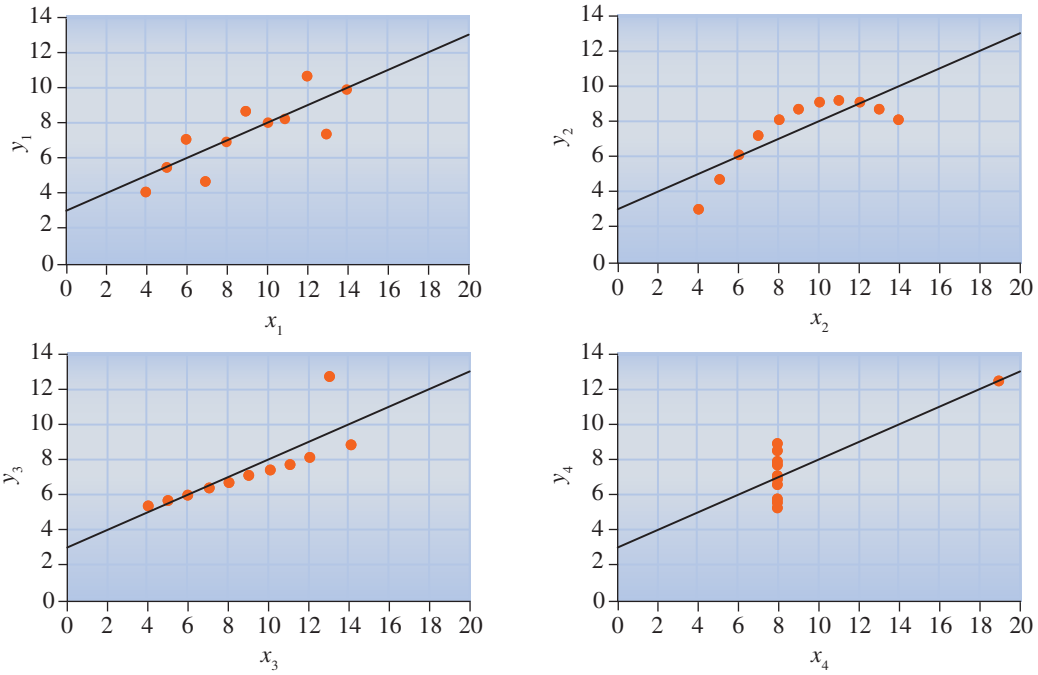


Figure 5.1.21

The Anscombe Quartet brings out an important data analytic principle. No matter what kind of numeric data is being analyzed, it is critically important to plot it before making any conclusions. For single variables, an analyst should at least be looking at histograms and box plots. When we begin to look for relationships in bivariate or multivariate data, creating scatterplots to display the relationship between the variables is more important than calculating summary measures of linear relationship, i.e., the correlation coefficient.

5.1 Exercises

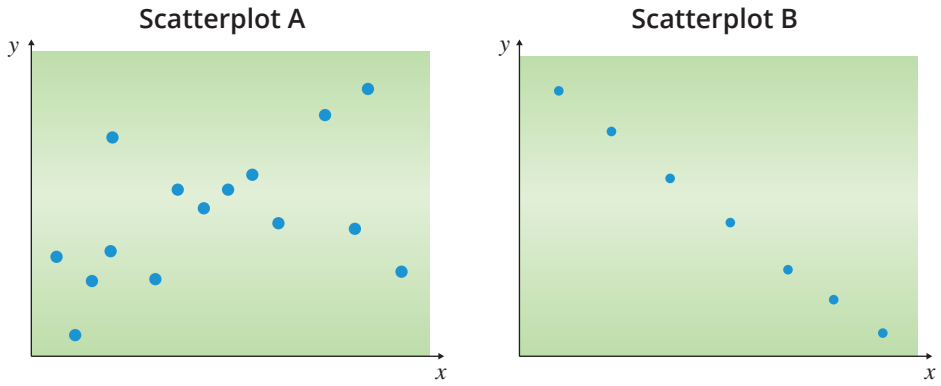
Basic Concepts

1. Give an example of a situation in which knowledge of a relationship between two variables is desired.
2. If a relationship can be uncovered, what are the potential benefits?
3. What is bivariate data? How is bivariate data different from univariate data?
4. What graphical tool is often used in the discovery of relationships?
5. What sort of questions should you ask when studying a graphical representation of bivariate data?
6. If bivariate data exhibits an inverse relationship, what does that mean?
7. How do you measure the exact relationship between two variables?
8. In what range is the value of r when bivariate data exhibits a positive relationship? A negative relationship?
9. If the value of r is small, does this always mean that no relationship exists? Explain.
10. What is confounding? Why is confounding a problem?

©HAWKES LEARNING

Exercises

11. Answer the following questions regarding the overall pattern of the data for each of the scatterplots.



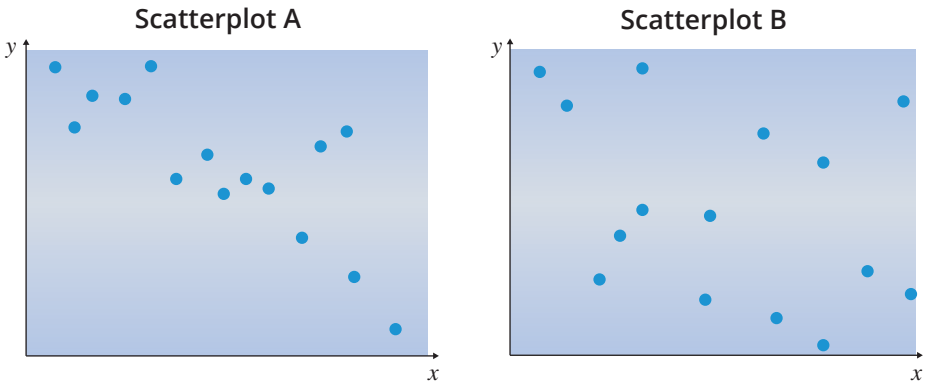
a. Does the pattern roughly follow a linear pattern?

b. Is the pattern upward sloping or downward sloping?

c. Are the data values tightly clustered in the pattern or widely dispersed?

d. Are there significant deviations from the pattern?

12. Answer the following questions regarding the overall pattern of the data for each of the scatterplots.



a. Does the pattern roughly follow a linear pattern?

b. Is the pattern upward sloping or downward sloping?

c. Are the data values tightly clustered in the pattern or widely dispersed?

d. Are there significant deviations from the pattern?

13. In the story of Moneyball, Billy Beane and Paul DePodesta initially calculated that 95 wins in a season was necessary for the Oakland A's to make it into the playoffs.[7] Next, they wanted to understand the relationship between overall season run differential (the number of runs scored in a season minus the number of runs allowed) and the number of games won in a season. Go to the companion website and download the Moneyball data set. Beane and DePodesta performed this analysis in 2002, so they only had data up to the year 2001. Subset the data set to only include data for the years 1962–2001.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Moneyball

©HAWKES LEARNING

a. Analyze the data collected for the study by answering the following questions:

i. What questions might Beane and DePodesta be trying to answer?

ii. Do the variables selected in the data set seem appropriate for answering the question?

iii. What biases or errors might be present in the data?

iv. What level of measurement (nominal, ordinal, interval, ratio) does each variable possess?

v. How is the data collected – through observation or controlled experiment?

b. Plot the data points for the variables *RD* (run differential) and *W* (wins) on a scatterplot.

c. Based on the scatterplot in part **b.**, answer the following questions regarding the overall pattern of the data.

i. Does the data roughly follow a linear pattern?

ii. Is the pattern upward sloping or downward sloping?

iii. Are the data values tightly clustered in the pattern or widely dispersed?

iv. Are there significant deviations from the pattern?

14. A pharmacist is interested in studying the relationship between the amount of a particular drug in the bloodstream (in nanograms per milliliter ng/ml) and reaction time (in seconds) of subjects taking the drug. Ten subjects are randomly selected and administered various doses of the drug. The reaction times (in seconds) are measured 15 minutes after the drug is administered with the following results.

Reaction Time of a Drug										
Amount of Drug (ng/ml)	1	2	3	4	5	6	7	8	9	10
Reaction Time (in sec)	0.5	0.7	0.6	0.7	0.8	0.8	0.9	0.6	0.9	1.0

a. Analyze the data collected for the study by answering the following questions.

i. Do the variables selected for measurement seem appropriate for the study of interest?

ii. What biases or errors might be present in the data? What confounding variables could impact the conclusion?

iii. What level of measurement (nominal, ordinal, interval, ratio) does the data possess?

iv. How is the data collected—through observation or controlled experiment?

b. Plot the data points on a scatterplot.

c. Based on the scatterplot in part **b.**, answer the following questions regarding the overall pattern of the data.

i. Does the pattern roughly follow a linear pattern?

©HAWKES LEARNING

ii. Is the pattern upward sloping or downward sloping?

iii. Are the data values tightly clustered in the pattern or widely dispersed?

iv. Are there significant deviations from the pattern?

15. Illustrate, using a scatterplot, a data set that would have a correlation coefficient of 1.

16. Illustrate, using a scatterplot, a data set that would have a correlation coefficient of -1.

17. Describe the relationships indicated by the correlation coefficients as tightly clustered in a positive linear fashion, tightly clustered in a negative linear fashion, loosely clustered in a positive linear fashion, loosely clustered in a negative linear fashion, or no linear relationship.

a. $r = 0.9$

b. $r = 0.5$

c. $r = -0.9$

d. $r = -0.5$

e. $r = 0$

f. What assumption did you make about the scatterplots in answering **a.** through **e.**?

18. Describe the relationships indicated by the correlation coefficients as tightly clustered in a positive linear fashion, tightly clustered in a negative linear fashion, loosely clustered in a positive linear fashion, loosely clustered in a negative linear fashion, or no linear relationship.

a. $r = 0.8$

b. $r = 0.4$

c. $r = -0.8$

d. $r = -0.4$

e. $r = 0.1$

f. What assumption did you make about the scatterplots in answering **a.** through **e**.?

19. Sometimes the following descriptions are assigned to the correlation coefficient:

$r = 0$	no linear relationship
$-0.5 < r < 0$	weak negative linear relationship
$0 < r < 0.5$	weak positive linear relationship
$-0.8 < r \leq -0.5$	moderate negative linear relationship
$0.5 \leq r < 0.8$	moderate positive linear relationship
$-1.0 < r \leq -0.8$	strong negative linear relationship
$0.8 \leq r < 1.0$	strong positive linear relationship
$r = 1$	exact positive linear relationship
$r = -1$	exact negative linear relationship

Describe the relationships indicated by the correlation coefficients below using the descriptions defined above.

a. $r = 0.9$

b. $r = 0.5$

c. $r = -0.9$

d. $r = -0.5$

©HAWKES LEARNING

e. $r = 0$

f. What assumption did you make about the scatterplots in answering **a.** through **e.**?

20. Describe the relationships indicated by the correlation coefficients below using the descriptions defined in problem 19 above.

a. $r = 0.8$

b. $r = 0.4$

c. $r = -0.8$

d. $r = -0.4$

e. $r = 0.1$

f. What assumption did you make about the scatterplots in answering **a.** through **e.**?

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > Super Bowl

21. Using the Super Bowl data set from the companion website, consider the following:

a. Construct a scatterplot using the variables *Winner_First Downs* and *Winner_Total Yards*.

b. Does there appear to be a negative or positive relationship between the variables?

c. Compute the correlation coefficient.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > OECD Better Life Index 2022

22. Using the OECD Better Life Index 2022 data set from the companion website, consider the following:

a. Construct a scatterplot using the variables *Safety Score* and *Life Satisfaction*.

b. Determine the correlation coefficient.

c. Describe the relationship indicated by the correlation coefficient and the scatterplot.

23. If the following variables have high negative linear correlations, is it reasonable to conclude that an increase in one variable causes a decrease in the other variable? Explain what could be causing this apparent relationship.

a. Time spent studying and number of followers on Instagram

b. Amount of caffeine consumed and academic performance

c. Amount of spending money and time to spend with friends

24. If the following variables have high positive linear correlations, can we conclude that an increase in one variable causes an increase in the other variable? Explain what could be causing this apparent relationship.

a. Sale of air conditioners and sale of tomatoes

b. Sale of greeting cards and sale of chocolates

c. Number of wrecks on a local highway and absenteeism from work

©HAWKES LEARNING