MAKING SENSE OF DATA

Essentials series

CORRELATION AND REGRESSION

Prerequisites
Descriptive statistics
Charts and graphs
The normal distribution
Surveys and sampling
Correlation and regression

Relationship with a scatterplot

A fair amount of research supports the claim that the frequency which a tree cricket chirps is related to temperature. This relationship (known as Dolbear's Law) can be used to predict the temperature using the number of times the crickets chirp per 15 seconds.

Number of Chirps (in 15 seconds)	Temperature (°F)
18	57
20	60
21	64
23	65
27	68
30	71
34	74
39	77

Table 1: Cricket chirps and temperature data

Notice that each observation is composed of two variables that are tied together, in this case the number of times the cricket chirped in 15 seconds (the X-variable), and the temperature at the time the data was collected (the Y-variable). Statisticians call this type of two-dimensional data *bivariate* data. Each observation contains one pair of data collected simultaneously.

Making a scatterplot

Bivariate data are typically organised in a graph known as a *scatterplot*. A scatterplot has two dimensions, a horizontal dimension (called the x-axis) and a vertical dimension (called the y-axis). Both axes are numerical - each contains a number line.

The x-coordinate of bivariate data corresponds to the first piece of data in the pair; the y-coordinate corresponds to the second piece of data in the pair. If you intersect the two coordinates, you can graph the pair of data on a scatterplot. Figure 1 shows a scatterplot of the data from Table 1.

Interpreting a scatterplot

You interpret a scatterplot by looking for trends in the data as you go from left to right:

- ✓ If the data show an uphill pattern as you move from left to right, this indicates a *positive relationship between X and Y*. As the *x*-values increase (move right), the *y*-values increase (move up) a certain amount.
- ✓ If the data show a downhill pattern as you move from left to right, this indicates a *negative relationship between X and Y*. That means as the *x*-values increase (move right) the *y*-values decrease (move down) by a certain amount.
- ✓ If the data don't resemble any kind of pattern (even a vague one), then no relationship exists between *X* and *Y*.

This topic focuses on linear relationships. A *linear relationship between X and Y* exists when the pattern of *x*- and *y*-values resembles a line, either uphill (with positive slope) or downhill (with negative slope).

Looking at Chart 1, there does appear to be a positive linear relationship between number of cricket chirps and the temperature. That is, as the cricket chirps increase, you can predict that the temperature is higher as well.

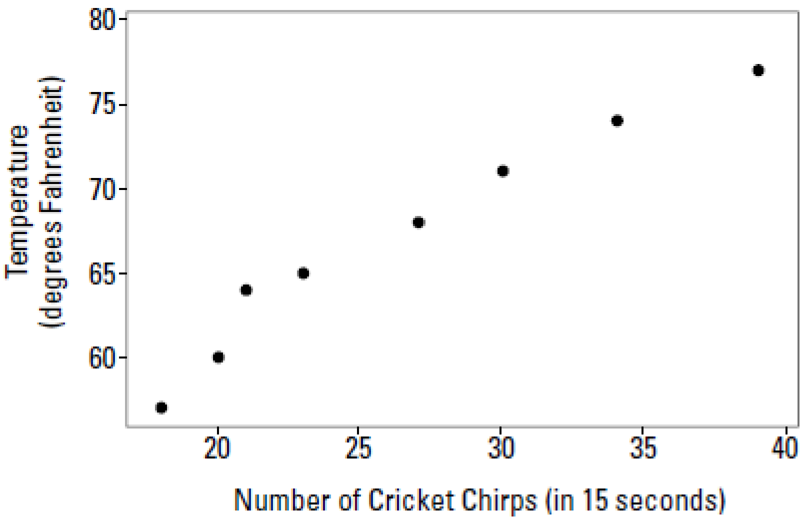

Chart 1: Scatterplot of cricket chirps versus outdoor temperature

Measuring relationships using the correlation

After the bivariate data have been organized, the next step is to do some statistics that can quantify or measure the extent and nature of the relationship.

Calculating the correlation

The pattern and direction of the relationship between X and Y can be seen from the scatterplot. The strength of the relationship between two numerical variables depends on how closely the data resemble a certain pattern. Although many different types of patterns can exist between two variables, this topic considers linear patterns only.

Statisticians use the *correlation coefficient* to measure the strength and direction of the linear relationship between two numerical variables X and Y. The correlation coefficient for a sample of data is denoted by r.

The formula for the correlation (r) is

$$r = \frac{1}{n-1}\sum\frac{(x-\bar{x})(y-\bar{y})}{s_x s_y}$$

where n is the number of pairs of data; $\bar{x}$ and $\bar{y}$ are the sample means; and s_x and s_x are the sample standard deviations of the x- and y- values, respectively.

For example, given the bivariate data set (3, 2), (3, 3), and (6, 4) what is the correlation (r)? Table 2 shows the x- and y- values.

x	y
3	2
3	3
6	4

Table 2: x- and y- values

<u>Step 1</u>: Find the mean of all the x-values ($\bar{x}$) and the mean of all the y-values ($\bar{y}$).

As the mean is given by $\sum x_i/_n$, $\bar{x} = {}^{(3+3+6)}/_3$ or $\bar{x} = 4$; similarly $\bar{y} = {}^{(2+9+4)}/_3$ or $\bar{y} = 3$.

<u>Step 2</u>: Find the standard deviation of all the x-values (s_x) and the standard deviation of all the y-values (s_y).

As the standard deviation (s) is given by

$$s = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$

we can construct this table using our x-values

x	$x - \bar{x}$	$(x-\bar{x})^2$
3	-1	1
3	-1	1
6	2	4

then applying this formula (n=3) we have

$$s_x = \sqrt{\frac{(1+1+4)}{(3-1)}} = \sqrt{\frac{6}{2}} = \sqrt{3} \;\; or \;\; 1.73$$

This same procedure applied to the y-values gives $s_y = 1.00$.

<u>Step 3</u>: For each (x, y) pair in the data set, take x minus $\bar{x}$ and y minus $\bar{y}$, and multiply them together.

The differences multiplied together are: (3 – 4)(2 – 3) = (−1)(−1) = 1; (3 – 4)(3 – 3) = (−1)(0) = 0; (6 – 4)(4 – 3) = (+2)(+1) = +2.

<u>**Step 4**</u>: Add up all the results from Step 3.

Adding the Step 3 results, you get $1 + 0 + 2 = 3$

<u>**Step 5**</u>: Divide the sum by $s_x \times s_y$.

Dividing by $s_x \times s_y$ gives $3/(1.73 \ast 1.00) = 3/1.73 = 1.73$.

<u>**Step 6**</u>: Divide the result by n – 1, where n is the number of (x, y) pairs.

Dividing the Step 5 result by 3 – 1 (which is 2) and you have the correlation $r = 0.87$.

Interpreting the correlation

The correlation r is always between $+1$ and -1. Here is how you interpret various values of r.

A correlation that is …	Indicates …
exactly -1	a perfect downhill linear relationship
close to -1	a strong downhill linear relationship
close to 0	no linear relationship exists
close to +1	indicates a strong uphill linear relationship
exactly +1	indicates a perfect uphill linear relationship

Table 3: Interpreting the correlation coefficient (*r*)

How “close” do you have to get to -1 or $+1$ to indicate a strong linear relationship? Most statisticians prefer to see correlations above $+0.60$ (or below -0.60) before getting too excited.

Properties of the correlation

Here are two important properties of correlation:

- ✓ The correlation is a unit-less measure. This means that if you change the units of X or Y, the correlation doesn’t change. For example, changing the temperature (Y) from Fahrenheit to Celsius won’t affect the correlation.
- ✓ The variables X and Y can be switched in the data set, and the correlation doesn’t change. For example, if height and weight have a correlation of 0.53, weight and height have the same correlation.

Finding the regression line

After you've found a linear pattern in the scatterplot, and the correlation between the two numerical variables is moderate to strong, you can create an equation that allows you to predict one variable using the other. This equation is called the *simple linear regression line*.

Which is X and which is Y?

When doing correlations, the choice of which variable is X and which is Y doesn't matter, as long as you're consistent for all the data; but when fitting lines and making predictions, the choice of X and Y makes a difference. In general, X is the variable that is the predictor. Statisticians call the X-variable the explanatory (or independent) variable, because if X changes, the slope tells you (or explains) how much Y is expected to change. The Y-variable is called the response (or dependent) variable because if X changes, the response (according the equation of the line) is a change in Y. Hence Y can be predicted by X if a strong relationship exists.

Note: In our cricket example, you want to predict the temperature based on listening to chirps. The real cause-and-effect is the opposite: as temperature rises, cricket chirps increase.

Checking conditions

In the case of two numerical variables, it's possible to come up with a line that you can use to predict Y from X, if (and only if) the following two conditions are met:

1) The scatterplot must find a linear pattern; and
2) The correlation, r, is moderate to strong (beyond ± 0.60)

Understanding the equation

For the crickets and temperature data (see Table 1), you see the scatterplot in Figure 1 shows a linear pattern. The correlation between cricket chirps and temperature is found to be very strong ($r = 0.98$). You now can find one line that best fits the data (in terms of the having the smallest average distance to all the points.). Statisticians call this technique for finding the best-fitting line a *simple linear regression analysis*.

The formula for the best-fitting line (or *regression line*) is $y = mx + b$, where m is the slope of the line and b is the y-intercept. The slope of a line is the change in Y over the change in X. For example, a slope of 10/3 means as the x-value increases (moves right) by 3 units, the y-value moves up by 10 units on average.

The y-intercept is that place on the y-axis where the line crosses. For example, in the equation $y = 2x – 6$ the line crosses the y-axis at the point -6 (see Chart 2). The coordinates of this point are (0,–6); when a line crosses the y-axis, the x-value is always 0. To come up with the best-fitting line, you need to find values for m and b that best-fit the pattern of data.

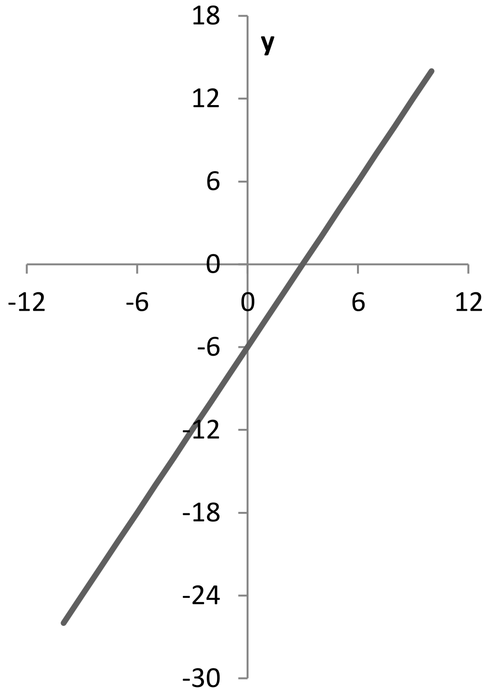

Chart 2: Best-fit line for y = 2x – 6

Finding the slope

The formula for the slope, m, of the best-fitting line is

$$m = r\left(\frac{s_y}{s_x}\right)$$

where

r = the correlation between X and Y
s_y = the standard deviations of the y-values
s_x = the standard deviations of the x-values

The correlation and the slope of the best-fitting line are not the same. The formula for slope takes the correlation (a unit-less measurement) and attaches units to it.

Finding the y-intercept

The formula for the y-intercept, b, of the best-fitting line is

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

where

$\bar{x}$ = the mean of the x-values
$\bar{y}$ = the mean of the y-values
m = the slope

Keep in mind that five well-known summary statistics are all you need to do all the necessary calculations: $\bar{x}, \bar{y}, s_x, s_y \; and \;\; r$.

Interpreting the slope and y-intercept

Even more important than being able to calculate the slope m and y-intercept to form the best-fitting regression line is the ability to interpret their values.

Interpreting the slope

The slope is interpreted as "rise over run". If the slope for example is 2, you can write this as $2/1$ and say as X increases by 1, Y increases by 2, and that's how you move along from point to point on the line. In a regression context, the slope is more important because it tells you how much you can expect Y to change as X increases.

In general, the units for the slope are the units of the Y-variable per units of the X-variable. It's a ratio of change in Y per change in X. Suppose in studying the effect of dosage level in milligrams (mg) on blood pressure a researcher finds that the slope of the regression line is -2.5. You can write this as $-2.5/1$ and say blood pressure is expected to decrease by 2.5 points on average per 1 mg increase in drug dosage.

Interpreting the y-intercept

The y-intercept is the place where the regression line $y = mx + b$ crosses the y-axis and is denoted by b. Sometimes the y-intercept can be interpreted in a meaningful way, and sometimes not. This differs from the slope, which is always interpretable.

There are times when the y-intercept makes no sense. For example, suppose rain is used to predict bushels per acre of wheat; if the regression line crosses the y-axis somewhere below zero (as it most likely will), the y-intercept will make no sense. You can't have negative wheat production.

The best-fitting line for the crickets

The five summary statistics for the cricket data (see Table 1) are shown below in Table 4.

Variable	Mean	Standard Deviation	Correlation
Chirps (x)	$\bar{x} = 26.5$	$s_x = 7.4$	$r = +0.98$
Temp (y)	$\bar{y} = 67$	$s_y = 6.8$	

Table 4: Five summary statistics for cricket data

The slope m for the best-fitting line for the cricket chirp vs. temperature data is $m = r\left(\frac{s_y}{s_x}\right) = 0.98\left(\frac{6.8}{7.4}\right) = 0.90$. So, as the number of chirps increases by 1 chirp per 15 seconds, the temperature is expected to increase by 0.90 degrees Fahrenheit on average. To get a more practical interpretation, you can multiply the top and bottom of the slope by 10 to get 9.0/10 and say that as chirps increase by 10 (per 15 seconds), temperature increases 9 °F.

Now, to find the y-intercept b you calculate $\bar{y} - (m \times \bar{x})$, giving $67 - (0.90 \times 26.5) = 43.15$. So the best-fitting line for predicting temperature from cricket chirps based on the data is

$$y = 0.90x + 43.15$$

where

y = the temperature (°F)
x = the number of chirps (in 15 seconds)

The y-intercept would try to predict temperature when there is no chirping at all. However, no data can be collected at or near this point, so we can't make predictions for temperature in this area.

Making predictions

After you have a strong linear relationship, and you find the equation of the best-fitting line $y = mx + b$, you use that line to predict y for a given x-value. This amounts to plugging the x-value into the equation and solving for y. For example, if your equation is $y = 2x + 1$, and you want to predict y for $x = 1$, then plug 1 into the equation for x to get $y = 2(1) + 1 = 3$.

Remember that you choose the values of (the explanatory variable) that you plug in; what you predict is Y, the response variable, which totally depends on X. By doing this, you are using one variable that you can easily collect data on, to predict a Y variable that is difficult or not possible to measure; which works well as long as X and Y are correlated. This is the big idea of regression.

Extrapolation

Just because you have a model doesn't mean you can plug in any value for X and correctly predict Y. For example, in the chirping data, there is no data collected for less than 18 chirps or more than 39 chirps per 15 seconds (see Table 1). If you try to make predictions outside this range you're going into uncharted territory; the farther outside this range you go with your x-values, the more dubious your predictions for y will get.

Making predictions using x-values that fall outside the range of your data is not recommended. Statisticians call this *extrapolation*; watch for researchers who try to make claims beyond the range of their data.

Correlation doesn't necessarily mean cause-and-effect

Scatterplots and correlations identify and quantify relationships between two variables. However, if a scatterplot shows a definite pattern and the data are found to have a strong correlation, that doesn't necessarily mean that a cause-and-effect relationship exists between the two variables. A *cause-and-effect relationship* is one where a change in X causes a change in Y. (In other words, the change in Y is not only associated with a change in X, it is directly caused by X.)

For example, suppose a well-controlled medical experiment is conducted to determine the effects of dosage of a certain drug on blood pressure. The researchers look at their scatterplot and see a definite negative linear pattern; they calculate the correlation and it is strong. They conclude that increasing the dosage of this drug causes a decrease in blood pressure. This cause-and-effect conclusion is sound because they controlled for other variables that could affect blood pressure in their experiment, such as other drugs taken, age, general health, and so on.

However, if you made a scatterplot and examined the correlation between ice cream consumption versus murder rates, you would also see a strong linear relationship (this time positive). Yet no one would claim that more ice cream consumption causes more murders to occur.

What's going on? In the drug example, the data were collected through a well-controlled medical experiment, which minimizes the influence of other factors that might affect blood pressure. In the second example, the data were only based on observation, and no other factors were examined. It turns out that this strong relationship exists because increases in murder rates and ice cream sales are both related to increases in temperature. (Temperature in this case is called a *confounding variable*; it affects both X and Y).

Whether two variables are found to be causally associated depends on how the study was conducted. Only a well-designed experiment or a large collection of several different observational studies can show enough evidence for cause-and-effect.