Association and Causation – Solutions

COR1-GB.1305 – Statistics and Data Analysis

Association and Correlation

1. Let us consider a test of children's ability to write French text correctly from dictation. For simplicity we may assume that the spoken French is standardized by using a tape recorder, that the scoring has been made uniform, and that the scorers can decipher all the handwritings involved. Let

$$\begin{aligned} y &= \text{the score on this dictation test,} \\ x &= \text{the weight of the child.} \end{aligned}$$

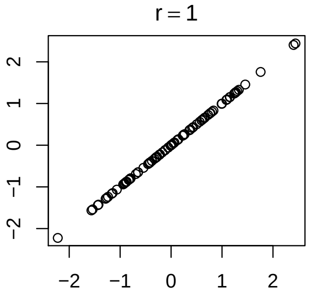
What is the relation between y and x?

Solution: The relation depends on the context (population). Consider the following contexts:

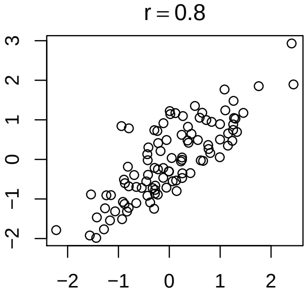
- *The ages of the children vary over a wide range,* say 5 to 15 years. In this case, older children will have been in school longer, and so they will perform better on French dictation, at least where French is taught or spoken. Older children also weight more than younger children. Thus, there will likely be a strong positive relation between weight x and dictation score y.
- *The ages of the children are nearly constant,* say 15 years plus or minus a few weeks. Now, there will still be differences between the children, for instance between girls and boys. At this age, girls tend to have better language profficiency than boys do. They also tend to weight less than boys. Thus, it seems likely that there will be a weak negative relation between x and y.
- *The ages of the children are nearly constant, but they come from a mix of countries,* Let's say we have children from France, Holland, and the U. S. A. If the French tend to be lighter than the Dutch, who tend to be lighter than the Americans, then we would get a strong negative relation between weight x and French dictation y.

2. Consider two variables, x and y. In the scenarios below, we have collected $n = 100$ measurements of (x, y) pairs. There is a scatterplot of the 100 points, along with the computed correlation between the two variables. Is there an apparent association between the two variables? Is the relationship positive or negative? Weak or strong?

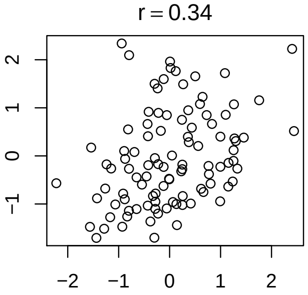
(a)



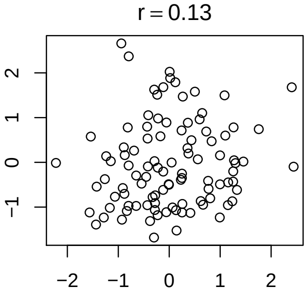
(b)



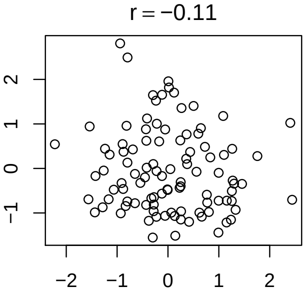
(c)



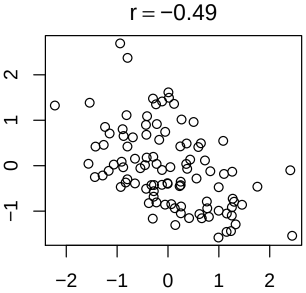
(d)



(e)



(f)



Solution: Looking at the scatterplots directly is the best way to get a sense of the association. The associations are a little subjective, but here are some plausible judgments: (a) extremely strong positive; (b) very strong positive; (c) moderate positive; (d) negligible; (e) negligible; (f) strong negative.

Causation

3. Consider a general x and y. How might one confirm the three requirements for causation?

   (a) Consistency?

   Solution: Consistency can be confirmed by observation alone. We might look at a variety of different populations and see whether the relationship between x and y is consistent in direction and amount.

   (b) Responsiveness?

   Solution: Responsiveness can sometimes be confirmed by experiment. If we can intervene and change x, we can see if y then changes. Note that it is not always possible to do this intervention.

   (c) A mechanism?

   Solution: A mechanism can only be confirmed by explicitly describing the mechanism, and supporting the correspondence between each step in the mechanism and that in the process under study.

Lurking Variables

In each of the following situations, a lurking variable can explain a strong association between two variables (some of these are taken from the Wikipedia article “Correlation does not imply causation”). Give an example of a potential lurking variable.

4. Sleepling with one’s shoes on is strongly correlated with waking up with a headache.

   Solution: Both are associated with going to bed drunk.

5. During the industrial revolution in the United States (1750–1850), there was a strong positive assocation between the amount of scotch whisky imported into New York and the number of ministers there as well.

   Solution: A common cause of both is the size of the population: as population increased, so did the amount of imports and the number of ministers.

6. As ice cream sales increase, the rate of drowning deaths increases sharply.

   Solution: Both increase during the summer.

7. There is a positive association between high school seniors’ GPAs and their SAT scores.

   Solution: It is implausible that someone’s SAT score affects their GPA; the other direction is implausible as well. There are likely many common causes of both of these (e.g. stress level).

8. Since the 1950s, both the atmospheric carbon dioxide level and obesity levels have increased sharply.

   Solution: Richer populations tend to eat more food and consume more energy.

9. An article in the May 13, 1999 issue of *Nature* found that young children who sleep with the light on are much more likely to develop myopia in later life. Explain why this does not imply that sleeping with the light on causes myopia.

Solution: There could be a lurking variable explaining the association. In fact, a later study at Ohio found a strong link between parental and child myopia, and noted that myopic parents were more likely to leave lights on in childrens' bedrooms.