Correlation and causality

Anastássios Perdicoúlis
Assistant Professor, ECT, UTAD (http://www.tasso.utad.pt)
Affiliate Researcher, CITTA, FEUP (http://www.fe.up.pt/~tasso)

Abstract

Correlation helps explore relationships in data series, and thus discover interesting and useful patterns. Such information may facilitate the formulation of causal hypotheses, with the proviso that the systems in question are sufficiently familiar and — even better — understood.

1 Introduction

Measurement is key in science (Perdicoúlis, 2013b): it is through measurement that we obtain data (Perdicoúlis, 2013a), which may then provide useful information such as relations between observation parameters. For instance, the size (e.g. in volume or weight) of a fish and the size (again, in volume or weight) of its liver are proportional, or in direct correlation: practically, the bigger the fish, the bigger its liver (Fowler et al., 1998, pp.130–141).

2 Correlation

The calculation and expression of correlation between data sets involves an elaborate study, both in the original fields — for instance, the natural or social sciences — and statistics. Correlation can be sought in *scatter graphs*, such as Figure 1, or in statistical measurements such as the *correlation coefficient* 'r'.

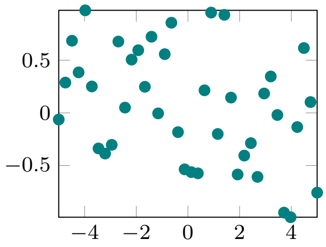

FIGURE 1 A scatter graph with a difficult case for correlation — in fact, a random distribution

The idea of the correlation coefficient 'r' is simple and elegant, with values of '+1' indicating a perfect positive (or direct) correlation, '−1' a perfect negative (or inverse) correlation, and zero indicating no correlation at all. In practice, however, for many different reasons and purposes, there are many versions of ‘r’. *Covariance*, for instance, is a fairly common measure of correlation (Fowler et al., 1998, p.135) — Equation 1.

$$Covar_{x,y} = \frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)} \tag{1}$$

Improving on this idea, the *product moment correlation coefficient* ‘r’ compensates the covariance for different units of measurement between the ‘x’ and ‘y’ variables (Fowler et al., 1998, p.135) — Equation 2.

$$r = \frac{\frac{\sum(x-\bar{x})(y-\bar{y})}{(n-1)}}{s_x \times s_y} \tag{2}$$

A particularly interesting case of correlation is *autocorrelation*: the study of time series data on their own. The corresponding graphic of ‘r’, showing the strength of correlation along time, is known as a *correlogram* (Begon et al., 1996, p.589–591) — Figure 2.

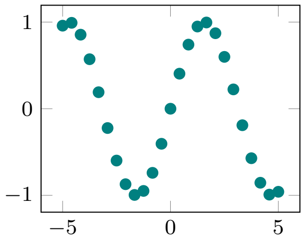

Figure 2 Oscillating autocorrelation ‘r’ pattern

Overall, the data analysis and search for patterns facilitated by correlation can be quite involving, insightful, useful, and even entertaining. However, the utility of correlation does not go as far as identifying causality, and most textbooks will acknowledge this (Hammond and McCullagh, 1978, p.219; Haynes, 1996, p.114; Fowler et al., 1998, p.130).

3 Causality

Causal enquiry is naturally qualitative (Perdicoúlis, 2013c), as it attempts to describe ‘how things are’ in dynamic situations — i.e. those characterised by activity and change — or ‘how things happen’ in terms of cause-and-effect explanations (Perdicoúlis, 2010, p.46). Drawing on information from observations and/ or case descriptions, people use various *heuristic* techniques (e.g. precedence, proximity, ‘sine qua non’, causal mechanisms) to discover and/ or verify causal relationships (Perdicoúlis, 2010, pp.51–55).

Being of qualitative nature, causal hypotheses can be formulated immediately, without a need for long or heavy investment in data collection. Proceeding without quantitative data may appear ‘non-scientific’, but in fact it is perfectly acceptable: science is concerned with the methodic *testing* of hypotheses, leaving considerable liberty for their formulation (Perdicoúlis, 2013d). So, technically it is possible to formulate causal hypotheses relatively quickly, at least to explain a phenomenon tentatively.

4 Discussion

Causal enquiry is creative work, based on knowledge and understanding, and can be aided by (a) studying similar phenomena in different contexts, whose causal model is [thought to be] known, and (b) drawing on information provided by correlation, even though this cannot be always trusted — for instance, causality is known to remain undetected by data correlation in cases of time delays (Sterman, 2000, p.697).

While it is possible and easy to demonstrate or prove correlation, it is very difficult or impossible to prove causality. Furthermore, the attribution of causality usually bears much responsibility, as evidenced remarkably in the courts of law. Let us imagine, for instance, that we were to explain the cause of the 'Great Chicago Fire' of 1871. In those circumstances, one blazing building was propagating the fire to the next building, and this was easy: most buildings were timber-frame constructions, and the fire took place on a windy night after a dry season. The 'cause', or the initial action that started the fire, is *anecdotally* mentioned to be the kick of a cow to a barn lantern — Figure 3. Accusing a cow (or her owner) as clearly as in Figure 3 is quite a bold statement, perhaps in disproportion to its substantiation: i.e. anecdotal evidence instead of solid 'proof'.

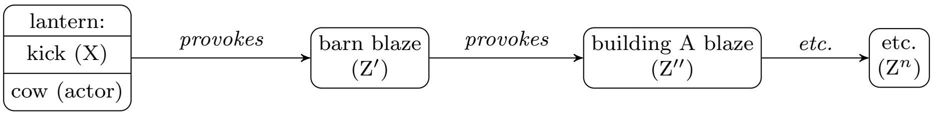

FIGURE 3 Causal scenario hypothesis for the Great Chicago fire of 1871

While correlations tend to give information about 'what' is happening, 'how much' and 'where', *understanding* and knowledge do not come before the causes are determined — at least tentatively, as hypotheses or mere suspicions. Often in science, as well as in the courts of law, most of the information that comes with some facility and confidence is correlation, accompanied by the necessary numerical backing. However, when correlation substitutes or overrides causal reasoning, very odd claims can be supported — as John Sterman demonstrates in his famous and purposely exaggerated 'ice cream and murder' example (Sterman, 2000, pp.141–142).

5 Challenges

There are two stages in explaining phenomena: (a) study measurements and establish correlations through deductive thinking; (b) establish causality through study and inductive thinking. The two can be performed one after the other, in that order. The former option is less 'scientifically risky' than the latter, dealing mainly with numerical data and performing all kinds of statistical tests. The latter is more creative, difficult to prove, but far more interesting as an explanation — in fact, it is the *only* explanation: correlation is merely an exploration.

It appears that in order to work safely with data and correlations, one must have an understanding of causality, or 'how things work' — i.e. at least a tentative explicit causal mental model of reality. The suggestions of correlation can 'prove' just about anything that can be demonstrated with numbers, if not checked for against the mental model of reality: 'does this make sense?' 'is this relationship causal, as expected?' Once again, without proper understanding we may be informed but 'remain in the dark' (Perdicoúlis, 2012a). Most likely this would not be satisfactory for anyone, even for statisticians.

References

Begon, M., J.L. Harper, and C.R. Townsend (1996) *Ecology.* Oxford: Blackwell Science.

Fowler, J., L. Cohen, and P. Jarvis (1998) *Practical Statistics for Field Biology* (2nd ed.). Chichester: John Wiley & Sons.

Hammond, R., and P.S. McCullagh (1978) *Quantitative Techniques in Geography* (2nd ed.). Oxford: Clarendon Press.

Haynes, R. (1996) Use of statistics. In: Watts and Halliwell (1996, pp.67–134)

Perdicoúlis, A. (2013d) The scientific qualifier. *oestros*, **11**.

Perdicoúlis, A. (2013c) On quality. *oestros*, **10**.

Perdicoúlis, A. (2013b) Shadow measurements. *oestros*, **9**.

Perdicoúlis, A. (2013a) People know. *oestros*, **8**.

Perdicoúlis, A. (2012b) Recreating established systems. *Systems Planner*, **10**.

Perdicoúlis, A. (2012a) Information and understanding. *oestros*, **2**.

Perdicoúlis A. (2010) *Systems Thinking and Decision Making in Urban and Environmental Planning.* Cheltenham: Edward Elgar.

Sterman, J.D. (2000) *Business Dynamics.* Boston: Irwin McGraw-Hill.

Watts, S., and L. Halliwell [eds.] (1996) *Essential Environmental Science: Methods & Techniques.* London: Routledge.