As a quick check, if you look at the original data, you see that for increasing x , y tends to decrease, so we certainly got the sign right.)

(b) *Steps 1 & 2.* Here are the means and standard deviations: $\mu_x = 2$, $\sigma_x = 1$, $\mu_y = 2$, $\sigma_y = 1$. Note that the x -values are the same as in (a) and that therefore the mean and standard deviation must be the same, and that the y -values are the same as the x -values “shuffled around,” which means that the mean and standard deviation of y will be the same as that of x .

Step 3. The table now looks like this:

x	y	x in standard units	y in standard units	Product
1	1	-1	-1	1
1	2	-1	0	0
1	1	-1	-1	1
1	3	-1	1	-1
2	1	0	-1	0
2	4	0	2	0
2	1	0	-1	0
3	2	1	0	0
3	2	1	0	0
4	3	2	1	2

Step 4. Again, take the average of the products:

$$\begin{aligned}
 r &= \frac{(1) + 0 + (1) + (-1) + 0 + 0 + 0 + (0) + (0) + (2)}{10} \\
 &= \frac{3}{10} = 0.3
 \end{aligned}$$

Thus, the correlation coefficient for the data shown in (b) is 0.3.

(c) *Steps 1 & 2.* Again, the x -values are the same as in (a), so we get $\mu_x = 2$ and $\sigma_x = 1$. The y -values are given by $y = 2x$, so we know from chapter 5 in FPP (“multiplying by a constant”) that $\mu_y = 2 \cdot 2 = 4$ and $\sigma_y = 1 \cdot 2 = 2$.

Step 3. Based on the previous steps, the table looks like this:

x	y	x in standard units	y in standard units	Product
1	2	-1	-1	1
1	2	-1	-1	1
1	2	-1	-1	1
1	2	-1	-1	1
2	4	0	0	0
2	4	0	0	0
2	4	0	0	0
3	6	1	1	1
3	6	1	1	1
4	8	2	2	4

Step 4. This time the average of the products is:

$$\begin{aligned} r &= \frac{1 + 1 + 1 + 1 + 0 + 0 + 0 + 1 + 1 + 4}{10} \\ &= \frac{10}{10} = 1 \end{aligned}$$

So, in this case we have a perfect positive correlation with $r = 1$.

(a) # 11

The correct answer is $r = -1$. Perhaps the easiest way to think about this question is to write down the following equation:

$$\text{total number of answers} = \text{right answers} + \text{wrong answers}$$

Alternatively, we can express this using the variables x for “number of right answers” and y for “number of wrong answers.”

$$10 = x + y$$

What happens now to y when x changes? Say, x goes up from 6.4 to 8. Then y must go down to $10 - 8 = 2$ (do you see why?). In other words, any change in x must lead to exactly the same change in y , only in the opposite direction. So, given that $\sigma_x = \sigma_y = 2$, we conclude that there is a perfectly negative correlation between the two variables, i.e., 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.

4 The best answer here is “somewhat higher.” By combining the two data sets we get a denser cloud of points around the center of the scatter, thus increasing the correlation. This can be seen if one follows the guidelines in drawing an oval about the cloud of points on page 125 in FPP given the data for this problem. The ovals that are produced for this data, are essentially lying on the same line. If we drew a line that best represented the relationship for men, women, and then the entire dataset as a

whole, the three would be roughly co-linear. So, from a graphical perspective, because the set of plotted points for both men and women are less dispersed about the line that we could draw describing the relationship for everyone relative to the range of data considered, the correlation coefficient would increase. Imagine increasing the size of either the men's or women's oval. If we scaled it up (keeping a fixed perspective) to the same length of our oval for the combined dataset, it would be much fatter than the oval for the combined dataset. If the oval is more narrow, then, the correlation is somewhat higher (i.e., better predictions).

In fact, if we simulate these distributions and plot the points in each case we can see this clear (see Figure 1 at the end of the document).

- # 7 No, the computed correlation of 0.5 is most likely not a fair measure of the extent to which Johnson received support from non-immigrants. The main problem here is that the calculation is based on county averages, which can be misleading because they tend to overstate the strength of an association. For more details on this problem, refer back to the section on “ecological correlations” on pp. 148-149 in FPP.
- # 8 The statement that “as you get older, you become less educated” is false. The years of education that a person has received cannot of course fall as they get older, so the only sensible explanation for the pattern that we see is that on average in earlier times women attended school for less years than they did from the 1970s on. In a nutshell, 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. We know that with an $r = -0.86$ there is a negative relation between the two variables. We also know that we can interchange the two variables without affecting the correlation coefficient, so we can state, “The higher the percentage of high-school seniors in a state who took the Math SAT, the lower the score.” The best explanation for this pattern is that in states where the percentage of those who take the Math SAT is low, only those with the highest motivation and best math skills decide to (voluntarily) take the Math SAT, while in other states the test might be compulsory and thus those high-school seniors with relatively poor math skills will take the test, too, dragging down

the state's average. (In the social sciences, the former behavior is known as "self-selection" and commonplace in educational and labor market settings.)

(b) False. The data do not conclusively show that the schools in Iowa do a better job at teaching math than the ones in Connecticut. One plausible explanation for the observed pattern is that all or at least some students from Connecticut are required to take the test because their school requires them to do so, while in Iowa there are no such requirements, which leads to the fact that only Iowa's best math students take the Math SAT (again, the problem in this case is self-selection), while in Connecticut students with average or even poor math skills do so. It might also be the case that the schools in Connecticut are doing as good a job as the ones in Iowa and that the scores differ even absent compulsory participation because the schools in Connecticut had worse students to start with. Only a study that measures the improvement of the same students, say from 6th to 8th grade, would allow us to assert that the schools in one state do a better job at teaching math than those in the other.

12 (a) The points make vertical and horizontal stripes because there is no such thing as 11.64 years of schooling completed. Years of schooling completed are always full years and therefore they "make stripes."

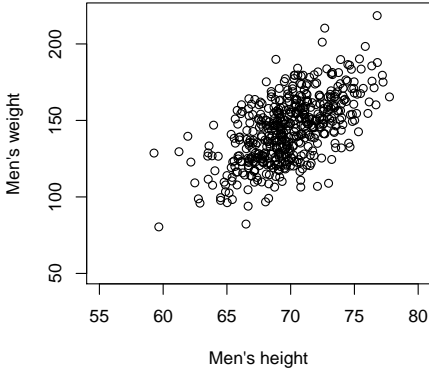
(b) The fact that we can count only 83 dots can simply be explained by considering that many couples have the same combination of years of schooling, e.g., (12,12) or (16,16). These dots are then on top of each other, but we can't see this pattern in a two-dimensional graph, giving rise to the impression that there are only 83 combinations.

(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.

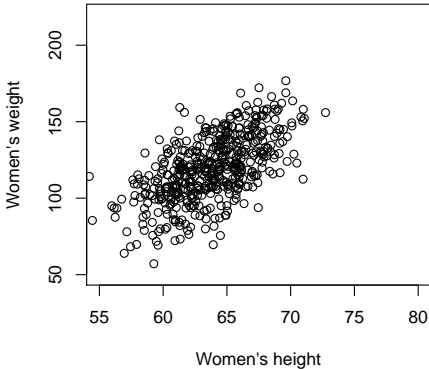
(ii) is the one left over.

(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).

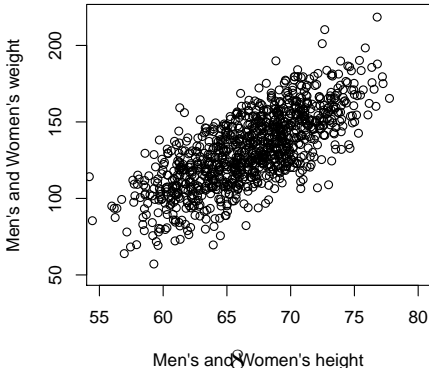
(iv) corresponds to A. The shaded area includes all those combinations with (12, y), where $y < 12$.



(a) Scatterplot for Men



(b) Scatterplot for Women



(c) Scatterplot of Combined Data

Figure 1: Correlations are: 0.6 (men), 0.61 (women), 0.73 (both).