University of Waterloo W. H. Cherry

# RELATIONSHIPS IN STATISTICS: Connections Among Three Variates

For *three* variates [two explanatory (**X** and **Z**) and one response (**Y**)] involving *two* causal relationships, there are *five* causal structures, as shown in the first two columns of Table HL65.1 at the right below; the first structure has *two* contexts, making six lines in the Table. The structures are five of the twelve cases given near the middle of page HL59.1 in Statistical Highlight #59 [cases (4), (6), (8), (9) and (10)] plus case $(1)_2$ from the upper right of page HL64.4 in Statistical Highlight #64 [also discussed in the second bullet (•) of Note 7 on page HL60.4 in Statistical Highlight #60]. [A reminder of two definitions from pages HL57.2 and HL68.1 in Statistical Highlights #57 and #68 is:

∗ **Confounding:** differing distributions of values of one or more *non*-focal explanatory variate(s) among two (or more) groups of elements or units [like (sub)populations or samples] with different values of the focal variate.

∗ **Interaction** of two factors $\mathbf{X}_1$ and $\mathbf{X}_2$ is said to occur when the effect of one factor on a response variate **Y** depends on the level of the other factor. Interaction means the combined effect of two factors is *not* the sum of their individual effects.]

Several matters are noteworthy.

| | **Variate *causal* connections** | **Table HL65.1** **Name** | **Association** **X-Y** | **X-Z** | **Z-Y** | **Causation** **X-Y** | **X-Z** | **Z-Y** |
|---|---|---|---|---|---|---|---|---|
| (8) | $\mathbf{Z} \to \mathbf{Y}$, $\mathbf{X} \to \mathbf{Y}$ ≡ $\mathbf{X}, \mathbf{Z} \to \mathbf{Y}$ | Confounding (type 2a) [A common response **Y**] | Yes | Yes | Yes | Yes | No | Yes |
| $(1)_2$ | $\mathbf{Z} \to \mathbf{Y}$, $\mathbf{X} \to \mathbf{Y}$ ≡ $\mathbf{X}_1, \mathbf{X}_2 \to \mathbf{Y}$ | Interaction | Yes | --- | Yes | Yes | No | Yes |
| (9) | $\mathbf{X} \leftarrow \mathbf{Z} \to \mathbf{Y}$ ≡ $\mathbf{Z} \to \mathbf{X}, \mathbf{Y}$ | Confounding (type 2b) [A common cause **Z**] | Yes | Yes | Yes | No | Yes | Yes |
| (10) | $\mathbf{X} \to \mathbf{Z}$, $\mathbf{X} \to \mathbf{Y}$ ≡ $\mathbf{X} \to \mathbf{Z}, \mathbf{Y}$ | A common cause **X** | Yes | Yes | Yes | Yes | Yes | No |
| (6) | $\mathbf{X} \to \mathbf{Z} \to \mathbf{Y}$ ≡ $\mathbf{X} \to \mathbf{Z} \to \mathbf{Y}$ | Causal chain **XZY** | Yes | Yes | Yes | Yes | Yes | Yes |
| (4) | $\mathbf{Z} \to \mathbf{X} \to \mathbf{Y}$ ≡ $\mathbf{Z} \to \mathbf{X} \to \mathbf{Y}$ | Causal chain **ZXY** | Yes | Yes | Yes | Yes | Yes | Yes |

○ What tends to distinguish the cases is the pattern of the *causal* relationships in the last column of Table HL65.1 – as shown in the *third* column, each variate is *associated* with each of the other two, except in the case of interaction, when the **X**-**Z** (*i.e.*, the $\mathbf{X}_1$-$\mathbf{X}_2$) relationship is not relevant.

○ *Confounding* and *interaction* have similarities (+) and differences (–).

- \+ both involve two explanatory variates which cause a response variate;
- \+ both have the same pattern of causal relationships and (except as noted above) associations in Table HL65.1;
- – their focus is different:
  - ⊙ confounding is concerned with the impact on investigating the **X**-**Y** relationship of (unwanted) *changes* in (confounder) **Z** as (focal variate) **X** changes;
  - ⊙ interaction is concerned with the impact on the $\mathbf{X}_1$-**Y** relationship *and* the $\mathbf{X}_2$-**Y** relationship of the *value* of (focal variates) $\mathbf{X}_2$ and $\mathbf{X}_1$ respectively (see also Notes 1 and 2 overleaf on the lower half of page HL65.2).

  The same components but different focus of confounding and interaction are somewhat reminiscent of the conditioning-ignoring distinction discussed in Note 3 on pages HL60.1 and HL60.2 in Statistical Highlight #60 because, in each case, statistical *mis*handling provides opportunity for comparison error to impose unnecessary (and so, possibly unacceptable) limitation on Answers.

The *manifestation* of confounding as comparison error may be (adversely) affected *if* there is *interaction* of **Z** and **X** – for example, (puzzling) 'inconsistencies' may be exhibited in the **X**-**Y** relationship – see Table HL60.1 and the discussion on pages HL60.3 and HL60.4 in Statistical Highlight #60.

○ The possibility of **Z** as *a common cause* of **X** and **Y** [case (9)] (so-called **common response**) is relevant when establishing the reason for an **X**-**Y** association, as discussed on pages HL60.2 and HL60.3 in Statistical Highlight #60.

- – The common cause **Z** responsible for *mis*identifying **X** as a cause of **Y** is our type 2b confounding – see the discussion near the top of page HL3.2 in Statistical Highlight #3 and on the lower half of page HL60.2 in Statistical Highlight #60.
  - \+ From *this* perspective, **Z** as a common cause of **X** and **Y** could be regarded as an extreme case of our type 2a confounding where **Z** is *solely* responsible for the change in **Y** as **X** changes.

○ **X** as *a common cause* of **Z** and **Y** [case (10)] is really the causal structure at the right below, because **Z** is an *explanatory* variate; thus, elsewhere we consider case (10) to involve *three* causal relationships, *not* two as in this Highlight #65. We have seen in earlier Highlights that case (10) with *three* causal relationships:

(10) $\mathbf{X} \to \mathbf{Z} \to \mathbf{Y}$, $\mathbf{X} \to \mathbf{Y}$

- – is *not* a viable basis for a comparative Plan, as discussed in Note 2 on page HL64.3 in Statistical Highlight #64;
- – can result in *biased* estimating of a treatment effect, as illustrated on pages HL10.3 and HL10.4 in the discussion of Table HL10.4 in Statistical Highlight #10.

○ As discussed on page HL62.1 in Note 1 in Statistical Highlight #62 (and also in the middle of page HL61.1 in Statistical Highlight #61), we think of causation of **Y** by **X** as proceeding via a (long) **causal chain** of explanatory variates leading to the response of interest. The Question context identifies (arbitrarily) *one* (focal) variate (**X**) in this chain as being of interest, but we recognize that this variate is *preceded* by and followed by other 'focal' variates; the context also (arbitrarily) defines the *end* of the chain in terms of a particular *response* variate (**Y**). However, this response can become part of an *explanatory* variate chain if a different Question context identifes a *different* (later) response variate. From this perspective:

- – The causal chain of case (6) is merely the upper branch of the (real) causal structure of case (10) shown above at the right;
  - ⊙ case (6) reminds us to distinguish **X** causing **Y** *via* **Z** from **X** and **Z** as *separate* causes of **Y** [case (8)].



(*continued overleaf*)

  – The causal chain of case (4) is really case (1) [= case (8)] – **Z** in case (4) is merely an explanatory variate *preceding* the focal variate **X** in the causal chain and so is (generally) of no statistical interest in the Question context.

The (surprising) number of statistical issues arising with relationships among *only three* variates is further complicated if the **X**-**Z**-**Y** relationship is modelled mathematically; such a model (for use in the Analysis stage of the FDEAC cycle) needs to consider:

- the *form* in the model (*e.g.*, first power, second power, square root, logarithm, product) of **X** and **Z**;
- the *distribution* of **Z** (*e.g.*, its mean and standard deviation) [and perhaps of **X**];
- the *relationship* of **X** and **Z** (*e.g.*, their correlation).

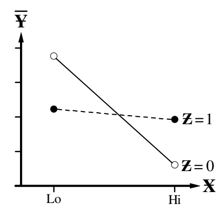
*Association* of (focal variate) **X** and (confounder) **Z** in case (8) is one feature of confounding, a source of comparison error and limitation on Answers from comparative Plans; however, this association (*e.g.*, non-zero correlation) of confounded variates is really an *incidental* feature of the phenomenon – association in the *usual* state of affairs for variates that change together.

- ○ *Zero* correlation of confounded variates is usually introduced by the investigator(s) – for instance, in a factorial treatment structure [see also the brief discussion of diagram (5) just before Note 1 at the bottom of page HL58.2 in Statistical Highlight #58].

Similarly, association among variates in the structural component (on the right-hand side) of a response model [like equation (HL57.1) on page HL57.1 in Statistical Highlight #57] is also a source of such limitation, manifested as *uncertainty* in the estimates of model parameters [*e.g.*, $\beta_1$ – the treatment effect for **X** – in equation (HL57.1)].

- ○ This uncertainty becomes apparent from **stepwise** model fitting, a process to assess [*e.g.*, based on the coefficient of multiple determination, a measure of the proportion of the variation in **Y** accounted for by the fitted model] *which* explanatory variates to include in the model. For instance, in the case of two (focal) variates $\mathbf{X}_1$ and $\mathbf{X}_2$, three models are fitted – one with *both* variates, one with $\mathbf{X}_1$ only and one with $\mathbf{X}_2$ only. The stronger the association (in the data) of $\mathbf{X}_1$ and $\mathbf{X}_2$, the greater the likely difference in the estimates of their coefficients $\beta_1$ and $\beta_2$ among the three models.
  – In the extreme situation where two variates $\mathbf{X}_i$ and $\mathbf{X}_j$ have correlation of magnitude 1 (*i.e.*, $\mathbf{X}_i$ and $\mathbf{X}_j$ are the *same* variate statistically), the model fitting process cannot be achieved computationally – the design matrix is not of full rank and so cannot be inverted.

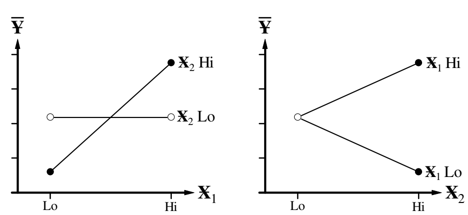
**NOTES:** 1. An idea, associated with the name of Taguchi, for *exploiting* interaction is illustrated by improvement of a process for manufacturing ceramic tiles; the diagram at the right for an **X**-**Y** relationship displays an interaction effect, because the *slope* of the (linear) relationship between **X** and the *average* of **Y** is *different* (here, smaller negative magnitude) when (non-focal) explanatory variate **Z** = 1 ('Hi') than when **Z** = 0 ('Lo'). In the tile-manufacturing process, if:




- **Y** is tile size *after* firing in an oven,
- **X** is oven temperature, whose variation from 'Lo' to 'Hi' over position within the oven causes tiles of the *same* initial size, but fired in different oven positions, to have different *final* sizes,
- **Z** is amount of clay in the ingredient mix used for the tiles,

by managing the amount of clay in the ingredient mix (*i.e.*, setting **Z** = 1), the manufacturing process is improved by making variation in tile final size *less* sensitive to variation in firing temperature due to tile position within the oven. This *indirect* approach exploiting interaction avoids the (more expensive) *direct* approach of making the temperature more uniform within the oven; of course, the properties of the tiles must remain acceptable when **Z** = 1 and clay must not be too expensive an ingredient.

2. The (equivalent) diagrams at the right show the effects of two (binary) focal variates $\mathbf{X}_1$ and $\mathbf{X}_2$ on (the average of) **Y**; one focal variate is on the horizontal axis of a diagram, the other distinguishes the two lines by its level.




- The *non*parallel lines show there is an $\mathbf{X}_1$-$\mathbf{X}_2$ *interaction.*
- The left-hand diagram shows that $\mathbf{X}_1$ and **Y** are *independent* when $\mathbf{X}_2$ is Lo (the relevant line has *zero* slope) but *not* when $\mathbf{X}_2$ is Hi.
- The right-hand diagram shows there is *no* independence of $\mathbf{X}_2$ and **Y** – *neither* line has zero slope.

  For these two dagrams, it is *wrong* to involve *conditional (probabilistic) independence* because it implies that we have undertaken the difficult task of formulating a probability model for this situation – see also Section 13 on page HL89.18 in Statistical Highlight #89.

Thus, different variate contexts can yield different assessments of (complicated) relationships.

Notes 1 and 2 (from pages HL68.2 and HL51.8) have more context in the respective Statistical Highlights #68 and #51.

In introductory statistics courses, emphasis on comparative Plans with *one* focal variate, together with similarities of confounding and interaction when there are three variates, should not be allowed to obscure the continuing importance of possible confounding in comparative Plans with *two or more* focal variates. With *three* variates and possible confounders $\mathbf{Z}_i$, $\mathbf{Z}_j$ and $\mathbf{Z}_k$, statistical issues like those in the foregoing discussion may arise for connections among:

○ $\mathbf{X}_1$, $\mathbf{Z}_i$ and **Y**, ○ $\mathbf{X}_2$, $\mathbf{Z}_j$ and **Y**, AND ○ $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{Z}_k$ and **Y** [$\mathbf{Z}_k$ may be *a common cause* of $\mathbf{X}_1$, $\mathbf{X}_2$ *and* **Y**].