University of Waterloo W. H. Cherry

RELATIONSHIPS IN STATISTICS: Causation – Statistical Issues

To define ***formally*** in statistics what it means to say (a change in) **X** *causes* (a change in) **Y** in a ***target*** population, we state three criteria (useful in practice when establishing causation or quantifying the effect of **X** on **Y**):

(1) **LURKING VARIATES:** Ensure *all other* explanatory variates $\mathbf{Z}_1$, $\mathbf{Z}_2$, ..., $\mathbf{Z}_k$ hold their (same) values for *every* population element when $\mathbf{X} = 0$ and $\mathbf{X} = 1$ (sometimes phrased as: *Hold all the* $\mathbf{Z}_i$ ***fixed*** *for* .....).

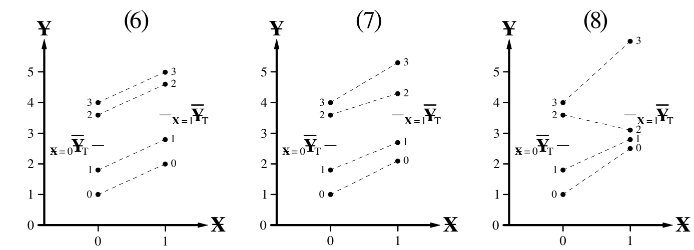
(2) **FOCAL VARIATE:** Observe the population **Y**-values, and calculate an appropriate attribute value, under *two* conditions:
- ⊙ with *every* element having $\mathbf{X} = 0$;
- ⊙ with *every* element having $\mathbf{X} = 1$.

(3) **ATTRIBUTE:** Attribute$(\mathbf{Y}$, perhaps some of $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k \,|\, \mathbf{X} = 0) \neq$ Attribute$(\mathbf{Y}$, perhaps some of $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k \,|\, \mathbf{X} = 1)$;
those of $\mathbf{Z}_1$, $\mathbf{Z}_2$, ..., $\mathbf{Z}_k$ *included* in the attribute will have the *same* values when $\mathbf{X} = 0$ and $\mathbf{X} = 1$ under (1).

The notation $\mathbf{X} = 0$ and $\mathbf{X} = 1$ for values of the focal variate is *symbolic* – 0 and 1 *represent* two *actual* values of **X** in a particular context; actual values of the focal variate are set in the **protocol for setting levels**, discussed in Statistical Highlight #68.

Three illustrations, involving only *one* lurking variate **Z**, of this formal definition are given at the right below for a target population of 4 elements with respective **Z** values (shown beside the points) of 0, 1, 2 and 3.

- ○ In diagram (6), **Y** values increase by 1 as **X** changes from 0 to 1 and, correspondingly, the *average* of **Y** (indicated by a short horizontal line) increases by 1 from 2.6 to 3.6.
- ○ In diagram (7), the **Y** values again increase as **X** changes from 0 to 1 but by *differing* amounts.
- ○ In diagram (8), three **Y** values *in*crease but one *de*creases as **X** changes, although the *average* of **Y** again increases by 1 from 2.6 to 3.6.



In contrast to diagrams (1) to (5) on the overleaf side (page HL58.2) of Statistical Highlight #58 where there *is* confounding, diagrams (6) to (8) illustrating our definition of causation have (of course) *no* confounding – the values of **Z** do *not* change as **X** changes, so there is *no* **X**-**Z** association (*zero* **X**-**Z** correlation). Also, the $\overline{\mathbf{Y}}$s have a subscript T denoting 'target population'.

NOTES: 1. The first two of the three criteria given above, which *we* take as a formal definition of causation in a *target* population, are *idealizations* – no Plan can fully satisfy these two criteria in practice. For example:

- For the Question: *Does smoking cause lung cancer?*, we can think of a (long) **causal chain** of explanatory variates leading to the response of interest (here, *lung cancer status*). The Question identifies (arbitrarily) *one* variate in this chain (here, *smoking status*), but we recognize that this variate is *preceded* by 'focal' variates (factors that caused the individual to decide to smoke) and it is *followed* by others [factors that describe the damage (at a cellular level, say) that is ultimately manifested as cancer]. When 'lurking variates' criterion (1) refers to *ensuring all other explanatory variates hold their (same) values for every target population element*, it does *not* include variates in the causal chain involving the 'main' focal variate.
  - The Question identifies one (focal) *explanatory* variate in the causal chain as being of interest; it also (arbitrarily) defines the *end* of the chain in terms of a particular *response* variate. However, this response can become part of an *explanatory* variate chain if a different Question identifes a *different* (later) response variate – for instance, *alive* or *dead* instead of *lung cancer status* in our example.
  - When a causal chain loops back upon itself, the situation is commonly referred to as **feedback**.
- In 'focal variate' criterion (2), the ideal of observing *all* elements of the target population under each of *two* values of the focal variate is attained more closely in practice in an *experimental* Plan – the two samples to which the investigator(s) assign equiprobably the two values of the focal variate stand in for the respondent population (and, hence, at two stages removed, for the target population) under the two values.
  - In an *observational* Plan, the two values of the focal variate define *sub*populations of the respondent (and the study) population and the two samples with the two values of the focal variate stand in only for these *sub*populations; this matter is discussed in Section 1 on the first side of Statistical Highlight #9.
  - In some investigations, there may, of course, be *more than* two focal variate values of interest.
  - Coming closer to meeting criterion (2) is one reason why an *experimental* Plan is preferred, where feasible.
- 'Attribute' criterion (3) defines causation in terms of (a change in) an *attribute*, not of individuals – this is consistent with the predominant concern of statistics with *populations*, not elements. A consequence of criterion (3) is that **X** need not bring about a change in **Y** for *every* element of the population for us to say **X** *causes* **Y**.
  - A rationalization of this departure from the intuitive idea that causation *always* produces an effect is [like criterion (1)] in terms of non-focal explanatory variates $\mathbf{Z}_i$ – there may be elements with (some) such variate(s) whose value(s) have the consequence that a change in **X** does *not* bring about a change in **Y**; we would normally think of these elements as being a *small* proportion of the population.

(*continued overleaf*)

NOTES: (cont.) 1. ● – + An illustration is there *may* be individuals for whom smoking would *never* bring about lung cancer; at our present level of (genetic) knowledge, we cannot identify such individuals (if they exist) but it is still good public health policy to discourage smoking in light of the known lung cancer *rates* for non-smokers and smokers.

There is further discussion of *statistical* issues involving causation in the previous Statistical Highlight #61.

2. Our three criteria overleaf at the top of page HL62.1 defining causation are framed in terms of the (target) *population* and an appropriate *attribute*, **not** elements and their variates. Criterion (1) specifies *all* non-focal explanatory variates (our $\mathbf{Z}$s) remain fixed; three approaches try to meet this criterion to manage comparison error in practice:
   - hold *some* $\mathbf{Z}$s fixed *physically* by blocking, matching or subdividing (see the following Statistical Highlight #63);
   - under probability assigning of elements' focal variate values, use statistical theory to manage *under repetition* differences among unblocked, unmeasured and unknown $\mathbf{Z}$s (see Statistical Highlight #10);
   - use a *response model* in the Analysis stage of the FDEAC cycle to hold some $\mathbf{Z}$s fixed *mathematically*, but even a quite elaborate model, like equation (HL57.1) on the first side of Statistical Highlight #57, cannot involve *all* possible $\mathbf{Z}$s in its structural component, and only those k variates included are reflected in the interpretation of β_1, the model coefficient of the *focal* variate. [The *stochastic* component of a response model like that in equation (HL57.1) deals mathematically with effects on $\mathbf{Y}$ of $\mathbf{Z}$s *not* included in the structural component.]

   The challenge in investigating statistical relationships is to come close enough to the ideal represented by the three criteria to obtain an Answer with limitations whose level of severity is acceptable in the Question context.

3. For 'focal variate' criterion (2), there are focal variates (like age and sex) whose values can*not* be *assigned* to elements by the investigator(s) in an experimental Plan. For such variates, we avoid the stronger language of saying *increasing age* ***causes*** *loss of visual acuity* in favour of *increasing age is* ***associated with*** *loss of visual acuity.*
   - Such associations are important in contexts like discrimination by sex or race where, for example, we compare the relevant population proportion with the proportion of women or a racial group in an employment or other category. *Causation* (in the sense of our three criteria) by sex or race is not the issue with such associations, because there is no intention to change the value of the focal variate.
     - We may also speak of the *reason* (rather than the *cause of*) why a population subgroup is under- or over-represented – for example, in an employment context we may consider relevant *qualifications.*
   - Some focal variates (like cigarette smoking) cannot *ethically* be assigned to human elements, which imposes limitations that arise from using animal elements in an experimental Plan or human elements in an observational Plan.

   These matters are pursued in a discussion of Simpson's Paradox in Statistical Highlight #51.
   - The ideal of criterion (2) ignores any *time* difference between the realization of the two conditions $\mathbf{X} = 0$ and $\mathbf{X} = 1$. In actual investigations, the two groups (usually samples) with elements having $\mathbf{X} = 0$ and $\mathbf{X} = 1$ are observed concurrently but, in a cross-over Plan (like the oat bran investigation described in Note 3 on page HL9.6 in Statistical Highlight #9), there *is* a time difference between $\mathbf{X} = 0$ and $\mathbf{X} = 1$ for both half samples; any changes in elements' *other* explanatory variates values over time may then be a source of comparison error.

4. 'Attribute' criterion (3) involves different attribute values for different values of the focal variate (but with relevant $\mathbf{Z}_i$s remaining the *same*); our definition therefore implies that if $\mathbf{X}$ *causes* $\mathbf{Y}$, there is *association* of elements' $\mathbf{X}$ and $\mathbf{Y}$ values over the target population under the two values of $\mathbf{X}$; we *hope* this association carries over into the study population, the respondent population and the sample.
   - If a cause has *more than one* effect (*e.g.*, smoking is a cause of several different cancers), 'attribute' criterion (3) must be broadened to include inequality of the attributes of *all* the relevant response variates. Extending the preceding argument for *one* response, the values of these (several) response variates will each be associated with the values of $\mathbf{X}$ over the elements of the target population under the two values of $\mathbf{X}$; the values of these $\mathbf{Y}$s with the common cause $\mathbf{X}$ will *also* be associated.

   This causation-association connection under our definition of causation in statistics is used in Statistical Highlight #60.

5. An example of the caveat in 'attribute' criterion (3) is: when using least squares estimates [equation (HL62.1) at the right] to *compare* simple linear regression *slopes*, the z values must be the *same* when $\mathbf{X} = 0$ and $\mathbf{X} = 1$.

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} y_j(z_j - \bar{z})}{\sum_{j=1}^{n}(z_j - \bar{z})^2} \quad \text{-----(HL62.1)}$$

6. Ideas about investigating $\mathbf{X}$-$\mathbf{Y}$ relationships are summarized at the right in Table HL62.1.

Table HL62.1: Summary of Ideas About Investigating $\mathbf{X}$-$\mathbf{Y}$ Relationships

Criterion (1): the ideal	Ensure all the $\mathbf{Z}_i$ hold their (same) values for every population element when $\mathbf{X} = 0$ and $\mathbf{X} = 1$
Criterion (3)	For causation, a relevant *attribute* must differ in value when $\mathbf{X} = 0$ and $\mathbf{X} = 1$
Confounding	Confounding arises when one or more of the $\mathbf{Z}_i$ change in value when $\mathbf{X} = 0$ and $\mathbf{X} = 1$
Comparison error	A difference, due to confounding, from the *real* or *intended* value of an *attribute* of a relationship.

   - The *difference* in attribute values in criterion (3) must be such as to be *practically important* in the Question context.
   - A danger of appropriating 'confounding' as statistical terminology is that a word for *failure* to meet criterion (1) may shift the focus away from this overriding ideal.