Chapter 5: Scatterplots and Correlation

1 Introduction

So far in the course, we have been analyzing univariate data (i.e. exploring each variables in a dataset separately). Now we will shift our focus towards bivariate analysis (the analysis of two variables to determine relationships between them). The most common way to display the relation between two quantitative variables is a scatterplot. In a scatterplot, usually the explanatory variable is on the x-axis and response variable is on the y-axis.

Example: According to a 2016 article published in the New York Times, when a survey was conducted of a representative sample of American voters, the types of foods these voters considered to be healthy didn't always align with the types of foods a panel of nutritionists identified as being healthy. A list was prepared of 52 food items, and the list was shared with 2000 registered voters and 672 nutrition experts. Figure below is a scatterplot that shows how the percentage of nutritionists who identified food items as being healthy compared with the percentage of American voters who identified these items as being healthy.

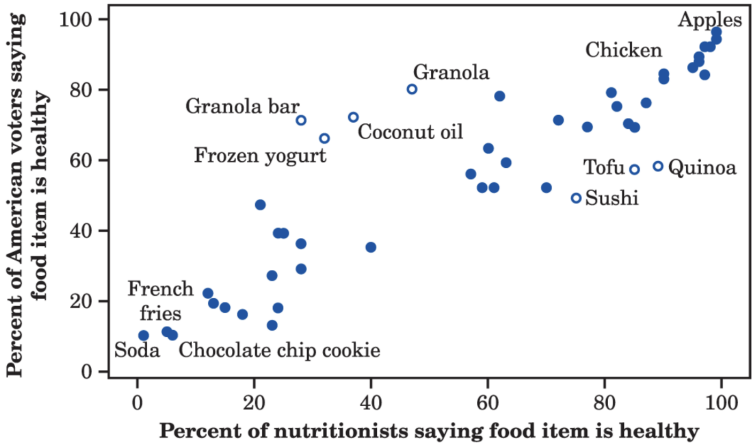


Figure 14.2 Scatterplot of percentage of American voters who think certain foods are healthy against percentage of nutritionists who think certain foods are healthy, Example 1.

- We want to see how the percentage of American voters who identify food items as healthy changes as the percentage of nutritionists who label these items as healthy changes, so we put the percentage of agreement among nutritionists (the explanatory variable) on the horizontal axis.
- We can see, for several of the food items, that as the percentage of agreement among nutritionists goes up, the percentage of agreement among American voters also goes up.

- We can also see that there isn't always agreement among nutritionists and American voters. For example, items such as frozen yogurt, granola, granola bars, and coconut oil were rated more often as healthy by American voters than by nutrition experts.
- Conversely, items like sushi, quinoa, and tofu were rated more often as healthy by nutrition experts than by American voters.

2 Scatterplots

Scatterplot shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

We always plot the explanatory variable, if there is one, on the horizontal axis (the x axis) of a scatterplot and response variable on the y-axis. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

Try it yourself: According to a 2021 article published in the Health Journal, when a survey was conducted with a representative sample of office workers, the types of activities these workers considered to be effective for stress relief didn't always align with the types of activities a group of mental health professionals identified as effective for stress relief. Attached below is the scatterplot. Interpret the scatterplot as in the example above.

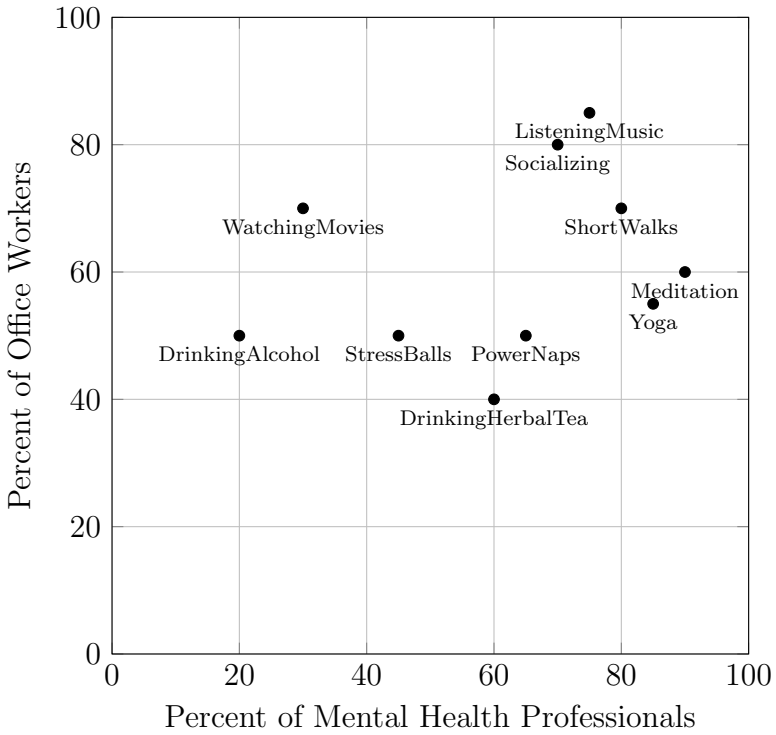


Figure: Scatterplot of percentage of mental health professionals who think certain activities are effective for stress relief against percentage of office workers who think certain activities are effective for stress relief.

3 Types of association

- Positively associated: Two variables are positively associated when one variable increases, other variable also increases. The scatterplot slopes upward as we move from left to right.

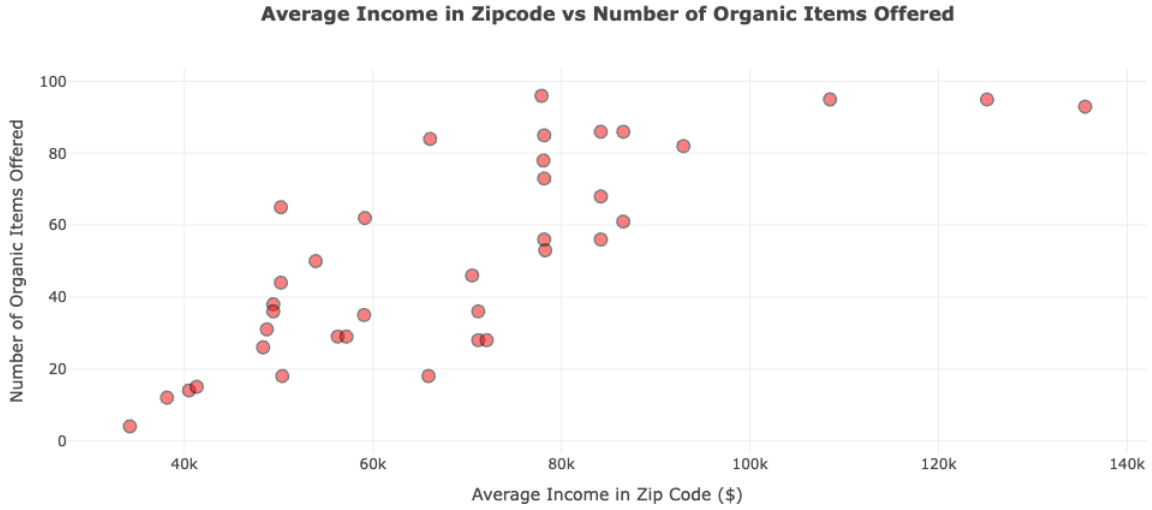
  Examples: Temperature and Ice Cream Sales, Study Time and Exam Scores, Exercise and Weight Loss, etc

- Negatively associated: Two variables are negatively associated when one variable increases other variable decreases, and vice versa. The scatterplot slopes downward from left to right.

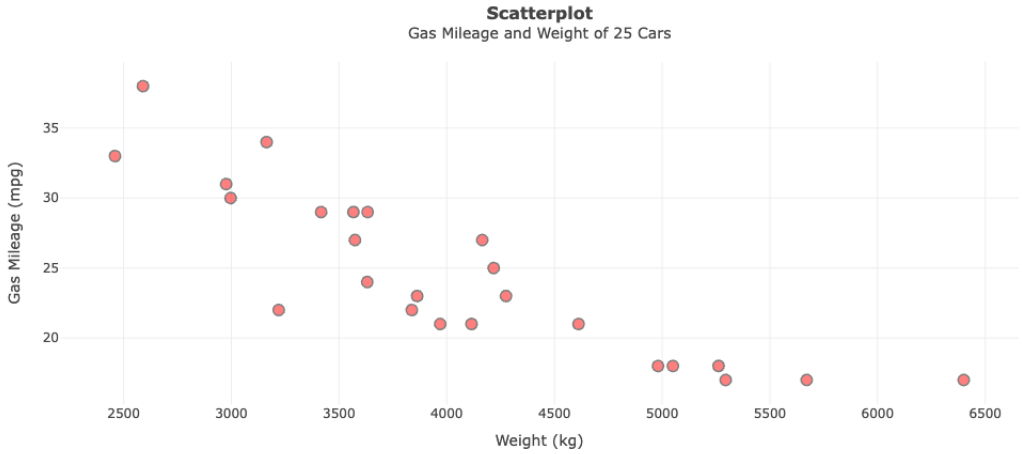
  Examples: Elevation and Temperature, Price of a Product and Quantity Demanded, etc.

Positive or negative association?

a.



b.



(a) is a positive association as when average income in zipcode increases, number of organic items offered also increases. Similarly, (b) is a negative association as when weight of a car increases, gsa mileage decreases and vice-versa.

4 Correlation

So far we have only eyeballed the relationship between two variables. But if we want to be robust in our analysis, we need to introduce a metric that can measure the association between two variables. Our eyes aren't a good judges of how strong a relationship is.

The correlation describes the direction and strength of a linear (straight-line) relationship between two quantitative variables. Correlation is typically measured using the correlation coefficient, denoted as r. **Note: Correlation measures the association between two variables only if they have a linear relationship.** The correlation coefficient ranges from -1 to 1 and is unitless. Values close to -1 indicate a strong negative linear relationship, values near 0 suggest no linear relationship between the variables, and values close to 1 indicate a strong positive linear relationship.

Correlation coefficient is calculated by using the formula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where: x and y are the observations and $\overline{x}$ and $\overline{y}$ are the mean of x and y respectively.

As you can see from the formula, x and y are treated symmetrically, and the order of x and y doesn't matter. Therefore, swapping x and y will not affect the value of the correlation coefficient. This property is important because it ensures that the correlation coefficient remains consistent regardless of how the variables are labeled or interpreted in a particular context. Hence the correlation between response variable and explanatory variable or the correlation between explanatory variable mean the same.

4.1 General interpretation of correlation values

The following table is a general rule-of-thumb interpretation of correlation coefficients.

Correlation Coefficient, r	**Rule-of-thumb Interpretation**
-1 to -0.7	Strong negative linear relationship
-0.7 to -0.3	Moderate negative linear relationship
-0.3 to -0.1	Weak negative linear relationship
-0.1 to 0.1	Negligible or no linear relationship
0.1 to 0.3	Weak positive linear relationship
0.3 to 0.7	Moderate positive linear relationship
0.7 to 1	Strong positive linear relationship

Exercise: Describe the strength of the linear relationships of the following correlation coefficients.

A. $r = -0.351$

B. $r = 0.575$

C. $r = 0.823$

D. $r = -0.981$

E. $r = 0.11$

Answers:

A. $r = -0.351$ - Moderate negative

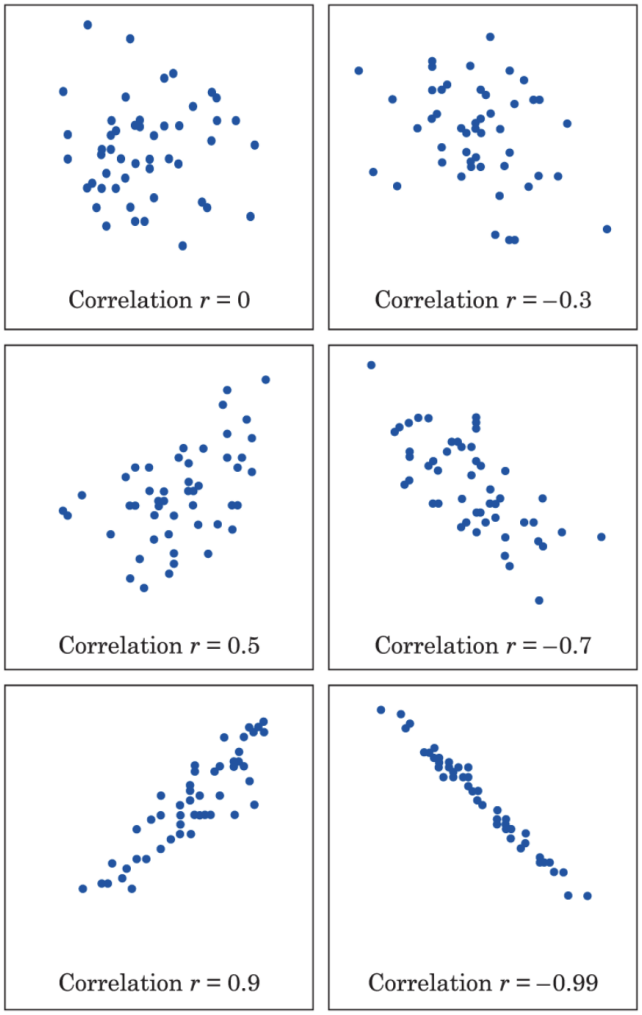
B. $r = 0.575$ - Moderate positive

C. $r = 0.823$ - Strong positive

D. $r = -0.981$ - Strong negative

E. $r = 0.11$ - Weak positive

Let's looks at some of the scatterplots and estimate the correlation coefficients.



Moore/Notz, *Statistics: Concepts and Controversies*, 10e, © 2020 W. H. Freeman and Company

5 Is correlation affected by outliers?

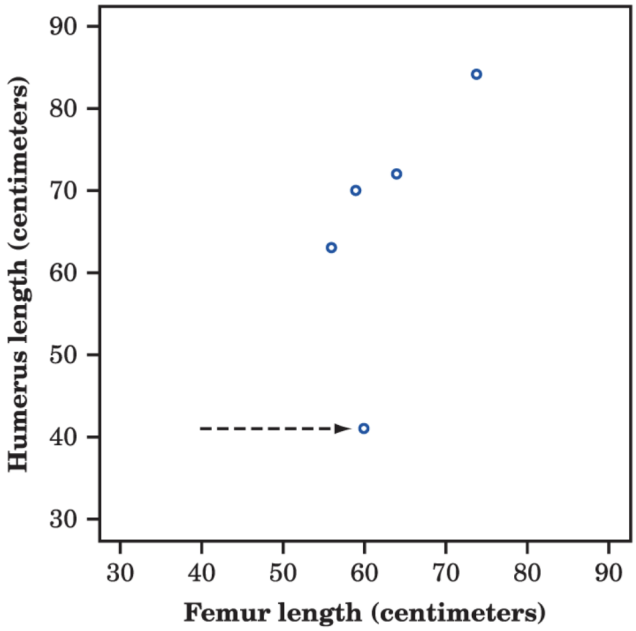
Let's recall the formula to calculate the correlation coefficient.

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where: x and y are the observations and $\overline{x}$ and $\overline{y}$ are the mean of x and y respectively.

As we see the formula to calculate correlation coefficient involves mean and mean is affected by outliers as discussed in previous chapter. Hence, correlation coefficient is also affected by outliers.

Let's look at an example,



Moore/Notz, *Statistics: Concepts and Controversies*, 10e,
© 2020 W. H. Freeman and Company

In the figure above, the scatterplot shows the relationship between femur length and humerus length. The correlation coefficient of the scatterplot without the outlier is $r = 0.994$. Just the inclusion of that outlier brings down the correlation coefficient to $r = 0.640$.

