Solutions to Worktext Exercises

Chapter 14

Visualizing Bivariate Data Analysis

Basic Learning Exercises

1. X is the *percent* of the state's population age 65 or above, and Y is the number of deaths from heart disease *per 100,000 population*. Yes, the variables are corrected for population size. Populous states (e.g., California) have more persons over 65 (X) and more heart deaths (Y) and conversely for small states (e.g., South Dakota). This would induce a high correlation without measuring the effect we are studying.

2. The scatter plot shows a strong positive relationship with somewhat unusual values on the right (Florida) and left (Alaska).

3. In percent of population aged 65 or over, Florida is the highest (18.5) and Alaska is the lowest (5.3). In death rate from heart disease, West Virginia is the highest (378.9) and Alaska the lowest (90.6). Alaska is a frontier state that attracts young people who probably do not have heart disease. Florida is a sunbelt retirement state. West Virginia's high heart disease rate may be related to poor economic conditions.

4. The correlation is 0.801, which suggests a direct (positive) association. The p-value is very small (below 0.01), which indicates a significant correlation.

5. The correlation coefficient falls to 0.5 or lower, although its p-value is still below 0.01. A single data point can be influential if it is in a highly-levered position.

6. X = Percent of population age 65 and over   Y = Death rate from heart disease
   Mean 12.69   Median 12.7   Mean 274.5   Median 280.40
   For both variables, the median falls near the middle of the box and mean and median are similar. For both variables, the lower tail of each box plot is slightly longer than the upper tail, but the difference is not striking. Neither variable is strongly skewed.

7. When the grid is chosen to form equal intervals, we get 14 states on the left and 36 on the right. When the grid cutpoints are chosen to equalize the group frequencies, we have 25 states in each group, which is an advantage. If the data were uniformly distributed, either option would yield the same results.

8. With 3 columns the left box plot disappears (only two points) and with 4 columns both the left and right column box plots disappear (1 and 3 points respectively). There are not enough data points in the end categories to find the quartiles.

9. The Equal Frequency option works well for 2, 3, or 4 columns. Although its cut points look strange, the Equal Frequencies option works better when there are sparse data in the tails. The Equal Width option is more intuitive, but fails when sparse tails exist.

10. The median heart disease rate is higher for the states with more population aged 65 and over. This same conclusion holds consistently as we increase the number of columns.

Intermediate Learning Exercises

11. Yes, it represents the data fairly well, although a few points are somewhat above (Mississippi) or below (Hawaii, Florida) the fitted regression line.

12. a) The estimated equation is Heart Death Rate = -29.05 + 23.92 (Pop 65 and Up %). This says that for every 1 percent increase in population aged 65 or over, a state may expect about 24 additional deaths per 100,000 from heart disease. b) A positive slope is believable (we expect more heart deaths in an older population). c) The intercept is not useful because no state has zero population over age 65. d) The slope might be a useful guide to planning health care in states with growing retiree populations.

13. The R^2 value (0.64) says that the model explains about 64 percent of the variation in Y (in other words, 36 percent of the variation in Y is unexplained).

14. a) The R^2 falls to under 0.25. b) The slope drops dramatically to under 15. c) A single data point can have a powerful effect on a fitted regression line.

15. The table is correct. It is difficult to count the data points within the grid because the markers sometimes sit on the grid line.

16. a) Yes, the row and column frequency totals are equal, as we requested. b) If the sample size were odd (say n = 51) the row and column totals couldn't be exactly equal. c) The categories were formed by finding the median of X and the median of Y and using them as cutpoints. "Low" would be any value below the median and "High" would be any value above the median. d) No, the cell frequencies are extremely unequal. The upper right and lower left cells have 19 states each, while the other two cells have only 6 states.

17. a) The observed and expected frequencies are very different. b) No, the variables are not independent. If they were, every cell would contain 12.5 states, but actually the cells contain either 6 or 19 states. The chi-square test statistic is large (13.520) and its p-value is small (less than 0.001) so we would reject the hypothesis that X and Y are independent. c) No, all expected frequencies exceed 5, so Cochran's Rule is not violated.

18. In this case the chi-square test is helpful because it is not so greatly affected by any extreme points that have such a strong impact on the regression. This is because all we are doing is classifying X and Y into categories (Low, High) using the median as the dividing point. The median is insensitive to extreme values.

19. a) The chi-square test statistic (20.141) has risen slightly. b) Its p-value is still low (less than 0.001) so the variables still are judged dependent. c) One expected frequency is below 5 (the upper right cell's expected frequency is 4.8). This is a slight violation of Cochran's Rule, though not a severe one (Cochran's Rule is only a rule of thumb).

20. No. If we increase the table size, all of the expected frequencies would drop below 5.

Advanced Learning Exercises

21.

Sample	1	2	3	4	5	6	7	8	9	10	Average
r	____	____	____	____	____	____	____	____	____	____	0.25
p-value	____	____	____	____	____	____	____	____	____	____	0.24

The average correlation is about 0.25 (range is –0.11 to 0.61). About 1 in 10 estimated correlations will be negative. There is great variation. About half the p-values will be below 0.10 (range is 0.01 to about 0.90). Most of the time, the scatter plot doesn't reveal the positive correlation clearly, and many scatter plots look quite random.

22.

Sample	1	2	3	4	5	6	7	8	9	10	Average
r	____	____	____	____	____	____	____	____	____	____	0.25
p-value	____	____	____	____	____	____	____	____	____	____	0.09

The average correlation is about 0.25 (range is from 0.01 to 0.44). It is very unlikely that any of the correlations will be negative. There is less variation than before. About 75 percent of the p-values will be below 0.10 (range is 0.01 to about 0.40), and about half of the p-values are shown as "p < 0.01". The scatter plots usually reveal the positive correlation, though it is hardly a strong relationship. A larger sample helps to see it. Larger sample size reduces variation in the estimates and reduces the p-values.

23. The slope is positive over 90 percent of the time. Yes, the regression line does help infer the positive correlation.

24. a) It doesn't take long to get a correlation coefficient close to 0.50. b) The first few drag-and-drop actions raise the correlation rapidly, but gains are harder to obtain as the exercise processes. c) When the correlation coefficient r reaches 0.99, the data are nearly a straight line. d) The t statistic for the slope is over 50 and the F statistic may exceed 3,000. e) No, such samples would be rare in most applications.

25. Although the intercept can be seen clearly, the data points are clustered farther to the right when this option is selected, so there is less detail on the scatter plot.

26. It is different because the regression model no longer has an intercept at all. Forcing a regression line through the origin is a major change in the model.

27. The scatter plot for the uniform looks similar to a normal distribution, but with less clustering in the middle. The right-skewed population tends to cause the sample data to cluster near the origin, while the left-skewed population causes the data to cluster at the right. Skewed populations might make a fitted regression line more vulnerable to extreme observations.