Mathematics Learning Support Centre

Statistics: Correlation

Richard Buxton. 2008.

1 Introduction

We are often interested in the relationship between two variables.

- Do people with more years of full-time education earn higher salaries?
- Do factories with more safety officers have fewer accidents?

Questions like this only make sense if the possible values of our variables have a natural order. The techniques that we look at in this handout assume that variables are measured on a scale that is at least *ordinal*. In discussing Pearson's correlation coefficient, we shall need to go further and assume that we have interval scale data - i.e. that equal intervals correspond to equal changes in the characteristic that we are trying to measure.

2 Plotting the data

The first step in looking for a correlation is to draw a scatterplot of the data. Figure 1 shows four examples[1].

2.1 Interpreting a scatterplot

These are some of the points to look for.

- How strong is the relationship?
  - In Figure 1, the relationship between gas consumption and outside temperature is very strong, while the relationship between Educational level and Crime rate is much weaker.
- Is the relationship increasing or decreasing?
  - In the 'Gas' example, higher outside temperatures are associated with *lower* gas consumption, but in the 'Ice cream' example, higher mean temperatures go with higher levels of ice cream consumption.

[1] Source of data: Hand(1994)

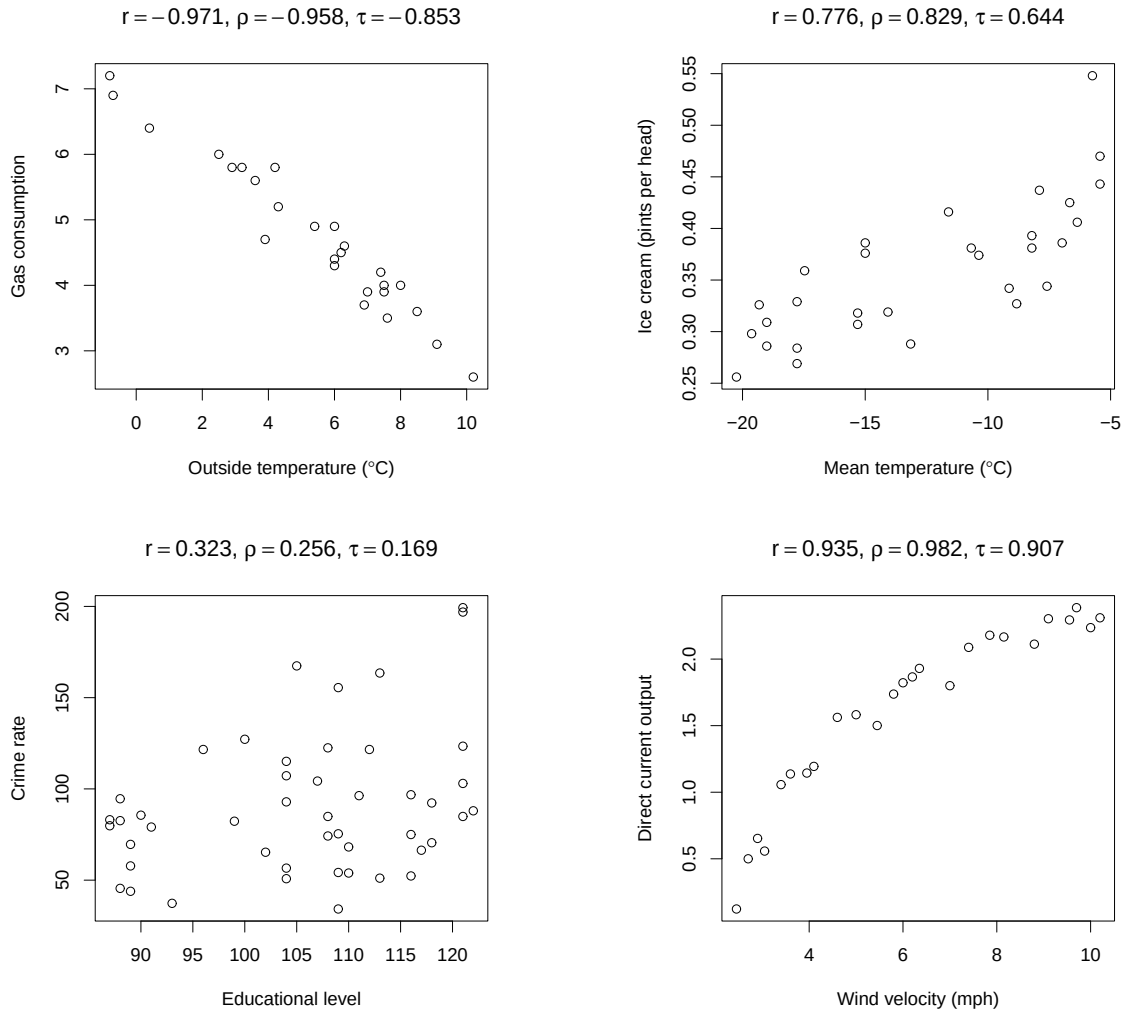

Figure 1: Scatterplots showing strong and weak relationships

- Is the relationship roughly linear?
  - The plot in the top left of Figure 1 shows a clear linear pattern, while the plot in the bottom right suggests a non-linear relationship with the initial steep slope leveling off as the wind speed increases.
- What is the *slope* of the relationship?
  - Is an increase in one variable associated with a small, or a large, increase in the other one? For example, factories with more safety officers may have fewer accidents, but is the reduction in accidents large enough to justify the cost of the additional safety officers?
- Are there any *outliers*?
  - Figure 2 shows a plot of Police expenditure per head against Population size

for 47 US states[2]. At first glance, there seems to be an increasing relationship, with larger states spending more per head on policing. But if you cover up the two outliers at the top right of the plot, the correlation seems to disappear. The evidence for a correlation comes almost entirely from these two points.

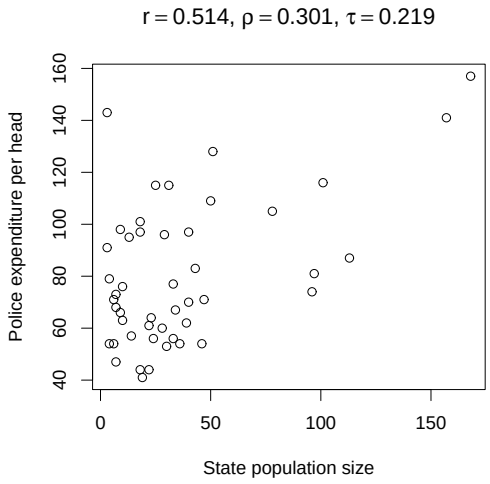

Figure 2: Effect of outliers

2.2 Scatterplots in SPSS

The simplest way to draw a scatterplot in SPSS is to use the *Chart Builder*.

—**Graphs**
—**Chart Builder**
—choose **Scatter/Dot**
—drag the **Simple Scatter** plot into the plotting region
—drag the variables that you want to plot into the **X-Axis** and **Y-Axis** boxes
—Click **OK**

If your data can be split into distinct groups - for example, be gender, you may find it helpful to use a **Grouped Scatter** plot, instead of a **Simple Scatter** plot. Put the two main variables on the x and y axes, as above, but then drag the grouping variable (e.g. gender) into the **Set Colour** box.

If you want to look at all pairwise correlations among a group of variables, use a scatterplot matrix. Drag the **Scatterplot Matrix** into the plotting region and drag all your variables into the **Scattermatrix** box.

[2]Source of data: Hand(1994)

3 Correlation coefficients

A correlation coefficient gives a numerical summary of the degree of association between two variables - e,g, to what degree do high values of one variable go with high values of the other one?

Correlation coefficients vary from -1 to +1, with positive values indicating an increasing relationship and negative values indicating a decreasing relationship.

We focus on two widely used measures of correlation - Pearson's r and Kendall's τ.

- Pearson's coefficient
  - measures degree to which a relationship conforms to a straight line
- Kendall's coefficient
  - measures degree to which a relationship is always increasing or always decreasing

Spearman's rank correlation coefficient, ρ behaves in much the same way as Kendall's τ, but has a less direct interpretation.

3.1 Which coefficient should I use?

- Interval scale data and interested in linear relationships - e.g. wish to build linear model
  - Use Pearson's coefficient
- Interested in *any* increasing/decreasing relationship
  - Use Kendall's coefficient

3.2 Pearson's coefficient

Suppose we have n data pairs (x_i, y_i)

Pearson's correlation coefficient is given by

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

... where$\overline{x}$ and $\overline{y}$ are the means of the x and y values.

In practice, we always use statistical software to do the calculations.

Looking back at Figure 1, notice how the absolute size of the coefficient drops towards zero as we get more and more scatter. While Pearson's r is good at measuring the strength of

a *linear* association, it can be quite misleading in the presence of curvature. Look at the wind turbine data in the bottom right plot of Figure 1. Pearson's r is 0.935, suggesting a strong linear association, but a linear model would clearly not be sensible here.

Because Pearson's r is based on the idea of linearity, it only makes sense for data that is measured on at least an *interval* scale. For ordinal data, use Kendall's τ or Spearman's ρ.

3.3 Kendall's coefficient

Kendall's τ can be used with any variables that are at least ordinal.

Each pair of data points is classified as concordant, discordant or tied.

- Concordant
  - Both variables increase or both variables decrease
- Discordant
  - One variable increases while the other one decreases
- Tied
  - One or both variables stays constant

Writing C, D and T for the number of concordant, discordant and tied pairs, Kendall's coefficient is given by...

$$\tau = \frac{C - D}{N}$$

...where $N = C + D + T$ (the total number of pairs).

The idea is that *concordant* pairs suggest an increasing relationship, while *discordant* pairs suggest a decreasing relationship. Kendall's τ is just the proportion of concordant pairs minus the proportion of discordant pairs.

The value of τ gives a measure of the degree to which a relationship is always increasing, or always decreasing - see Figure 1.

3.4 Modification of Kendall's coefficient for tied data

If some of the pairs of observations are tied, Kendall's coefficient cannot reach the limiting values of ± 1 even if all untied pairs are concordant (discordant). This is a particular problem in the analysis of contingency tables, where there will usually be a large number of ties. Kendall proposed the following as an alternative to the simpler coefficient defined above.

$$\tau_b = \frac{C - D}{\sqrt{(n(n-1)/2 - t_x)(n(n-1)/2 - t_y)}}$$

... where t_x is the number of tied x values and t_y is the number of tied y values.

This version of Kendall's τ is the one used by SPSS.

3.5 Interpreting a correlation coefficient

It's easy to misinterpret a correlation coefficient. These are some of the points to watch.

- A correlation coefficient can be badly affected by one or two outlying observations. For the 'police expenditure' data in Figure 2, the value of Pearson's r is 0.514, but if the two outliers at the top right of this plot are removed, the correlation drops to 0.237. Always look at a scatter plot before calculating a correlation coefficient!

- Correlation is not the same as causality. For example, factories with more safety officers may have fewer accidents, but this doesn't prove that the variation in accident levels is attributable to the provision of safety officers. The correlation may be a spurious one induced by another factor such as the age of the factory.

  One possible approach here is to use *partial correlation*. We 'adjust' our two variables to remove any variation that can be accounted for by our third variable (age of factory) and then look at the correlation between the two adjusted variables.

- Even if a relationship is genuine, a strong correlation doesn't necessarily imply that a change in one variable will produce a large change in the other one. The two sets of data shown in Figure 3 give the same correlation coefficient, but say quite different things about the effect of engine capacity on fuel economy.

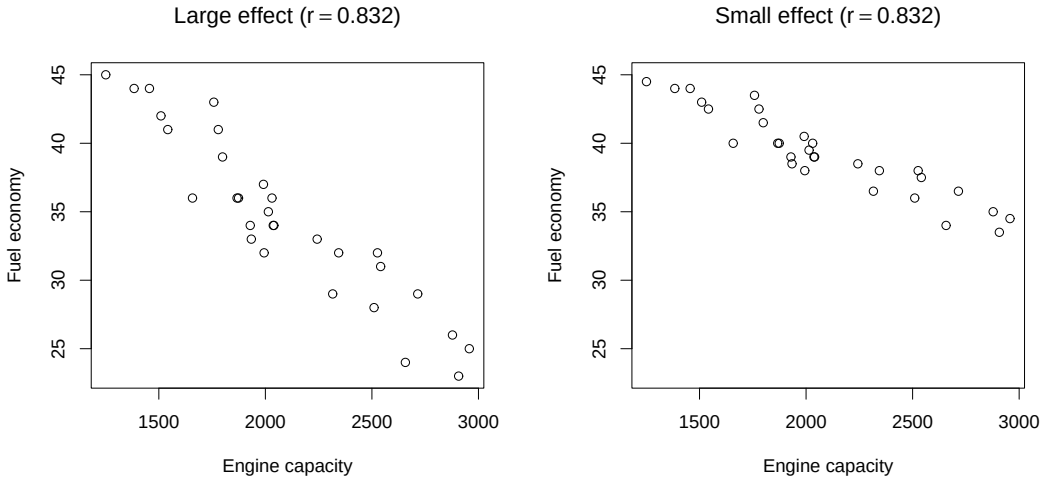

Figure 3: Correlation and size of effect

- Correlation coefficients are subject to sampling variation and may give a misleading picture of the correlation in the population we're sampling. We can quantify the uncertainty in an estimate of a correlation by quoting a confidence interval, or

range of plausible values. For the 'Ice cream' data in Figure 1, the 95% confidence interval for Pearson's r is 0.576 to 0.888, so we can be fairly sure that the population coefficient lies in this range.

For details of how to calculate confidence intervals for correlation coefficients, see Howells(1994) and Hollander(1999).

3.6 Testing for zero correlation

Most statistical software packages allow us to check whether a sample correlation is compatible with zero correlation in the population we're sampling. The test that is carried out here first assumes that the population correlation is zero and calculates the chance of obtaining a sample correlation as large or larger in absolute size than our observed value - this chance is given as the *p value*. If the *p value* is very small, we conclude that our sample correlation is probably incompatible with zero correlation in the population.

The limitation of a test for zero correlation is that it doesn't tell us anything about the *size* of the correlation. A correlation can be nonzero, but too small to be of any practical interest. For example, if we test for zero correlation with the data in the plot in the bottom left of Figure 1, we obtain p value of 0.027, which gives strong evidence for a nonzero correlation. But would a relationship as weak as this be of any practical interest?

3.7 Correlation coefficients in SPSS

—**Analyze**
—**Correlate**
—**Bivariate**
—Drag the two variables that you want to correlate into the **Variables** box
—Select the required correlation coefficients
—Click **OK**

Table 1 shows the SPSS output for the Ice cream data shown in Figure 1. This table relates to Pearson's coefficient - the output for Kendall's τ and Spearman's ρ is similar.

Correlations

		Consumption	Temperature
Consumption	Pearson Correlation	1.000	.776
	Sig. (2-tailed)		.000
	N	30	30
Temperature	Pearson Correlation	.776	1.000
	Sig. (2-tailed)	.000	
	N	30	30

Table 1: SPSS Correlation output

Each box of the table contains the information on the correlation between the corresponding row and column variables. Looking at the top right box, Pearson's r is 0.776, suggesting a moderately strong increasing relationship. The second figure is the p-value for a test of the hypothesis that the population correlation is zero. The figure here has been rounded to 3 decimal places, so the figure of 0.000 tells us that the p-value is less than 0.0005. This is a very small probability, so we can be almost certain that the population correlation is not zero. The third figure tells us the number of observations in our sample.

Unfortunately, SPSS does not provide confidence intervals for correlation coefficients. One package that does offer confidence intervals for both Pearson's and Kendall's coefficients is the package *StatsDirect* - see StatsDirect (2008).

4 References

For a simple introduction to correlation, see Moore (2004). For a more comprehensive treatment, see Howell (2002).

Hand, D.J. (1994). A Handbook of Small Data Sets Chapman and Hall, London.

Hollander, M. and Wolfe, D.A. (1999). Nonparametric Statistical Methods, Wiley, New York.

Howell, D.C. (2002). Statistical methods for psychology, Wiley, New York.

Moore, D.S. (2004). The basic practice of statistics, W.H.Freeman, New York.

StatsDirect (2008) See website at **www.statsdirect.com**