University of Waterloo W. H. Cherry

RELATIONSHIPS IN STATISTICS: Association Among Variates and Causation

Statistical Highlight #58 deals with *association* of two *explanatory* variates, like the *focal* variate **X** and a lurking variate (or confounder) **Z**; we now distinguish four reasons ('cases') for such associations, which are also shown symbolically at the right, where an arrow denotes causation.

- ∗ **X** causes **Z**;
- ∗ **Z** causes **X**;
- ∗ $\mathbf{Z}_j$ causes **X** *and* $\mathbf{Z}_i$ – we (unhelpfully) say **X** *and* $\mathbf{Z}_i$ are a **common cause** to $\mathbf{Z}_j$;
- ∗ coincidence [which often means both **X** and **Z** are associated with *time* – *i.e.*, coincidence is often case (3) where $\mathbf{Z}_j$ is time (whatever 'causation' by time means) – recall Note 3 on the overleaf side (page HL62.2) of Statistical Highlight #62].

(1) **X** → **Z** (**X** causes **Z**)

(2) **Z** → **X** (**Z** causes **X**)

(3) $\mathbf{Z}_j$ → **X**, $\mathbf{Z}_j$ → $\mathbf{Z}_i$ ($\mathbf{Z}_j$ causes **X** *and* $\mathbf{Z}_i$)

(4) **X** **Z** (coincidence)

If *extra*-statistical knowledge can rule out coincidence, two (explanatory} variates are associated for only *two* reasons:
○ direct causation [cases (1) and (2)], **OR:** ○ common cause [case (3)].

NOTE: 1. 'Common response' for case (3) above seems *confusing* terminology – **X** and $\mathbf{Z}_i$ are *two* 'response*s*' (not one response) to $\mathbf{Z}_j$. *Common cause* is a more natural description of this causal structure (because it is *why* **X** and $\mathbf{Z}_i$ are associated), although the phrase is already used in process improvement, where Deming distinguished *common* and *special* causes of (excessive) variation in process output. Nevertheless, to emphasize its essential feature, case (3) is sometimes described in these Materials as 'a common cause' with the relevant variate name (here $\mathbf{Z}_j$) included in the phrase.

- A more evocative use of 'common response' would be variate **Y** in case (8) below; regrettably, this is *not* its useage.

The four causal structures above can be extended to include the response variate **Y**; there are now *twelve* cases, in which:

- ∗ **X** and **Y** are associated in all *twelve*;
- ∗ **Z** (or $\mathbf{Z}_i$) and **Y** are associated in the last *nine*.
- ∗ **Z** (or $\mathbf{Z}_i$) and **X** are associated in the last *nine* [except perhaps in case (8)].

In the discussion below, the twelve cases are reduced to eight by assuming extra-statistical knowledge is sufficient to:

- ○ rule out 'coincidence' in case (3), in case (5) [which then becomes case (1)] and case (7);
- ○ enable the adjectives *explanatory* and *response* to be *correctly* applied to the variates **X** and **Y** and so rule out case (2).

(1) **X** → **Y** (**X** causes **Y**)

(2) **Y** → **X** (**Y** causes **X**)

(3) **X** **Y** (coincidence)

(4) **Z** → **X** → **Y** (**Z** causes **X** causes **Y**)

(5) **Z** **X** → **Y** (coincidence and **X** causes **Y**)

(6) **X** → **Z** → **Y** (**X** causes **Z** causes **Y**)

(7) **X** **Z** → **Y** (coincidence and **Z** causes **Y**)

(8) **X** → **Y**, **Z** → **Y** (**X** *and* **Z** cause **Y**)

(9) **Z** → **X**, **Z** → **Y** (**Z** causes **X** *and* **Y**)

(10) **X** → **Z**, **X** → **Y** (**X** causes **Z** *and* **Y**)

(11) $\mathbf{Z}_j$ → **X** → **Y**, $\mathbf{Z}_j$ → $\mathbf{Z}_i$ → **Y** ($\mathbf{Z}_j$ causes $\mathbf{Z}_i$ *and* **X** which cause **Y**)

(12) $\mathbf{Z}_j$ → **X**, $\mathbf{Z}_j$ → $\mathbf{Z}_i$ → **Y** ($\mathbf{Z}_j$ causes **X** *and* $\mathbf{Z}_i$ which causes **Y**)

The diagrams for the remaining eight cases illustrate directly two possibilities and, indirectly, a third:

- **+** **X** and **Y** are *associated* **and** **X** *causes* **Y**: cases (1), (4), (6), (8), (10) and (11);
- **+** **X** and **Y** are *associated* **but** **X** does *not* cause **Y**: cases (9) and (12).

Thus, key statistical issues in association and causation are:

- ∗ if **X** causes **Y** [cases (1), (4), (5), (6), (8), (10) and (11)], **X** and **Y** will be *associated*;
- ∗ if **X** and **Y** are associated [cases (1) to (12)] and coincidence can be ruled out, there *is* causation involving **Y** [all cases except (3)] ***but not necessarily*** by **X** [cases (7), (9) and (12)].

However, the context for the two precepts above – an *observed* association – needs to be extended to one of *no* observed association when, as discussed in Section 4 on page HL60.3 in Statistical Highlight #60:

- **+** **X** and **Y** are *not* associated **but** **X** *does* cause **Y**: case (8) in some situations of confounding by lurking variate **Z**.

The twelve causal structures above illustrate possible association-causation connections but a number of them are *not* relevant in practice to Plans for comparative data-based investigating of an observed **X**-**Y** association.

- Association due to coincidence is seldom of statistical interest, eliminating cases (3), (5) and (7).
  - – Case (7) is also case (8) when the **X**-**Y** relationship is coincidence.
- Correct identification of the response and explanatory variates eliminates case (2).
- All associations can be thought of in terms of causal chains – see the first bullet (●) in Note 1 on page HL62.1 of Statistical Highlight #62 – but investigating other steps in the **X**-**Y** chain is seldom of statistical interest, eliminating cases (4) and (6).
- Case (8) is case (1) with lurking variate **Z** shown explicitly and so is covered under case (1) [and under case (11)].
- Because **Z** is an *explanatory* variate, case (10) is really the causal structure at the right, which is investigated as case (1) or case (11) [see also Note 2 at the top of page HL64.3 in Statistical Highlight #64 and the discussion of Table HL10.4 at the top of page HL10.4 in Statistical Highlight #10].

(10) **X** → **Z** → **Y**, **X** → **Y** (**X** causes **Y** *and* **Z** which causes **Y**)

- Case (12) is both: – case (9) ['common cause'] with an intermediary variate shown in the $\mathbf{Z}_j$-**Y** branch,
  – case (11) for the Question *Is* **X** *a cause of* **Y**? when the Answer is *No.*

(*continued overleaf*)

This leaves cases (1), (9) and (11); we discuss cases (1) and (9) in Section 2 page HL63.1 in Statistical Highlight #63. We pursue them and cases (8) and (11) on pages HL64.1 to HL64.3 in Statistical Highlight #64.
[The structure of case (11) is also of *incidental* interest because of its common cause $\mathbf{Z}_j$ of **X** and $\mathbf{Z}_i$, which could account for **X** and **Z** changing *together* in diagrams (1) to (8) discussed on pages HL60.3 and HL60.4 in Statistical Highlight #60].

The foregoing discussion shows why, in statistics, we distinguish *association* from *causation*: to remind us that, just because we observe (for instance, in a scatter diagram) that **X** and **Y** are *associated*, we can**not** say, without further investigating, that a change in **X** will *bring about* (or *cause*) a change in **Y**.

- Statistical Highlight #66 discusses *correlation* as a measure of the tightness of clustering of the points of a scatter diagram about a straight line; correlation is therefore one way of quantifying magnitude ('strength') of association between **X** and **Y** as seen in a scatter diagram. For this reason, the distinction between association and causation may also be referred to elsewhere as the distinction between correlation and causation, although this wording is better avoided.
- When referring to an **X**-**Y** relationship, phrases used in statistics like *association is not (necessarily) causation* and *correlation is not (necessarily) causation* encompass *three* possibilities:
  - the **X**-**Y** relationship is a *coincidence* – this may pique our curiosity but is seldom of practical importance;
  - **X** and **Y** are *associated* but **X** does not *cause* **Y**;
  - **X** *is* a (or possibly *the*) cause of **Y**.

  Undue emphasis on the second possibility (*e.g.*, in introductory statistics teaching) can obscure three matters:
  + association *does* imply causation if coincidence can be ruled out; BUT:
  + the causation *may* be, but is not *necessarily*, between **Y** and **X**, the variates *observed* to be associated.
  + *Lack* of association of **X** and **Y** does *not* rule out causation of **Y** by **X** – as **X** changes, a confounder **Z** may change in such a way that **Y** remains *un*changed – see diagrams (5) to (8) at the top of page HL60.4 in Statistical Highlight #60.

NOTES: 2. When (a change in) an explanatory variate **U** (a focal variate **X** or a confounder **Z**) *causes* (a change in) a variate **V** (a response variate **Y** or a focal variate **X**), several matters determine the *strength* of the association (as quantified by the correlation, say, of **U** and **V**, if they are *quantitative* variates).

$\mathbf{U} \to \mathbf{V}$
$\mathbf{X} \to \mathbf{Y}$
$\mathbf{Z} \to \mathbf{X}$

- If **U** is the *only* cause of **V** and acts on a time scale that is **short** relative to the period of observation, there is a *high* correlation of **U** and **V**; in the absence of measurement error, the magnitude of r would be 1.
  = An illustration is force **X** causing acceleration **Y**.
- *Weaker* association of **U** and **V** can occur for several reasons, as illustrated by the data for the occurrence of lung cancer **Y** in relation to smoking status **X** in three non-smokers and three smokers in Table HL59.1 at the right. The *strong* ('perfect') association in case (A) can weaken because:

Table HL59.1:

Unit	Smoking status	Lung cancer (A)	(B)	(C)
1	Non-smoker	No	No	No
2	Non-smoker	No	No	No
3	Non-smoker	No	Yes	No
4	Smoker	Yes	Yes	No
5	Smoker	Yes	Yes	Yes
6	Smoker	Yes	Yes	Yes

  - one *non*-smoker in case (B) acquired lung cancer from **another cause** (*e.g.*, asbestos inhalation);
  - the smoker with*out* lung cancer in case (C) may: yet develop lung cancer, OR: die before doing so, OR: be in a population subgroup for which **X** does *not* cause **Y**;

  the first two possibilities have a time scale for causation that is **long** relative to the period of observation and the third involves our **definition of causation** (at the top of page HL62.1 in Statistical Highlight #62) in terms of an *attribute*.

In these ways, we account for differing strengths of *association* observed in *causal* **X**-**Y** relationships or, expressed another way, we account for why (a change in) **X** *causes* (a change in) **Y** but, for *some* population elements:
- **Y** changes when **X** does *not* change (*e.g.*, some *non*-smokers get lung cancer), OR:
- **Y** does *not* change when **X** changes (*e.g.*, some smokers do *not* get lung cancer).

3. Association is a straight-forward idea (we can *see* it), causation much less so; the two causal structures at the right [cases (8) and (9) from page HL59.1 overleaf] give insight into their difference. As discussed in Note 4 on page HL62.2 in Statistical Highlight #62, under our definition of causation at the top of page HL62.1 in Highlight #62:

(8) $\mathbf{X} \to \mathbf{Y} \leftarrow \mathbf{Z}$

(9) $\mathbf{X} \leftarrow \mathbf{Z} \to \mathbf{Y}$

- in the causal structure of case (8) [a common *response* **Y**], there is *association* of **X** and **Y** and of **Z** and **Y** but *no* necessary association of the (unconnected) causes **X** and **Z**; BUT:
- in the causal structure of case (9) [a common *cause* **Z**], there is *association* of **Z** and **Y** and of **Z** and **X** so there is *necessarily* association of **Y** and **X**.

The *difference* between the two structures lies in the *direction* of the arrows denoting causation – if their direction is *reversed* in either diagram, they are the *same* causal structure, apart from the variate names. *Our* definition of causation thus suggests that causation is *directed association*, although it is questionable whether this (model) concept provides much insight into the *real world* difference between association and causation.

Cases (8) and (9) and three other similar causal structures are compared in Statistical Highlight #65.