Correlation

8

Contents

What You Will Learn in This Chapter

In Chaps. 3 and 6 we investigated the effect of latitude on tree height by measuring trees at 2 and 3 locations, respectively, and testing for differences in mean heights. As we will see, a better way to answer this question involves testing tree heights at even more locations of latitude. Computing an ANOVA is not a good idea for this situation because the ANOVA does not take the ratio scale properties of the latitude into account. The ANOVA treats each location as nominal (see Chap. 7). Correlations allow us to include the ratio scale aspect of the information and thereby summarize the effect of latitude into one value, r.

8.1 Covariance and Correlations

Let us first visualize correlations. If there were a perfect negative correlation, then an increase in one variable corresponds to a consistent decrease in another variable, for example, tree height decreases as latitude increases. If we plot latitude on the x-axis and tree height on the y-axis, the points fall on a straight line as in Fig. 8.1a (perfect negative

© The Author(s) 2019

M. H. Herzog et al., *Understanding Statistics and Experimental Design*,
Learning Materials in Biosciences, https://doi.org/10.1007/978-3-030-03499-3_8

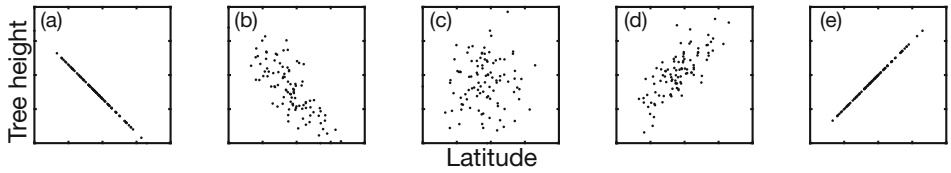

Fig. 8.1 Tree heights versus latitude for five different scenarios, (**a**)–(**e**). Each data point shows the height of one tree at one location. Correlations measure the linear relationship between the two variables

correlation). On the other hand, if there is no relationship between the variables, the data looks like a diffuse cloud of points as in Fig. 8.1c (no correlation). If tree height increases as latitude increases, there is a perfect positive correlation (Fig. 8.1e). Usually, we find cases in between the three basic scenarios (Fig. 8.1b, d).

This linear relationship is captured by the covariance equation:

$$\mathrm{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X}) \times (y_i - \overline{Y})}{n - 1} \tag{8.1}$$

where, for example, the latitude data are x_i, the tree heights are y_i, and the $\overline{X}$ and $\overline{Y}$ are the respective mean values, i.e., the mean latitude and the mean tree height, respectively. The data consists of n pairs of latitudes and tree heights. The covariance generalizes the concept of variance because $\mathrm{cov}(x, x)$ is the variance of x.

A disadvantage of covariance is that it depends on the scale. For example, if you measure tree height in meters the covariance is smaller than if you measure it in centimeters. For this reason, we normalize the covariance by the standard deviation of x and y and arrive at the correlation:

$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y} \tag{8.2}$$

This type of correlation is called Pearson's correlation. Correlation values range between -1.0 and $+1.0$, where -1 indicates a perfect negative correlation, 0 indicates no correlation, and $+1$ indicates a perfect positive correlation (see Fig. 8.1a, c, and e). Values between these limits indicate intermediate strengths of the relationship between the variables.

8.2 Hypothesis Testing with Correlations

Figure 8.2 shows a sample ($n = 50$) of tree height data from many latitudes. Each point corresponds to a single tree. Obviously, there is not a perfect correlation, but the correlation seems to be different from zero. We use hypothesis testing to look for a significant

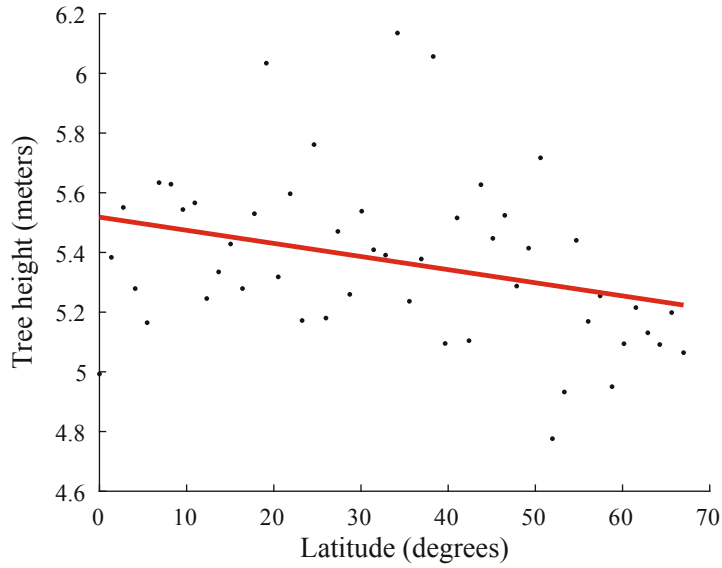

Fig. 8.2 Tree heights versus latitude for a sample of 50 trees. The correlation is $r = -0.312$. The red line is a best fitting straight line

correlation. Our null hypothesis is:

$$H_0 : \rho = 0$$

where ρ corresponds to the population correlation.

We do not need to go into the details, but if the null hypothesis is true, then the standard deviation of the sampling distribution of a sample correlation is:

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}} \tag{8.3}$$

and the appropriate test statistic is a t value computed as:

$$t = \frac{r - 0}{s_r} \tag{8.4}$$

with degrees of freedom $df = n - 2$. The typical statistical software output for the data in Fig. 8.2 would look something like that shown in Table 8.1.

Table 8.1 Typical statistical software outputs for a correlation

r	t	df	p
−0.312	−2.28	48	0.027

Since the p value is less than 0.05, we conclude there is a significant correlation. The fact that the r-value is negative indicates that taller trees are found at lower latitudes

8.3 Interpreting Correlations

Assume we found a significant correlation between variables x and y, what does it tell us? First, it does not tell us that x causes y. This can be simply understood by noting that

$$\text{cov}(x, y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X}) \times (y_i - \overline{Y})}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \overline{Y}) \times (x_i - \overline{X})}{n-1} = \text{cov}(y, x) \tag{8.5}$$

which, if interpreted improperly, would suggest that x causes y and that y causes x. A significant correlation can occur for four reasons:

1. x causes y
2. y causes x
3. some intermediate variable z causes x and y
4. the correlation is spurious

An example for an intermediate variable (reason 3): it is not the latitude that determines tree heights. Rather factors related to latitude directly influence tree heights, such as water supply. Spurious correlations (reason 4) can occur by random. For example, for years 2000–2009 the correlation is $r = 0.947$ between US per capita consumption of cheese and the number of people who died by becoming tangled in their bedsheets. If scientists find such a high correlation in an experiment, they open a bottle of champagne! Spurious correlations are inevitable if you look across large enough sets of data.

It is important to note that because correlations only measure linear relationships, a non-significant correlation does not mean there is no relationship (or causation) between x and y. For example, air temperature systematically changes with time of day in a sinusoidal fashion (it goes up and down during the day-night cycle), but a correlation between time of day and temperature might produce $r \approx 0$.

It is always a good idea to look at a graph of data in addition to computing a correlation. Data of very different types can give rise to the same r-value (Fig. 8.3), so knowing

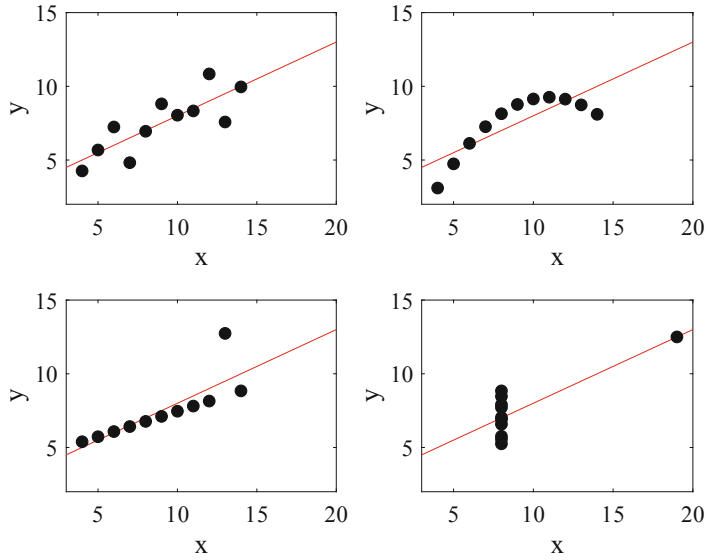

Fig. 8.3 Anscomb's quartet. Each data set has the same r-value ($r = 0.816$) despite looking very different when plotted

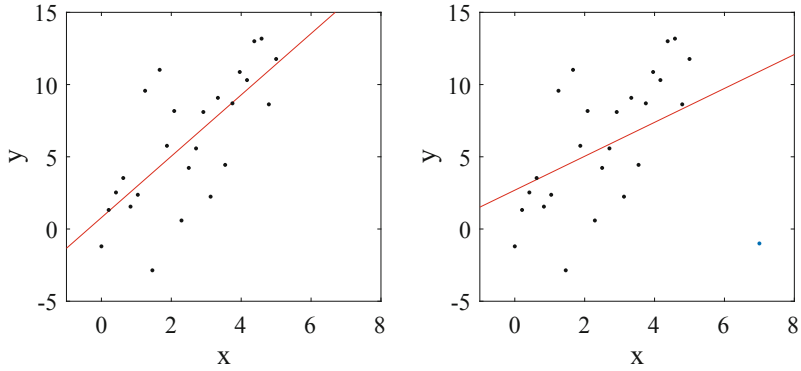

Fig. 8.4 Outliers can have a substantial impact on correlation. Left: original data set with $r = 0.71$. Right: a single outlier has been added (blue point at the bottom right) causing a big decrease in the correlation ($r = 0.44$)

only the correlation value provides only partial information about the data set. Moreover, correlations are very sensitive to outliers (Fig. 8.4), and a single data point added or removed from a data set can dramatically change the correlation value.

Table 8.2 Effect size guidelines for $|r|$ according to Cohen

	Small	Medium	Large
Effect size	0.1	0.3	0.5

8.4 Effect Sizes

Correlation is often used as a measure of effect size that indicates how much one variable is related to another variable. In particular, the square of a correlation, r^2, indicates the proportion of variability in one score (e.g., tree height) that can be explained by variability in the other score (e.g., latitude). This is the same kind of information provided by η^2, which we covered in Chap. 6. According to Cohen, an r-value of less than 0.1 is considered a small effect and the very same is true for values lower than -0.1 (Table 8.2).

8.5 Comparison to Model Fitting, ANOVA and t-Test

In Chap. 7 we fit a linear model to the learning data and focused on the slope, which is similar to computing a correlation because the correlation is a measure of linear relationships. A hypothesis test for a non-zero slope gives the same result as a hypothesis test for a non-zero correlation.

As mentioned in Chap. 7, it is not a good idea to use an ANOVA when the independent variable is on a ratio scale because the ANOVA treats the independent variable as being on a nominal scale. By taking full advantage of the ratio scale an analysis based on correlation has higher power than an ANOVA.

One could also use the t-test by splitting the data into, for example, smaller and larger than median latitudes, i.e., half the data go into a North group, the other half into a South group. In general, such approaches are not as good as an analysis based on the correlation because they (again) do not include the ratio scale nature of the independent variable. For example, in Fig. 8.5 the data from Fig. 8.2 are split into lower and higher latitude regions. The t-test does not produce a significant result. Thus, if we analyze the data with these subsets, we fail to note the significant difference found by looking at the correlation in the original data set (Table 8.1).

In some way, a correlation may be seen as a generalization of the ANOVA and the t-test.

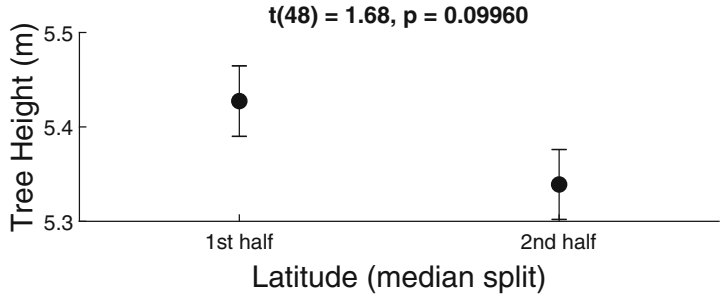

Fig. 8.5 Data produced from a median split of the data in Fig. 8.2. A t-test investigating differences between the means is not significantly different

8.6 Assumptions and Caveats

Hypothesis tests for correlations hold several assumptions.

1. As always, data need to be independent and identically distributed.
2. The y-variable is Gaussian distributed when conditioned on any given x-value. That is, if we were to take all the y-values at a single x-value and make a histogram of them, the histogram would be Gaussian distributed.
3. Both variables are interval or ratio scaled.
4. Sample size is fixed before the experiment.

If data are on an ordinal scale, correlations can be computed with the Spearman's ρ, which uses ranks (ordinal scale) rather than the ratio scale. Spearman correlations are the non-parametric equivalent of the parametric Pearson correlations.

8.7 Regression

In this subsection, we quickly sketch the relationship between correlations and regressions. The hasty reader may skip it. Regression will play no role in the following chapters.

A correlation tells us about how tightly packed the data are around the best fitting line. For example a correlation of 1.0 tells us that all data points are perfectly on the line. However, what is this best fitting line? Regression gives us the equation of that best fitting line, which has one parameter for the slope (m) and one for the y-intercept (b; i.e., where the line hits the y-axis). The slope of the regression line is the standard deviation in the y-direction divided by the standard deviation in the x-direction, weighted by the r value from Eq. 8.2:

$$m = r\frac{s_y}{s_x} \tag{8.6}$$

Table 8.3 Typical statistical software outputs for a regression

Parameter	Coefficient value	t	p
Intercept (constant)	12.146	4.079	0.00017
Slope (latitude)	−0.147	−2.275	0.027

This means for every standard deviation we walk in the x-direction, we step up by the standard deviation in the y-direction multiplied by the r-value.

The intercept b is:

$$b = \bar{y} - m\bar{x}$$

For the tree height data presented in Fig. 8.2, the slope is $m = -0.1473$ and the intercept is $b = 12.1461$. This means that at a latitude of zero degrees, the average tree height is 12.1461 m, and that for every degree of latitude that we go North of that, we increase in tree height by −0.1473 m (in other words, tree heights go down as we increase our latitude). These results are typically summarized in statistical software as shown in Table 8.3

Here, in addition to the regression line slope and intercept, the statistical software also outputs a t- and p-value for the slope and intercept, the so-called regression coefficients. These statistics test the null hypothesis that the slope and intercept are equal to zero. In this example, the p-values are smaller than 0.05, and so both are significantly different from zero. In such a situation, the corresponding correlation (r-value) is typically significantly different from zero. An intercept that is not significantly different from zero means that the regression line roughly crosses the point (0, 0) on the graph.

Take Home Messages

1. Correlations are the preferred choice if both the x- and y-axis are ratio or interval scaled.
2. Causation and correlation should never be confused.
3. Very different sets of data can lead to the same r.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.