Figure 10.7. ASSOCIATION AND CAUSATION IN STATISTICS: Reality and Illusion

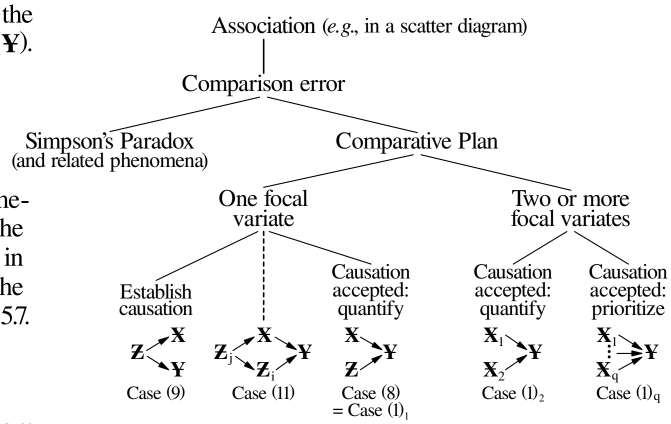
1. Data Visualization

A precept of investigating relationships in statistics is *looking at the data*; that is, *visualizing* the relationship – for example, the course STAT 442 has the title *Data Visualization*. A caveat to this precept is that deciding *how* to visualize the data and how to *interpret* the display(s) must be informed by adequate statistical understanding – seeing may be illusory. In this Figure 10.7, we deal in turn with the three main types of Question which arise when investigating statistical relationships:

1. Is there an *association* of **X** and **Y** and, if so, what is its *form* and/or *magnitude* and/or *direction*?
2. What is the *reason* for an association – *e.g.*, can we *establish* that **X** *causes* **Y**?
3. *Accepting* a relationship as causal: what is its *form* – *e.g.*, what is the *effect* of **X** on (the average of) **Y**?
   which explanatory variate is the *most important* cause of (variation in) **Y**?

This 'natural' order of such Questions may differ appreciably from the one in which they are discussed in an introductory course; also, the emphasis in introductory discussion is usually on the relationship of *one* focal variate (**X**) to *one* response variate (**Y**).

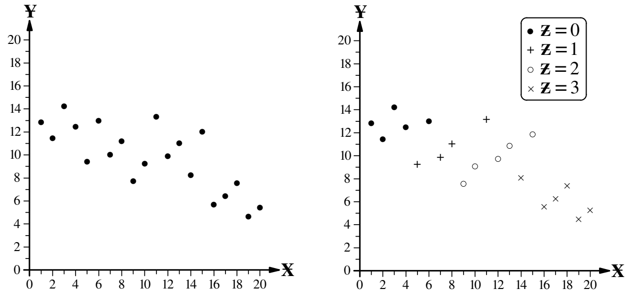
The main issue in our discussion of reality and illusion in data-based investigating of statistical relationships is the limitation imposed on Answers by comparison error arising from confounding by lurking variate(s [*e.g.*, **Z**]. A framework for our present discussion is shown in the schema at the right; it starts with an observed association (*e.g.*, of **X** and **Y** in the simplest case) [or with *lack* of (expected) association]. The five causal structures are from pages 5.34 and 5.36 in Figure 5.7.



2. Questions of the First Type

One way to investigate association is with a **scatter diagram**, – in its simplest form, Cartesian axes with dots (or other symbols), the coordinates of whose centre are the **X** and **Y** values of each bivariate observation. However, when examining such diagrams, it is easy to overlook the limitation on an Answer about the **X**-**Y** relationship imposed by different points on the scatter diagram having *differing* values of a lurking variate **Z**. This matter is illustrated by the two versions of the *same* scatter diagram at the right below:

- ∗ in the left-hand version in which **Z** values are *ignored*, we see an **X**-**Y** relationship that could reasonably be modelled by a straight line with a *negative* slope.
- ∗ in the right-hand version, where different symbols for the points denote four different values of some (non-focal) explanatory variate **Z**, the straight-line **X**-**Y** relationship can have a slope which is (close to) zero (when **Z** is 0), positive (when **Z** is 1 or 2) or negative (when **Z** is 3).



NOTES: 1. When looking at a scatter diagram of bivariate data to assess an **X**-**Y** relationship, we recognize that experience *out*side statistics with diagrams involving Cartesian axes provides poor preparation for statistics – it is difficult for later statistical training to overcome a mindset (*un*concerned with lurking variates) that arises from more formative earlier experience with such diagrams, starting in elementary school, with on-going exposure in the media, and continuing up to post-secondary level courses, including calculus and algebra.

2. Looking at multivariate data to try to detect *patterns* which answer Questions about relationships can be aided by statistical software that shows a point cloud in three dimensions on a computer screen with options like:
   - rotating the point cloud in real time;
   - using colour to distinguish subsets of the points;
   - linking points (*e.g.*, by using colour) across scatter diagrams which show point clouds for different subsets of the variates – *e.g.*, see Program 10, *Multidimensional Data Analysis* in *Against All Odds: Inside Statistics.*

3. The foregoing discussion and scatter diagrams in this Figure 10.7 draw attention to the distinction between *conditioning* on **Z** and *ignoring* **Z** when investigating relationships.
   - ∗ Conditioning is *subdividing*, as discussed at the upper left of page 11.13 in Section 2 of Figure 11.6.
   - ∗ A **marginal** (probability) distribution, referred to in Section 6 on page 11.14 of Figure 11.6 and illustrated in Tables 11.6.10 to 11.6.13, is an example of 'ignoring' the variate which is absent from the marginal distribution – for instance, in Table 11.6.11, $\mathbf{X}_2$ is absent, in Table 11.6.12, $\mathbf{X}_1$ is absent, and in Table 11.6.13, **Y** is absent.

(*continued overleaf*)

NOTES: (cont.) 3. The right-hand scatter diagram overleaf on page 10.13 shows the marginal distribution of **X** and **Y** if we think of the **Z** direction as coming vertically up from the page. With the **Z** values as given at the upper right of the right-hand version of the diagram and thinking of the page as the plane **Z** = 0, the first five points of the cloud would lie *on* the page; the remaining 14 points would then lie progressively further *above* the page in groups as one moves to the right across the diagram. This discussion reminds us that a marginal distribution is a *projection* – we *see* the marginal distribution of **X** and **Y** if we look vertically *down* on the diagram (*i.e.*, we look along the **Z** axis) to project the three-dimensional point cloud on to the plane of the page.

- It is interesting to speculate on the extent to which the ideas of conditioning and marginalizing (or projecting) provide a basis for understanding the ways in which mathematical models *approximate* reality, bearing in mind the maxim of the late Dr. George E. P. Box, a respected U.S. statistician: *All model are wrong, some are useful.*

4. Figure 11.6 on pages 11.11 to 11.18 about Simpson's Paradox and related phenomena shows how interpreting *tabulated* data and their visual representation must be informed by adequate statistical understanding, including formulating a *clear* Question and the role of lurking variates.

- Anscombe's four regression data sets in Figure 16.5 on pages 16.21 and 16.22 illustrate the statistical hazards of working with numerical data summaries without also *looking* at the data.

3. Questions of the Second Type

For the second type, we distinguish four reasons ('cases') for *association* of variates **X** and **Y** in the presence of a possible confounder (or lurking variate) **Z**; the four relevant cases from page 5.34 of Figure 5.7 are shown symbolically at the right, where an arrow denotes causation:

* **X** causes **Y** (in the presence of confounder **Z**);
* **Y** causes **X**;
* **Z** causes **X** *and* **Y** – that is, **Z** is a **common cause** of **X** and **Y**;
* coincidence [which often means that both **X** and **Y** are associated with *time* – *i.e.*, coincidence is often case (9) in which **Z** is time (whatever, if anything, 'causation' by time means)].

(1) $\mathbf{X} \to \mathbf{Y}$ (**X** causes **Y**) = (8) $\begin{tikzcd}[row sep=tiny] \mathbf{X} \arrow[dr] & \\ & \mathbf{Y} \\ \mathbf{Z} \arrow[ur] & \end{tikzcd}$ (9) $\begin{tikzcd}[row sep=tiny] & \mathbf{X} \\ \mathbf{Z} \arrow[ur] \arrow[dr] & \\ & \mathbf{Y} \end{tikzcd}$ (**Z** causes **X** *and* **Y**)

(2) $\mathbf{Y} \to \mathbf{X}$ (**Y** causes **X**) (3) **X** **Y** (coincidence)

A Question about the *actual* reason for an observed **X**-**Y** association can be answered by a process of *elimination*:

- coincidence [case (3)] requires extra-statistical knowledge to rule it out;
  - in a universe with an almost uncountable number of variates changing in value over time, coincidental associations are likely *numerous* but may largely go *un*noticed;
- identifying correctly which is the response and which the explanatory variate can rule out case (2);
- the remaining two reasons are direct causation [case (1) = case (8)] and common cause [case (9)] – an investigation with a comparative Plan that yields acceptable limitation on the Answer imposed by comparison error (due to type 2b confounding – see Figure 11.7) can then be used to try to rule out case (9); if successful, this leaves (direct) causation [case (1)] as the (likely) reason for the association.
  - The Plan that manages confounding by a common cause for **X** and **Y** [case (9)] will also manage possible confounder(s) **Z** (or $\mathbf{Z}_i$) changing as focal variate **X** changes [case (8)] in a way that makes *acceptable* the limitation imposed by comparison error (due to type 2a confounding); the common theme of managing type 2 confounding [cases (9) and (8)] is holding **Z** *fixed* as **X** changes in value.
    - \+ Case (8) is, of course, case (1) with confounder **Z** shown explicitly – **Y** is a **common response** to **X** and **Z**.

(9) $\begin{tikzcd}[row sep=tiny] & \mathbf{X} \\ \mathbf{Z} \arrow[ur] \arrow[dr] & \\ & \mathbf{Y} \end{tikzcd}$

(8) $\begin{tikzcd}[row sep=tiny] \mathbf{X} \arrow[dr] & \\ & \mathbf{Y} \\ \mathbf{Z} \arrow[ur] & \end{tikzcd}$

For Questions of the second type, comparison error results in an Answer that *mis*identifies the cause of **Y** – for instance:

- an Answer which says **X** *is* a cause of **Y** when (in reality) it is not [case (9)],
- an Answer which says **X** is *not* a cause of **Y** when (in reality) it is [case (8)].

Diagrams (4) to (8) on the facing page and on page 10.16 illustrate these matters in more detail. A real-world example occurred in *The Globe and Mail* on March 7, 2015, pages M1 and M5, in the article: **$42 AN HOUR** WHY CANADA'S YOUNG ACADEMICS ARE ON THE PICKET LINES.

The excerpt (from page M5) relevant to this discussion is given at the right. The three variates are:

X: proportion of the Canadian population with a PhD,
Y: Canada's level of productivity and innovation,
Z: Canada's level of development.

The assumption is *that more graduate students equal increased productivity and innovation.* BUT: *It could well be that countries at high levels of development produce both innovation and PhDs.*

Thus, there is a need to establish whether the applicable causal structure is our case (1) [= case (8)] or case (9).

In addition to the *causal* Question, there are also statisti-

Do we have too many PhDs?

Yet, if graduate programs have grown, it has been because governments have waved money at university administrations, operating under the assumption that more graduate students equal increased productivity and innovation.

In 2011, Ontario announced that it would fund an additional 6,000 master's and doctoral spots even as its undergraduate per-student funding, adjusted for inflation, has been declining for decades.

Turns out that the current instructions for building a PhD student may not translate into the economic gains that governments assume.

"It is an open question whether PhDs are necessary to produce innovation and productivity, or can you do that with MAs", said Daniel Munro, an analyst at the Confer-

(1) $\mathbf{X} \to \mathbf{Y}$

(8) $\begin{tikzcd}[row sep=tiny] \mathbf{X} \arrow[dr] & \\ & \mathbf{Y} \\ \mathbf{Z} \arrow[ur] & \end{tikzcd}$

(9) $\begin{tikzcd}[row sep=tiny] & \mathbf{X} \\ \mathbf{Z} \arrow[ur] \arrow[dr] & \\ & \mathbf{Y} \end{tikzcd}$

Figure 10.7. ASSOCIATION AND CAUSATION IN STATISTICS: Reality and .... (continued 1)

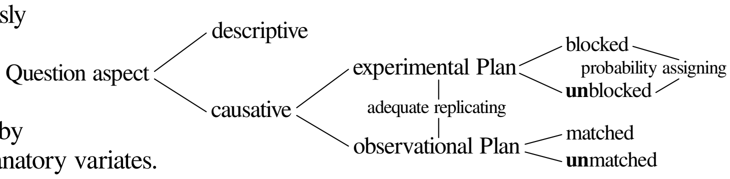
cally challenging matters of *measuring* a country's productivity and innovation (number of patents?) and level of development.

ence Board of Canada who will be releasing a report he co-authored on PhDs in the Canadian labour market this spring.

Dr. Munro says it is true that countries with a higher percentage of PhDs have higher level of productivity and innovation, and tend to produce more patents. However, the relationship between very advanced education and innovation is uncertain. It could well be that countries at high levels of development produce both innovation and PhDs.

4. Questions of the Third Type

For the third type of Question, provided that the **X**-**Y** relationship *is* causal [case (1) on the facing page 10.14], acceptable limitation imposed by comparison error (due to type 2a confounding) can usually be achieved by a comparative experimental (but *not* an observational) Plan developed and executed as previously discussed at length in Figure 5.7. The schema at the right, from the upper right of page 5.37 in Figure 5.7, reminds us of Plan components available to investigators to manage possible confounding by known and by unknown and unmeasured *non*-focal explanatory variates.



For Questions of the second and third types, which involve an inference from association to causation, the limitation imposed by comparison error (due to type 2 confounding – see Figure 11.7) on Answers can be summarized in two precepts:

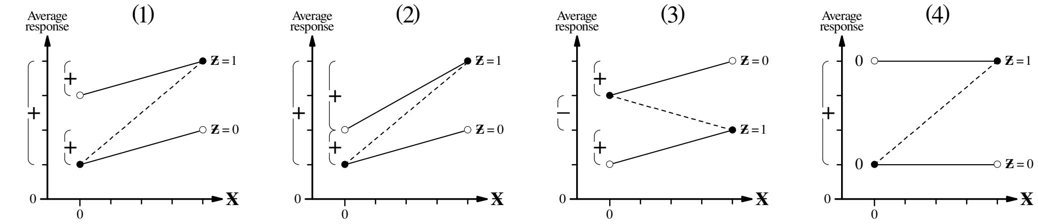
∗ **Partial or complete false positive:** when *coincidence* can be ruled out as a reason, an **X**-**Y** *association* indicates *causation* of **Y** but ***not necessarily*** (only or even in part) by **X**.
∗ **False negative:** association of **X** and **Y** may be *absent* even when they have a *causal* relationship.

These precepts formalize the ideas in Note 30 on page 5.35 in Figure 5.7 and are illustrated in more detail in eight diagrams below and overleaf on page 10.16, which show how the *average* of **Y** changes with **X** in the presence of a (binary) confounder **Z**; except for diagram (4) [case (9) on the facing page 11.14], the causal connections are those in case (8). In these eight diagrams:

- the circles (○) represent response variate averages that do *not* occur under the Plan because these averages would require the units' **Z** values to be *different* from their actual values (counterfactuals – recall the middle of page 5.50 in Figure 5.7).
- the dots (•) represent response variate averages for units whose **Z** values mean that these averages *are* observed (data),
- the solid lines are the *un*observed 'reality,' ● the dashed lines are the observed 'illusion';
- the horizontal axis has a *quantitative* scale, rather than showing **X** as an indicator variate with values of 0 and 1.

○ In the left-hand diagram (1) below, the (positive) effect of **X** on (the average of) **Y** is represented by the two braces to the *right* of the **Y**-axis; however, if there is confounding (**Z** *changing* from 0 to 1 as **X** changes), the effect of **X** on **Y** would be *observed* as the brace to the *left* of the axis. Thus, in diagram (1):
  – the dashed line involving comparison error yields a *wrong (exaggerated) magnitude* for the effect of **X** on **Y**.
  – the slope of the solid lines showing the relationship of **X** to the average of **Y** are *un*affected by the value of **Z** – that is, there is no *interaction* of **X** and **Z**.

○ Diagram (2) is like diagram (1) except there *is* interaction of **X** and **Z** – the solid lines have *different* slopes.



  – Diagrams (1) and (2) illustrate partial false positives; association of **X** and **Y** *does* correspond to causation of **Y** by **X** but confounding by **Z** distorts reality (creates illusion).

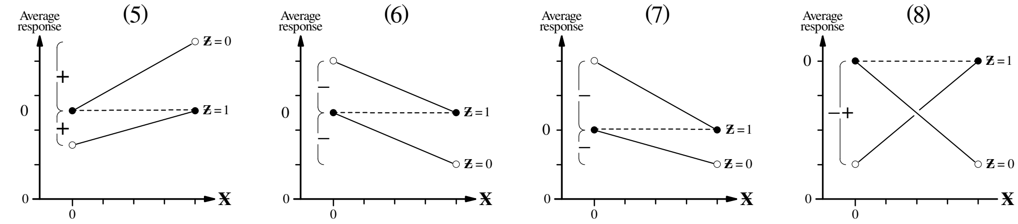
○ In diagram (3) with a *different* (negative) **Z**-**Y** relationship, comparison error is *wrong direction* for the effect of **X** on **Y**.
  – An Answer which, due to comparison error, is a wrong *direction* for an **X**-**Y** relationship is also a case of wrong *value* (and, usually, wrong *magnitude*) but, as in Simpson's Paradox (see Figure 11.6 on pages 11.11 to 11.18), the more dramatic (directional) manifestation of comparison error is usually emphasized.

○ The right-hand diagram (4) illustrates (with*out* interaction) a (complete) *false* positive Answer – there is **X**-**Y** *association* without **X**-**Y** *causation*; this is the situation when **Z** is a common cause of **X** and **Y** [case (9) on the facing page 10.14].
  – The two lines in diagram (4) have *zero* slope so we say **X** and **Y** are *independent conditional on* **Z** – for (unit) change in **X**, there is *no* change in (the average of) **Y** *provided* **Z** is (held) fixed – see also Notes 8 and 7 on page 11.18 in Figure 11.6.

⊙ A false negative Answer, as in diagram (5) overleaf, can occur in other ways; in diagram (6), such an Answer again arises from confounding with*out* interaction, and in diagrams (7) and (8) [as in diagram (5)] from *both* confounding *and* interaction.
  – In diagrams (2), (5) and (7), interaction of **X** and **Z** is *incidental* to the manifestation of comparison error when quantifying the magnitude and/or direction of the effect of **X** on **Y**; only in diagram (8) is the interaction *essential* to this manifestation.

(*continued overleaf*)

+ (Complete) false negatives involve the special case of *exact* cancellation of the effects of **X** and **Z** on **Y**.



The foregoing discussion of diagrams (1) to (8), involving type 2 confounding, is summarized in Table 10.7.1 at the right.

Table 10.7.1
X-Average response RELATIONSHIPS IN EIGHT DIAGRAMS

Diagram	Relationship	Interaction	Comparison error
(1)	Positive	None	Wrong magnitude – exaggerated
(2)	Positive	Incidental	Wrong magnitude – exaggerated
(3)	Positive	None	Wrong direction: wrong value
(4)	None	None	Wrong value: false positive
(5)	Positive	Incidental	Wrong value: false negative
(6)	Negative	None	Wrong value: false negative
(7)	Negative	Incidental	Wrong value: false negative
(8)	Mixed	Essential	Wrong value: false negative

NOTES: 5. For Questions of the third type, concerned with quantifying the effect of **X** on **Y**, the eight **X**-Average response diagrams overleaf on page 10.15 and above are only *illustrative* of comparison error because its particular manifestation depends on the interplay of several matters affecting diagram appearance, including:

- the magnitude(s) and direction(s) of the slope(s) of the **X**-**Y** relationships;
- the magnitude of the **Z**-**Y** relationship, reflected in the *vertical separation* of the solid lines;
- the absence or presence of interaction (of **X** and **Z** in their effects on **Y**);
- the *forms* of the **X**-**Y** and **Z**-**Y** relationships (*e.g.*, linear or *non*linear).

6. The *two* descriptions in the last column of Table 10.7.1 above for diagrams (4) to (8) overleaf on page 10.15 and above remind us that, for quantitative variates, comparison error is [despite the second (more dramatic) descriptions] a *wrong value* (and, usually, a wrong *magnitude*) for the effect of **X** on (the average of) **Y**. As a consequence, diagrams (4) to (8) illustrate the occurrence of comparison error for Questions of *both* the second and third type.
   - Similarly, Simpson's Paradox (discussed in Figure 11.6 on pages 11.11 to 11.18) is, under subdivision on an additional explanatory variate, a *change in value* for a 'treatment' effect, resulting in a (dramatic) change of direction of the 'effect' [analogous to diagram (3) overleaf on page 10.15.]

7. For the (unattainable) ideal [criterion (1) near the top of page 5.32 in Figure 5.7] of *all* non-focal explanatory variates remaining fixed when **X** changes to make apparent its relationship to **Y**, the Answer has *no* limitation imposed by comparison error but, when a possible confounder **Z** does *not* remain fixed, comparison error *does* impose a limitation on the Answer about the **X**-**Y** relationship. These limitations, from the foregoing discussion in this Figure 10.7, are summarized in Table 10.7.2 below; it reminds us that the nature of the limitation depends on the Question context.

Table 10.7.2

Question context	Question	Comparison error	Confounding type	Other name
Comparing proportions	Does **Y** increase or decrease with **X**?	Wrong direction for an **X**-**Y** association	3	Simpson's Paradox
Establishing causation	Is **X** the (or a) *cause* of **Y**?	Wrong cause identified for **Y**	2b	Common cause
Quantifying a treatment effect	What is the effect of **X** on **Y**?	Wrong magnitude for effect of **X** on **Y**	2a	Confounding
		Wrong direction for effect of **X** on **Y**	2a	Confounding
Improving a process	Is **X** the *most important* cause of **Y**?	Wrong main cause identified for **Y**	2a	Confounding
More than one focal variate	What is the effect of each $\mathbf{X}_i$ on **Y**?	Wrong effect of one or more $\mathbf{X}_i$ on **Y**	1, 2a	-----

- If Simpson's Paradox (in the first line of Table 10.7.2) is observed in a *sample* but the Question involves the corresponding *population*, limitation is imposed on the Answer by *sample* error as well as by *comparison* error.
  – Regardless of whether the investigation involves a (respondent) population census or sample, study, non-response and measurement error also likely need to be managed in the Plan – *e.g.*, see page 5.84 in Figure 5.7.
- The second-last line of Table 10.7.2 deals with **process improvement** – prioritizing the explanatory variates responsible for variation in **Y**; the relevant causal connections can then be thought of as case $(1)_2$ [repeated at the right from page 5.36 in Figure 5.7]. Prioritizing a set of m explanatory variates $\mathbf{X}_l$ (l = 1, 2, ..., m) can be achieved by successively prioritizing *pairs* of variates $\mathbf{X}_i$ and $\mathbf{X}_j$. The effects of confounding are the *same* as for *quantifying* the effect of **X** on **Y** and the manifestation of comparison error is identifying the **wrong *main* cause** of (variation in) **Y**.

  $(1)_2$ $\mathbf{X}_i \searrow \mathbf{Y}$, $\mathbf{X}_j \nearrow \mathbf{Y}$

  – Because prioritizing explanatory variates involves *two or more* focal variates, the Plan should make provision for estimating *interaction* effect(s).
    + ['Perfect or type 1 confounding' among some treatment effects arising from using a *fractional* factorial treatment structure may be acceptable in the Question context – recall Note 47 on page 5.45 in Figure 5.7 and see the discussion of type 1 confounding on pages 11.19 and 11.20 in Figure 11.7.]
- In the *last* line of Table 10.7.2, again involving estimating treatment effects for two or more focal variates, the Plan should provide for estimating both main *and* interaction effects.