Figure 10.6. CAUSATION: An Introduction

When we hear that a house fire was *caused* by an electrical fault, most people have little trouble understanding what is being said or recognizing the implication that, if the electrical fault had *not* occurred, there would have been no fire. It is tempting to conclude from such every-day experiences that causation is a straight-forward idea but, when we try to make more precise the notion of what is meant by causation, complexities arise; an overview of the issues is given in the *Encyclopædia Britannica*, 1957 Edition, William Benton, Chicago, Volume 5, pages 60-63 (*Causality or Causation*). The matters below are not a definitive discussion but are useful when dealing with *statistical* aspects of causation.

- We think initially of a change in a *response* variate $\mathbf{Y}$ (*i.e.*, an effect) being *caused* by a change in a *focal* (explanatory) variate $\mathbf{X}$; in the illustration above, $\mathbf{Y}$ is the fire status of the house (not on fire or on fire) and $\mathbf{X}$ is the status of its electrical system (working properly or faulty) and we say (a change in) $\mathbf{X}$ *caused* (a change in) $\mathbf{Y}$.
  - There may also be *lurking* explanatory variate(s) $\mathbf{Z}$ (or $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k$).
- Causes of $\mathbf{Y}$ *other than* $\mathbf{X}$, if they exist, may need to be considered.
- When *coincidence* can be ruled out as a reason, an $\mathbf{X}$-$\mathbf{Y}$ *association* indicates *causation* of $\mathbf{Y}$ but ***not necessarily*** by $\mathbf{X}$.
  - Stating this precept as *association is not causation* or *correlation is not causation* is more easily *mis*understood.

  The association-causation distinction is a revered topic in many statistics courses.
- In a causal relationship, if the *cause* is absent, the response does not occur (from that cause); if the *response* is absent, the cause *may* still be present.
- We may need to consider the *time scale* of a causal relationship – for example, there is a *short* time scale for the movement of a bat causing a baseball to move, there is a *long* time scale for cigarette smoking causing lung cancer. Implicit in the idea of a time scale is that the cause always *precedes* the occurrence (if any) of the response.
  - Even in something as simple as the baseball illustration, it is easy to digress into a discussion of what *caused* the batter to hit the ball, by which we usually mean the *reasons* (s)he did so; such discussion would then probably invoke the behaviour and characteristics of both the pitcher and the batter (among other things), quite *different* issues from those in our illustration and generally not relevant to our present concerns.
- There is a level at which any causal relationship can be seen to be a multi-step sequence of events (a causal chain); this sequence is usually complex and imperfectly understood, but may help us rationalize the time scale we observe for the causation.
  - For a bat causing a baseball to move, the sequence is at a molecular level.
  - For cigarette smoking causing lung cancer, before we go to a molecular level, there is the sequence of changes induced in lung tissue by the chemicals in tobacco smoke [incidentally, such changes explain why some *former* smokers get lung cancer *after* the cessation of smoking as a cause].
- A cause may *not* produce a response if:
  - the *intensity* of the cause is too low – a bat may not hit a ball hard enough to make it move appreciably;
  - the time scale is *long* in relation to the period of observation; this idea can be called *censoring* – a smoker may not be observed to get lung cancer because (s)he dies of some *other* cause *before* the lung cancer occurs.

  The two factors may impinge on each other in that higher intensities of the cause may shorten the time scale; this is related to the idea of *dose-response* – the incidence of lung cancer *in*creases with *level* of cigarette consumption.
- The discussion of causation may need to be framed somewhat *differently* in the two cases where:
  - the cause makes the response *happen* – a bat causes a baseball to move;
  - the cause *prevents* the response from happening – seatbelts cause a *reduction* in fatalities in automobile accidents.

○ To define *formally* in statistics what it means to say $\mathbf{X}$ *causes* $\mathbf{Y}$ in a (target) population, we state three criteria:

(1) **LURKING VARIATES:** Ensure *all other* explanatory variates $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k$ hold their (same) values for *every* population unit 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 $\mathbf{Y}$-values, and calculate an appropriate attribute value, under *two* conditions:
- ⊙ with *every* unit having $\mathbf{X} = 0$;
- ⊙ with *every* unit 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).

For example, the z values must be the *same* when using least squares estimates [as given in equation (10.6.1) at the right] to *compare* simple linear regression *slopes* when $\mathbf{X} = 0$ and $\mathbf{X} = 1$.

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} y_j(z_j - \overline{z})}{\sum_{j=1}^{n}(z_j - \overline{z})^2} \qquad \text{-----(10.6.1)}$$

These criteria are framed in terms of the (target) *population* and an appropriate *attribute*, **not** units and their variates.

- The combination of other causes, a long time scale and censoring may make it *difficult* to establish causation; for example, dietary fat (as the proportion of the caloric intake it contributes to the diet) has been implicated as a cause of breast cancer in women. However, dietary fat is still only classed officially as a *risk factor* for breast cancer; the same idea is conveyed by saying fat is a *weak* cause of breast cancer in women, but *risk factor* is preferred wording.
  - *Measuring* issues can also impede establishing causation – for instance, when trying to establish possible harmful effect(s) of watching 'undesirable' TV programs, how do we quantify the 'harmful effect(s)' and 'undesirability' of a program?

(*continued overleaf*)

- The three criteria overleaf on page 10.11 imply that an Answer about causation from an *exprimental* Plan will usually have (substantially) fewer limitations due to comparison error than an Answer from an *observational* Plan.
- Interest in causation in statistics is seldom limited merely to whether **X** *causes* **Y**; more usual concerns are with the *direction* and/or the *extent* of the relationship; for example, does an increase in **X** cause an *in*crease or a *de*crease in the attribute for **Y** and/or by how much (or by what proportion) does the attribute value change when **X** = 0 and **X** = 1.
- The notation **X** = 0 and **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 values**.

The four numbered statements below, and the points which follow, illustrate the foregoing matters in relation to statistical issues involving causation; the statements are worded concisely by excluding assumptions or context – for example:
+ in 2, sex is assumed to be the *only* cause of pregnancy;
+ in 3, smoking refers to *cigarette* smoking;
+ in 4, seat belt usage is considered in the context of *automobile* accidents.

Also, outside the present discussion, the word *cause* might not be used – for instance, statement 4 might be *seat belts reduce fatalities.* Characteristics of the causal relationships described in the four statements are summarized in Table 10.6.1 at their right.

1. Force causes acceleration.
2. Sex causes pregnancy.
3. Smoking causes lung cancer.
4. Seat belts cause a reduction in fatalities.

Table 10.6.1:

#	Cause X	Response Y	Time scale	Other causes?
1	force (quantitative)	acceleration (quantitative)	short	no
2	sex (categorical)	pregnancy (categorical)	intermediate	no
3	smoking (quantitative)	lung cancer (categorical)	long	yes
4	seat belt (categorical)	fatality (categorical)	intermediate	yes

- When the response **Y** is *absent*, it does *not* imply the cause **X** is absent; specifically, we recognize that:
  - no acceleration does *not* imply no force – buildings and bridges are, of course, engineered to withstand (*i.e.*, *not* accelerate under) *some* of the forces they usually experience;
  - no pregnancy does *not* imply no sex;
  - no lung cancer does *not* imply a non-smoker;
  - death in a car accident does *not* imply a person was not wearing a seat belt.
- When the cause **X** is *absent*, the response **Y** will *also* be absent *if* **X** is the *only* cause of **Y** (statements 1 and 2) but *not* necessarily if there are *other* causes (statements 3 and 4).
  - However, if a person has lung cancer, under current estimates that around 90% of lung cancer is due to smoking, it as highly *probable* the person has been (or is) a smoker.
- Establishing or recognizing a causal relationship is usually *easier* when the time scale is *short* and it tends to become *harder* as the time scale gets *longer*; loss of *biological* units from observation (*e.g.*, due to death) compounds this difficulty.
  - *Short* time scales are more common for causation in the physical sciences, where causality is often relatively unambiguous on the basis of even small numbers of instances.
  - *Long* time scales are common in the biological and medical sciences, and causality must then routinely be approached on the basis of relatively large aggregates of units and an attribute of these aggregates.

  This may imply that the steps in the multi-step sequence (or causal chain) tend to be more *numerous* and/or to proceed more *slowly* in biological than in physical systems.
- Specific factors may complicate the process of establishing causation in particular instances.
  - Cancer due to smoking occurs at sites (*e.g.*, the urinary bladder) *remote* from the site directly exposed (the lungs).
  - Seat belts do not prevent *all* fatalities and they occasionally *cause* fatalities that might otherwise not have occurred.
  - A skeptic could argue that seat belts do not *themselves* cause fewer fatalities but, rather, wearing them reminds people to drive more carefully; even if this were true, it does *not* remove the causation or mitigate against the desirability of wearing seat belts.
- Five criteria used to establish smoking as a *cause* of lung cancer (from an assessment of the information from over 6,000 investigations, predominantly with *observational* Plans) in the 1964 U.S. Surgeon General's Report are given in Program 11 of *Against All Odds: Inside Statistics* (about 22 minutes into the video) as questions about characteristics of the association:
  - **consistency:** do the different methods of studying the association provide *consistent* results?
  - **strength:** are the lung cancer rates for smokers *much* greater than those for non-smokers?
  - **specificity:** can smoking habits be predicted from lung cancer incidence and can lung cancer incidence be predicted from smoking habits?
  - **temporal relationship:** does the presumed cause (smoking) always *precede* the presumed effect (lung cancer)?
  - **coherence:** does the association make sense in light of what we know about the history and biology of the disease?

NOTE: The notation used in this Figure is **X** for the *focal* variate and **Z** for other *non*-focal explanatory ('lurking') variates; elsewhere, you may see the meaning of **X** and **Z** interchanged. There can also be *more than one* focal variate.