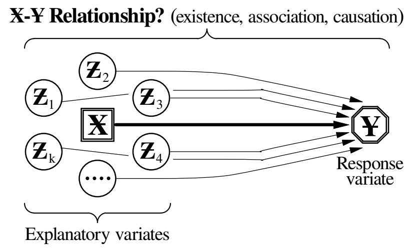
University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans

Statistical Highlight #88 provides an overview of the FDEAC Cycle. To avoid extending its already-lengthy discussion of the Design stage, the additional statistical issues that arise when investigating *relationships* are described in *this* Highlight #89, which assumes knowledge of Sections 4 to 10 of Statistical Highlight #88. The goal here is to provide an understanding of the key concepts of *experimental* Plans and *observational* Plans and their distinctions.

The material in this Highlight #89 is, with slight adaptation, that in Statistical Highlights #57 to #59, #61 to #63 and #68 (except for Section 13 on page HL89.18, a systematic presentation of ideas arising incidentally in scattered places elsewhere in these Materials). For illustrations and elaborations of *this* Highlight #89, refer to Statistical Highlights #60, #64 to #67, #69 and #70.

1. Investigating Statistical Relationships: Changing and Comparing [Statistical Highlight #57]



Relationships occur in most (perhaps all) areas of human endeavour and come in many forms. In statistics, we cast relationships in terms of *variates* – in the simplest case, between one explanatory variate (**X**, say, which *we* call the **focal** variate) and one *response* variate **Y**, over the elements of a population. However, as portrayed pictorially at the right, in statistics we can seldom ignore *other* (**non**-focal) explanatory variates (denoted **Z**$_1$, **Z**$_2$, ..., **Z**$_k$) when answering a Question about an **X**-**Y** relationship, because the Answer is predicated on **Z**$_1$, **Z**$_2$, ..., **Z**$_k$ remaining *fixed* when **X** *changes* to make apparent its relationship to **Y**. This idea arises mathematically when, to analyze data for the k+2 variates of each unit in a sample of n units, we use the response model (HL89.1) in which **Y**

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 z_{1j} + \dots + \beta_{k+1} z_{kj} + R_j, \quad j = 1, 2, \dots, n, \quad \begin{matrix} R_j \sim N(0, \sigma), \\ \text{pr. indep., EPS} \end{matrix} \qquad \text{-----(HL89.1)}$$

has a first-power (or 'straight-line') relationship to each explanatory variate; the interpretation of β_1 (the coefficient of the *focal* variate in the model) is the change in the average of **Y** for unit change in **X** while **Z**$_1$, **Z**$_2$, ..., **Z**$_k$ *all remain fixed in value.* [The interpretation of *any* of the k+2 coefficents in the structural component of (HL89.1) requires a similar caveat, of course.] A framework of terminology for discussing data-based investigating of statistical relationships is given in the schema below.

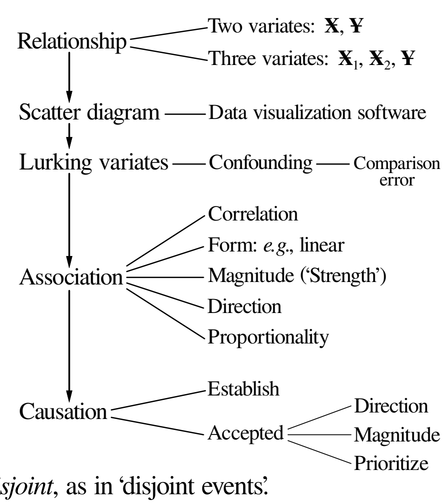
NOTE: 1. The method of investigating an **X**-**Y** relationship in statistics is by *changing* and *comparing* – we compare values of **Y** as the value of **X** changes, described above as **X** changing to make apparent its relationship to **Y**. This is why experimental and observational Plans are described as **comparative**.

- Changes in the focal variate **X** may be those that occur naturally in the population or they may be changes imposed by the investigator(s) under an experimental Plan.
- After *two* variates, the next level of complication is the relationships among *three* variates: *two* explanatory variates **X**$_1$ and **X**$_2$ and a response variate **Y** [or two responses to *one* explanatory variate, called '**common cause**'].

Experimental Plan: a comparative Plan in which the *investigator(s)* (*actively*) assign the value of the focal (explanatory) variate to each unit in the sample (or in each block);

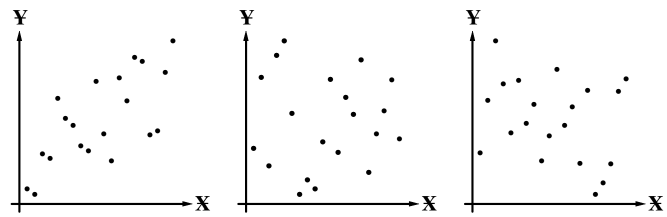
Observational Plan: a comparative Plan in which, for each unit selected for the sample, the focal (explanatory) variate (*passively*) takes on its 'natural' value ***un****influenced* by the investigator(s).

It is investigators' *in*ability to assign elements' focal variate values (*e.g.*, of age, sex, marital status and income – changes in people's dietary or exercise habits can be imposed but compliance is difficult to achieve) that restricts choice of Plan type and so weakens ability to manage comparison error; this matter is pursued in detail in Sections 9 and 10 on pages HL89.12 to HL89.15 (and in Statistical Highlights #10 and #9).



∗ A **relationship** in statistics arises from the following sequence of happenings.
- We observe that the value of a *response* variate **Y** *changes* (*i.e.*, shows *variation*) over the elements or units of a group, such as a target population, a study population, a respondent population or a sample.
  - ○ It is implicit that there are one or more *cause(s)* of these changes (*i.e.*, of this variation) in **Y**.
- We wish to account for these changes (*i.e.*, for this variation) – we introduce the idea of an *explanatory* variate **X**.
- We look for *association* between the values of **Y** and **X** (*e.g.*, using a scatter diagram – see below) – a relationship is the *connection* (if any) between *changes* in **X** and *changes* in **Y** (or in the *average* of **Y**).
  - ○ If (data establish that) **Y** remains *un*changed while **X** changes (or *vice versa*), there is *no* **X**-**Y** relationship, an idea of *un*connectedness captured by one sense of the word **independent** – see page HL89.18.
    - **+** We should recognize the distinction between the 'behavioural unconnectedness' of *independence* and the 'spatial separateness' captured by *disjoint*, as in 'disjoint events'.

∗ A **scatter diagram** is a Cartesion plot with a response variate (or estimated residual) on the vertical axis, an explanatory variate on the horizontal axis.
- A scatter diagram – a graphical attribute – is a useful way to *look* at data for an **X**-**Y** relationship. Each element appears as a dot (or other appropriate symbol) located at the co-ordinates determined by its **X** and **Y** values; three examples are shown at the right.



○ The task of looking at *multi*variate data (*i.e.*, data for three or more variates) to detect *patterns* which answer Questions about relationships can be aided by statistical software that shows, on a computer screen, a point cloud in three dimensions, with additional possibilities like:
+ using colour to distinguish subsets of the points; + rotating the point cloud in real time.

Program 10 of *Against All Odds: Inside Statistics*, entitled *Multidimensional Data Analysis*, shows such software in use. [Taking account of lurking variates when interpreting a scatter diagram is discussed in Statistical Highlight #29.]

2. Comparison Error – Lurking Variates and Confounding [Statistical Highlight #57]

As background to an **X**-**Y** relationship, **Z**$_1$, **Z**$_2$, ..., **Z**$_k$ in the schema at the upper right overleaf on page HL89.1 are called **lurking variates**, a phrase that means lurking *explanatory* variates in that each **Z** accounts, at least in part, for changes from element to element in the value of the response variate. The importance of lurking variates is that if the distributions of their values *differ* between groups of elements [like (sub)populations or samples] with different values of the focal variate, an Answer about the **X**-**Y** relationship may differ from the true state of affairs unless the differences in the values of the relevant **Z**s are taken into account.

A practical difficulty for data-based investigating of an **X**-**Y** relationship is that lurking variates are often *numerous* and so:
- important **Z**s or their differing distributions for different values of the focal variate can easily be overlooked, AND:
- substantial resources may be needed to measure values on the sampled units for those **Z**s deemed to be important.

Variates other than **X** and **Y** that *are* measured on the sampled units can be assessed by:
+ looking at a scatter diagram of y against z_i to check if **Z**$_i$ *is* an explanatory variate, AND:
+ comparing boxplots of z_i values for the different values of x to check for differences among these different focal variate values.

The *same* statistical issue raised by lurking variates is involved, with different terminology, in **confounding**; the difference is that the behaviour of lurking variates (the entity responsible) is *why* confounding (the statistical issue) occurs.

An explanatory variate responsible for confounding is called a **confounder** or **confounding variate**; these two terms are synonyms for a lurking variate whose distribution of values (over a group of units) differs for different values of the focal variate.

The following definitions summarize the foregoing discussion:

∗ **Lurking variate:** a non-focal explanatory variate whose differing distributions of values (over groups of elements or units) for different values of the focal variate, if taken into account, would meaningfully change an Answer about an **X**-**Y** relationship.

∗ **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.
- **Confounder (confounding variate):** a non-focal explanatory variate involved in confounding.

'Confounding' and 'confounder' have the convenience of being one-word terminology rather than the multi-word phrases involving 'lurking variates' which convey the same ideas.

∗ **Comparison error:** for an Answer about an **X**-**Y** relationship that is based on comparing attributes of groups of elements or units with different values of the focal variate, comparison error is the difference from the *intended* (or *true*) state of affairs arising from:
– differing distributions of lurking variate values between (or among) the groups of elements or units OR – confounding.

The alternate wording of the last phrase accommodates the equivalent terminologies of lurking variates and confounding; in a particular context, we use the version of the definition appropriate to that context:
- 'lurking variates' can more readily accommodate phenomena like Simpson's Paradox – see Statistical Highlight #51;
- 'confounding' is more common in the context of comparative Plans (*e.g.*, as in Section 8 on pages HL89.10 to HL89.12, but the variety of usage of 'confounding' can be a source of difficulty – see Statistical Highlight #3.

3. Association – Statistical Issues [Statistical Highlight #58]

The foregoing description of a relationship in statistics refers to the *association* of **Y** and **X**; this Section 3 defines association in statistics. In doing so, it provides background for discussing association between (or among) explanatory variates, and association between them and the response variate, in Section 6 on pages HL89.8 to HL89.10.

∗ **Association:** if a scatter diagram shows a clustering of its points about, say, a line with positive slope (*i.e.*, we see that, as **X** increases, **Y** also tends to increase), we say **X** and **Y** show a (positive) *association*; there is *moderate* positive association of **X** and **Y** in the left-hand diagram of the three scatter diagrams at the upper right on this page HL89.2; the right-hand

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans (continued 1)

diagram shows *weak negative* association and the middle diagram shows *little or no* association.

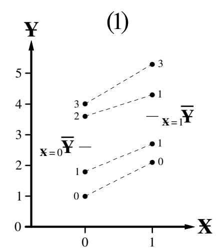
- **Correlation:** a numerical measure of *tightness of clustering* of the points on a scatter diagram about a straight line – historically, correlation is denoted r (c would have been a better choice) and its values lie in the interval [–1, 1]; the respective correlations are about +0.7, 0 and −0.25 for the three scatter diagrams at the upper right of the facing page HL89.2.
  - + If the points of a scatter diagram lie *on* a straight line with positive slope, r = +1;
  - + if the points of a scatter diagram lie *on* a straight line with negative slope, r = –1;
  - + if the points of a scatter diagram are haphazardly spread over its rectangular area, r is zero or close to it.

  Correlation is discussed and illustrated in detail in Statistical Highlight #66.

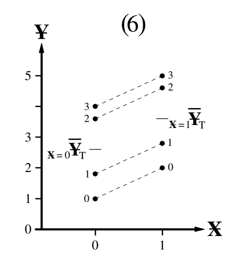
Three questions commonly of statistical interest about an association are:

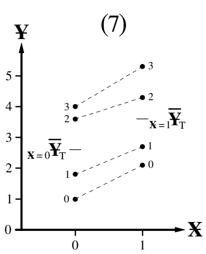
- what is its **form**? – for example, can the trend be modelled by a *straight line* (*i.e*., is the trend *linear*)?
- what is its **magnitude**? – for linear association, what is the magnitude of the *slope*?
- what is its **direction**? – for linear association, is the slope *positive* or *negative*?
  - + **Proportionality** refers to a straight-line **X**-**Y** association *through the origin.*
  - + The sign of the direction (positive or negative) of a linear association is *also* the sign of correlation, but the connection between the *magnitudes* of slope and correlation is more complicated – see Section 8 in Statistical Highlight #66.

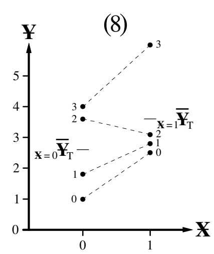
The background discussion for comparison error [*e.g*., in Section 2 near the middle of the facing page HL89.2] refers to a group of elements (or units) with a lurking variate (**Z**) whose distribution of values differs, over the elements (or units) of the group, for different values of the focal variate **X**. A consequence of this behaviour of **Z** is that the values of **X** and **Z** are *associated*, as illustrated in the scatter diagrams given below, for respondent populations with 4 or 9 elements and **Z** values (shown beside the points) like 0, 1, 2 and 3. {*Distinct* **Z** values for *all* population elements, as in diagrams (1) [at the right below] and (5) [overleaf on page HL89.4], is rare in real-world populations.}



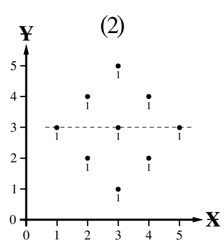
- ○ In diagram (1) at the right, the element with **Z** = 2 when **X** = 0 has **Z** = 1 when **X** = 1; thus, the change in the average of **Y** (indicated by a short horizontal line) from 2.6 to 3.6, as **X** changes from 0 to 1, no longer reflects *only* the effect of changing **X**; a limitation is therefore imposed on the Answer about the **X**-**Y** relationship by comparison error due to the behaviour of **Z** not being taken into account (or due to confounding by **Z**).
  - + Because **Z** changes with **X**, there is a (weak) **X**-**Z** association, quantified by a correlation of about −0.11 over the eight (**X**, **Z**) values; by contrast, when **Z** does *not* change with **X**, the **X**-**Z** correlation is *zero*; this is the case in diagrams (6), (7) and (8) at the right below diagram (1) – these three diagrams are used in Section 6 on pages HL89.8 and HL89.9 to illustrate the formal definition of what *we* mean by causation in statistics (the short horizontal lines again show the average of **Y** for each **X** value).

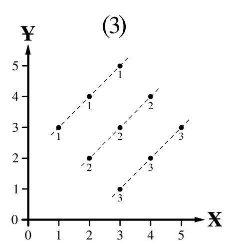


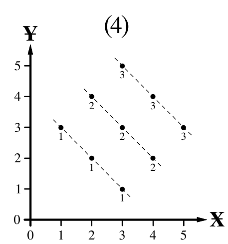




An extension of the illustration in diagram (1) is to the case of repeated values involving *more than two* **X** values.





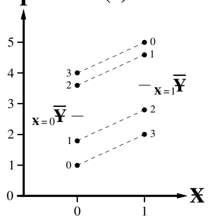


- ○ In diagram (2), if **Z** has the *same* value (say 1) for all nine units whose **X** and **Y** values yield this scatter diagram, there is *no* **X**-**Y** relationship in the sense that the **X**-**Y** correlation is zero.
  - + This *lack* of **X**-**Y** relationship is also reflected by the slope of *zero* for the straight line (shown dashed) which summarizes the trend in the points of the scatter diagram.
  - + When interpreting a scatter diagram like (2), it is easy to confuse *explicit* knowledge that there is the same **Z** value among the elements, with *assuming* this to be the case by *ignoring* the elements' **Z** value(s) – see Statistical Highlight #29.
  - + In diagrams (6) to (8) above, reminiscent of an *experimental* Plan with *two* values of the focal variate, we can accommodate *different* values of the potential confounder **Z** among the elements; by contrast, in diagram (2) above, reminiscent of an *observational* Plan, the elements must have the *same* **Z** value to meet the requirement for **Z** to remain fixed to avoid the limitation imposed on an Answer about an **X**-**Y** relationship by comparison error due to this lurking variate.
- ○ Diagram (3) above is visually the *same* as diagram (2) but the **Z** values *change* with **X** – the association of **X** and **Z** can be quantified as a correlation of about +0.7; as indicated by the dashed lines, there is now a (strong) *positive* **X**-**Y** association among points for which **Z** values are held fixed (*i.e*., for points with the *same* **Z** value).

(*continued overleaf*)

- In diagram (4) [overleaf on page HL89.3], again visually the same as diagrams (2) and (3), a *different* distribution of the *same* set of **Z** values as in diagram (3) yields a (strong) *negative* **X**-**Y** association – the **X**-**Z** correlation is again about +0.7.
  - + In diagrams (3) and (4), the **X**-**Y** relationship is the *same* for the three values of **Z**; the matter of *different* **X**-**Y** relationships for different **Z** values is pursued in Statistical Highlight #29.
  - + Like diagram (1), diagrams (3) and (4) illustrate, in a broader context, the limitation imposed on an Answer about an **X**-**Y** relationship by comparison error, when the elements' **Z** values do not remain fixed (are not the same) as **X** changes, and this behaviour is *not* taken into account (*e.g.*, when interpreting an **X**-**Y** scatter diagram).

(5)



A special case is when **Z** changes with **X** but in such a way that their values have *zero* correlation; an illustration is shown at the right in diagram (5), which is adapted from diagram (6) overleaf on page HL89.3 In such a situation, *despite* the confounding, it *is* possible (under an assumption of *additive* effects) to estimate the effect of **X** on the average of **Y**.

- This idea is exploited in Design of Experiments (DOE) when investigating a relationship with *two or more* focal variates – see Notes 24 to 27 (starting at the botom of page HL89.16).

NOTE: 2. The discussion above shows that, when looking at a scatter diagram of bivariate data to assess an **X**-**Y** relationship, experience *out*side statistics with diagrams involving Cartesian axes provides poor preparation for statistics – in calculus and algebra courses, for example, the issue of another variate affecting the interpretation of what we see in a diagram seldom (or never) arises.

- As discussed in Statistical Highlights #60, #51 and #29, lurking variates can affect the interpretation of the presence, *or* the absence, of an observed association (in even the simplest situation) involving **X** and **Y**.

4. Association Among Variates and Causation [Statistical Highlight #59]

The foregoing Section 3 deals with *association* of two *explanatory* variates, like the *focal* variate **X** and a lurking variate (or confounder) **Z**; in this Section 4, we 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 say $\mathbf{Z}_j$ is the **common cause** of **X** *and* $\mathbf{Z}_i$;
- ∗ 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) – see Note 8 on page HL89.9.

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

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

(3) **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)].

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), so-called **common response**].

(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** (**X** *and* **Z** cause **Y**)

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

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

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).

The diagrams for the remaining eight cases illustrate two possibilities:
- + **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)].
- ∗ *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.

(*continued*)

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans (continued 2)

The twelve causal structures on page HL89.4 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 6 on the lower half of page HL89.8 – 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)], although it is wise to keep case (8) in mind in discussions of case (1).
- 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*.

This leaves cases (1), (9) and (11); we discuss cases (1) and (9) in Section 7 page HL89.10. These and cases (8) and (11) are also pursued on pages HL64.1 to HL64.3 in Statistical Highlight #64.

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: 3. 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).

U → **V**
X → **Y**
Z → **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 HL89.1 at the right. The *strong* ('perfect') association in case (A) can weaken because:

Table HL89.1:

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 bottom of page HL89.6 and the top of page HL89.7) 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).

(*continued overleaf*)

NOTES: 4. Association is a straight-forward idea (we can *see* it), causation much less so; the two causal
(cont.) structures at the right [cases (8) and (9) from page HL89.4] give insight into their difference. As discussed in Note 9 on pages HL89.9 and HL89.10, under *our* definition of causation at the bottom of page HL89.6 and the top of page HL89.7:

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

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

- in the causal structure of case (8) [common *response*], 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) [common *cause*], 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.

5. Causation – An Introduction [Statistical Highlight #61]

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 **Y** (*i.e.*, an effect) being *caused* by a change in a *focal* (explanatory) variate **X**; in the illustration above, **Y** is the fire status of the house (not on fire or on fire) and **X** is the status of its electrical system (working properly or faulty) and we say (a change in) **X** *caused* (a change in) **Y**.
  - There may also be *lurking* explanatory variate(s) **Z** (or $\mathbf{Z}_1$, $\mathbf{Z}_2$, ..., $\mathbf{Z}_k$).
- Causes of **Y** *other than* **X**, if they exist, may need to be considered.
- When *coincidence* can be ruled out as a reason, an **X**-**Y** *association* indicates *causation* of **Y** but ***not necessarily*** by **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 **X** *causes* **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 element when $\mathbf{X} = 0$ and $\mathbf{X} = 1$ (sometimes phrased as: *Hold all the* $\mathbf{Z}_i$ ***fixed*** *for* .....).

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans (continued 3)

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

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

For example, the z values must be the *same* when using least squares estimates [as given in equation (HL89.2) at the right] to *compare* simple linear regression *slopes* when **X** = 0 and **X** = 1.

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} y_j(z_j - \bar{z})}{\sum_{j=1}^{n}(z_j - \bar{z})^2} \quad \text{-----(HL89.2)}$$

- These criteria are framed in terms of the (target) *population* and an appropriate *attribute*, **not** elements and their variates.
- 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 levels** (see pages HL89.16 and HL89.17).
- 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?
- The three criteria above and at bottom of the facing page HL89.6 imply that an Answer about causation from an *exprimental* Plan will usually have (substantially) less severe limitation 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 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 HL89.2 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 HL89.2:

#	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* elements 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 elements 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).

(*continued overleaf*)

– 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?

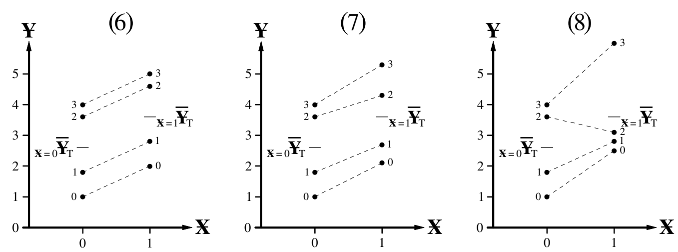
NOTE: 5. The notation used in this Highlight #89 is **X** for the *focal* variate and **Z** for other *non*-focal explanatory ('lurking') variates; elsewhere, you may see the meaning of **X** and **Z** interchanged. There can also be *more than one* focal variate.

6. Causation – Statistical Issues [Statistical Highlight #62]

Our formal definition of what it means in statistics to say (a change in) **X** *causes* (a change in) **Y** in a ***target*** population is given in Section 5 at the bottom of page HL89.6 and the top of page HL89.7.

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 pages HL89.3 and HL89.4 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: 6. The first two of the three criteria given on pages HL89.6 and HL89.7, 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 (page HL9.1) of Statistical Highlight #9.

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans (continued 4)

NOTES: (cont.) 6. ● – 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 an *attribute*, not individuals – this is consistent with the predominant concern of statistics with *populations*, not elements. A consequence of criterion (3) is that $\mathbf{X}$ need not bring about a change in $\mathbf{Y}$ for *every* element of the population for *us* to say $\mathbf{X}$ *causes* $\mathbf{Y}$.

– 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 $\mathbf{X}$ does *not* bring about a change in $\mathbf{Y}$; we would normally think of these elements as being a *small* proportion of the population.

+ 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.

7. Our three criteria at the bottom of page HL89.6 and the top of page HL89.7 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 (*e.g.*, see Section 8 on pages HL89.10 to HL89.12;

● 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 (HL89.1) on the upper half of page HL89.1, 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 (HL89.1) 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.

8. 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.

9. '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.

(*continued overleaf*)

NOTES: **(cont.)** 10. Ideas about investigating $\mathbf{X}$-$\mathbf{Y}$ relationships are summarized at the right in Table HL89.3.

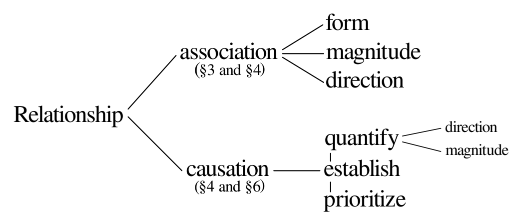
Table HL89.3: Summary of Ideas About Investigating $\mathbf{X}$-$\mathbf{Y}$ Relationships

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.

7. Investigating Statistical Relationships: Three Types of Causal Questions [Statistical Highlight #63]

Relationships investigated in statistics, which *we* describe in terms of variates, are often encountered as *associations*; investigating associations includes identifying their characteristics and/or the reasons (causal or otherwise) for them (as discussed on pages HL89.1 to HL89.4). This Section 7 is concerned with comparative Plans for investigating relationships where causation is to be established or *is* involved; the focus on the $\mathbf{X}$-$\mathbf{Y}$ relationship being *causal* means that a *change* can (potentially) be induced in $\mathbf{Y}$ by *changing* $\mathbf{X}$. These matters are summarized in the schema at the right, which reminds us that:



∗ association is usually characterized by its *form*, *magnitude* or *direction*;
  – correlation (see Statistical Highlight #66) is one measure of magnitude ('strength') for a straight-line association; form can also be *non*-linear;

∗ it is useful to distinguish three types of Questions with a causative aspect:
  – **Establishing** whether $\mathbf{X}$ *is* a cause of $\mathbf{Y}$, usually with a view to manipulating $\mathbf{X}$ to produce a (desired) change in $\mathbf{Y}$ – the quintessential example is whether cigarette smoking is a cause of lung cancer (and other life-threatening diseases), the topic of tens of thousands of data-based investigations over several decades starting in the 1940s. Establishing that an observed association of $\mathbf{X}$ and $\mathbf{Y}$ is causation of $\mathbf{Y}$ by $\mathbf{X}$ is answering the Question whether the relevant causal structure (shown again at the right from the upper right of page HL89.6) is case (1) [remembering case (8)] or case (9) [= case (12)].

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

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

  – **Quantifying** the relationship between $\mathbf{X}$ (or, more commonly, $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_q$) and $\mathbf{Y}$; this arises in the statistical area of *Design of Experiments* (DOE) – for example, the effect of temperature, humidity, light, fertilizer and insecticide levels on the growth of seedlings in a greenhouse. Quantifying a causal relationship is, in essence, investigating the case (1) causal structure – the subscripts on the case number now remind us that the Plan needs to reflect the number of focal variates involved.

    $(1)_1$ $\mathbf{X} \rightarrow \mathbf{Y}$

    $(1)_2$ $\mathbf{X}_1 \rightarrow \mathbf{Y}$, $\mathbf{X}_2 \rightarrow \mathbf{Y}$

  – **Prioritizing** causes by the size of their effect is the domain of (data-based) process improvement – trying to identify the *most important* cause (usually of excessive variation in the process output, $\mathbf{Y}$) from among many causes $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_q$.

    $(1)_q$ $\mathbf{X}_1, \ldots, \mathbf{X}_q \rightarrow \mathbf{Y}$

Questions which involve *establishing* and *quantifying* causal relationships are typically part of the *same* investigation. For example, in the Physicians' Health Study (described in Figure 10.2 of the STAT 231 Course Materials) of the effect of taking aspirin on heart disease, *two* Questions, in the context of an appropriate target population, are:

- does taking aspirin reduce heart-attack risk?
- is the reduction in heart-attack risk due to taking aspirin large enough to be practically important?

The Physicians' Health Study had to answer *both* Questions; in *in*formal discussion, it is easy to consider only *one* of the Questions and overlook the other.

Similarly, when *prioritizing* causes in process improvement investigations, investigators should:

- verify that the suspected (most important) cause *is* a cause of the (variation in the) response variate(s);
- validate that the proposed Answer *does* address the Question – that the proposed 'solution' *does* solve the 'problem'.

NOTE: 11. In STAT 231, *establishing* causation is discussed in Chapter 11 of the 2004 Course Notes but the emphasis is on *quantifying* the relationship between *one* focal variate and a response variate – for example, see Chapters 7, 10 and 15; extension to more than one focal variate is taken up in STAT 332. *Prioritizing* causes is pursued in STAT 435.

8. Terminology for Comparative Plans – The Protocol for Choosing Groups [Statistical Highlight #63]

The three criteria defining what *we* mean by causation, given at the bottom of page HL89.6 and the top of page HL89.7, involve observing a *population* under two values of the focal variate: with *all* the elements having $\mathbf{X} = 0$ and with *all* the elements having $\mathbf{X} = 1$. We try to approach this ideal in a *sampling* context by having *two* samples, one with its units having $\mathbf{X} = 0$ and the other with its units having $\mathbf{X} = 1$; each sample 'represents' the population under one of the two conditions, in the usual statistical sense of sample attributes being *estimates* of respondent population attributes. When the two samples are *compared* to quantify the change in (the average of) $\mathbf{Y}$ corresponding to a change in $\mathbf{X}$, each *non*-focal explanatory variate must have the *same* value in both samples; otherwise, there is (likely to be) comparison error. For comparative Plans for quantifying relationships, we distinguish [as introduced in Note 1 near the middle of page HL89.1]:

(*continued*)

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans (continued 5)

- ∗ an **experimental** Plan – a comparative Plan in which the *investigator(s)* (*actively*) assign the value of the focal (explanatory) variate to each unit in the sample (or in each block);
- ∗ an **observational** Plan – a comparative Plan in which, for each unit selected for the sample, the focal (explanatory) variate (*passively*) takes on its 'natural' value ***un****influenced* by the investigator(s).

This distinction reflects two types of populations encountered in data-based investigating of relationships.

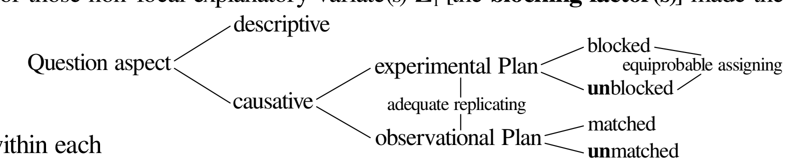
- ● A population in which all (or most) elements have *one* value of a focal variate of interest, whose value it *is* feasible to change.
  - – An example is a new drug to treat a serious disease – no one would already be taking the drug but it could be given to some participants ($\mathbf{X} = 1$) and withheld from others ($\mathbf{X} = 0$) in a clinical trial (an *experimental* Plan – see Note 17 on page HL89.13).
- ● A population in which each element has one of *two (or more)* values ($\mathbf{X} = 0, 1, ...$) of a focal variate of interest, whose value it is *not* feasible to change for any element – see Note 19 on pages HL89.14 and HL89.15.
  - – Instances of such focal variates are age, sex, marital status and income – their investigation necessarily involves an *observational* Plan; changes in people's dietary or exercise habits can be imposed but compliance is difficult to achieve.

It is investigators' *in*ability to assign elements' (or units') focal variate values that restricts choice of Plan type and so weakens ability to manage comparison error; this matter is pursued below (and also in Section 3 of Statistical Highlight #10 on pages HL10.3 to HL10.6, and in Sections 4 to 6 on pages HL9.2 to HL9.7 in Statistical Highlight #9).

For comparative Plans to answer a Question with a causative aspect, the **protocol for choosing groups** specifies whether the units of the sample will be selected so they form groups that can be used to reduce the limitation imposed on an Answer(s) by comparison error – relevant Plan components are shown in the schema at the right below.

- ∗ **Blocking** in an *experimental* Plan: forming groups of units (the **blocks**) with the *same* values of one or more non-focal explanatory variates; units within a block are then assigned *different* values of the *focal* variate. THUS:

  Blocking meets 'lurking variates' criterion (1) for those non-focal explanatory variate(s) $\mathbf{Z}_i$ [the **blocking factor(s)**] made the same within each block. SO THAT:

  Whether the Question involves establishing causation or quantifying a treatment effect, blocking *prevents confounding* of the focal variate with the $\mathbf{Z}_i$ made the same within each block, reducing the limitation imposed on Answer(s) by *comparison* error.
  - ● By making their values the *same* for one or more $\mathbf{Z}$s (*i.e.*, holding their values 'fixed') within blocks in an experimental Plan, blocking reduces variation in $\mathbf{Y}$ and so has the additional benefit of decreasing *comparing* imprecision.
    - – This additional benefit of blocking is analogous to that of *stratifying* in reducing *sampling* imprecision, as indicated in last lines of the two branches of the schema at the centre right of page HL89.15 in Note 20. [This analogy is sometimes interpreted as showing that stratifying in survey sampling is merely an instance of blocking, but this interpretation (unhelpfully) downplays the different contexts and intents of blocking and stratifying.]
- ∗ **Equiprobable assigning (EPA) [random assigning** or **randomization]:** using a probabilistic mechanism (described in the protocol for choosing groups) in an *experimental* Plan to assign the values of the focal variate with *equal* probability:

  \+ across the units of each block in a blocked Plan; \+ to each unit in the sample in an *un*blocked Plan.

  Equiprobable assigning provides a basis for theory which relates comparing imprecision to level of replicating; thus, EPA, *in conjunction with EPS and adequate replicating*, provides for quantifying comparing imprecision arising from unblocked, unknown and unmeasured non-focal explanatory variates and so allows a particular investigation to set group sizes which are likely to yield an Answer(s) with limitation imposed by comparison error that is acceptable in the Question context.
- ∗ **Matching** in an *observational* Plan: forming groups of units with the *same* values of one or more non-focal explanatory variates but *different* values of the *focal* variate. THUS:

  Matching meets 'lurking variates' criterion (1) [at the bottom of page HL89.6] for those non-focal xplanatory variate(s) $\mathbf{Z}_i$ made the same within each group. SO THAT:

  Whether the Question involves establishing causation or quantifying a treatment effect, matching *prevents confounding* of the focal variate with the $\mathbf{Z}_i$ made the same within each group, thus decreasing comparing imprecision and so reducing the limitation imposed on Answer(s) by *comparison* error.
  - – **Subdividing:** a form of *matching* used in an *observational* Plan in which the each value of the focal variate for the units of the sample is *subdivided* on the basis of the values of one or more *non*-focal explanatory variates that may be *confounded* with the focal variate under the Plan – see Table HL89.6 and its discussion on pages HL89.13 and HL89.14.

    We can think of *subdividing* as *matching* at an *aggregate* (rather than an *individual*) level; subdividing therefore has the *same* statistical benefit as matching for the non-focal explanatory variate(s) that are the basis for the subdividing.
    - ○ If subdividing is going to manage *only one* non-focal explanatory variate that is a (potential) source of comparison error, it *may* not be cost effective to devote the resources needed to obtain the relevant additional data.

(*continued overleaf*)

NOTES: (cont.) 12. Where the definitions given overleaf on page HL89.11 of blocking and matching refer to values of non-focal explanatory variates being the *same*, in practice the values may only be *similar.*

13. The groups of elements (or units) are called *blocks* in an experimental Plan but there is no such general term in an observational Plan; however, when the groups contain *two* elements (or units), they may be referred to as *matched pairs* – see Table HL89.4 at the right – but a *block* of two elements (or units) may also be referred to as a 'pair.'

Table HL89.4
Terminology for Comparative Plans

Plan	Process	Group
Experimental	Blocking	Block
Observational	Matching	(Matched pair)

- A comparative Plan involving pairing is usually our first encounter with the concepts of blocking or matching, to illustrate their role in managing comparison error.

14. In DOE, non-focal explanatory variate(s) made the same within blocks are called **blocking factor(s)**; in data-based investigating to improve industrial processes, typical blocking factors are days, shifts, batches of raw material, machine spindles or filler heads, moulding machines, moulds, or cavities within moulds.
    - The values of a blocking factor among blocks should be chosen to make its sample attribute (*e.g.*, its average or distribution) similar to its respondent (or study) population attribute.
    - An entity that is the same both within *and* among blocks (like the measuring process) is *not* a blocking factor but is part of what *defines the study population/process* – for example, data for an investigation collected on *one* day and *one* production shift. *If* such factors as day or shift have an appreciable effect on the response, the limitation imposed on the Answer by *study* error is more severe (comparison error is traded for study error).

15. Just as equiprobable *selecting*, in conjunction with *adequate replicating*, provides a theoretical basis for quantifying the likely size of sample error when estimating a (respondent) population average, so equiprobable *assigning*, in conjunction with *both* EPS *and* adequate replicating, provides the *same* benefit when estimating an average difference in (two) populations in an experimental Plan. This and other parallels between EPS and EPA are discussed in Note 20 on page HL89.15.

16. *We* use *different* terms for two processes which are similar but are used to manage different categories of error.
    - *Subdividing* (of a sample) on an *explanatory* variate to manage comparison error due to confounding by this variate, usually in an observational Plan used to answer a Question with a causative aspect.
    - *Stratifying* (of a population) on a *response* variate (or, in practice, on an explanatory variate that *stands in* for it) to make an Answer(s) more useful and/or to manage sample error – see Statistical Highlight #85.

    Elsewhere, *both* processes may be called 'stratifying.' There is further discussion of subdividing in Section 10 on pages HL89.13 to HL89.15 and of stratifying on the lower half of page HL89.14 later in Note 18.

9. Experimental Plans – Sample selecting and Blocking [Statistical Highlight #63]

The statistical ideal for sample selecting in *any* Plan is to have a *known* inclusion probability for each element of the respondent population; an example is *equi*probable selecting. For a Question with a *descriptive* aspect, if this ideal is not met, severe limitation is imposed on an Answer by *sample* error. However, experimental Plans to answer Questions with a *causative* aspect commonly do *not* use probability selecting because it is not feasible to implement it.

∗ For example, in data-based investigating to improve a manufacturing process (*e.g.*, by identifying and removing causes of excessive variation in the process output), the items manufactured by the process are often shipped away from the manufacturing plant as they are made and investigators are then forced (quickly) to use recent production, or a subset of it, as the sample – a *sample of convenience.* Three factors alleviate this *statistically* unsatisfactory state of affairs:
  – With *stable* processes [where the distribution(s) of the output response variate values remain (essentially) the same from one time period to another], a 'snapshot' of the process in time (like recent production) may often have attribute values that are *close* to those of the process in the long-term.
  – Answers are derived from *differences* in sample attributes; such Answers may have less severe limitation imposed by sample error than Answers based on sample attribute values which do *not* involve taking a difference.
    + An illustration is the Physicians' Health Study (of the effect of aspirin on heart disease), which used about 22,000 male doctors as the sample – half the doctors took aspirin and half took a placebo. It is likely that the incidence of heart attacks among doctors differs appreciably from that for the target population of all males, but the *difference* in incidence of heart attacks caused by taking aspirin may be much more similar among doctors and all males. (Two newspaper reports of this investigation are reprinted in Figure 10.2 of the STAT 231 Course Materials.)
  – Investigators may have a level of (extra-statistical) process knowledge that enables them to assess how close relevant attributes of recent production are likely to be to the corresponding long-term process attributes – informed human judgement seems to be better at sample selecting in such situations than it does when answering a Question with a *descriptive* aspect, but it is still far from the statistical ideal. [This matter is pursued in Statistical Highlight #83.]

The use of judgement selecting in the sampling protocols of comparative Plans illustrates the divergence between the statistical ideal and statistical practice under real-world constraints. Limitations imposed by the use of judgement selecting are:

(*continued*)

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans (continued 6)

- we can no longer say increased replicating reduces *sampling* imprecision; that is, we can no longer say increased sample size reduces the *likely* magnitude of sample error – see Appendix 3 on pages HL74.5 to HL74.8 in Statistical Highlight #74;
- more generally, the theoretical basis is gone for interpreting formal methods of data analysis like confidence intervals and tests of statistical significance. [Regrettably, both limitations are commonly ignored in practice.]

When answering a Question with a *causative* aspect, statistical best practice to manage comparison error (to reduce the limitation it imposes on the Answer) is to:

- block (to the extent that is feasible in the Question context) on known and measured lurking variates,
- use EPA to manage unblocked, unmeasured and unknown lurking variates,

[summarized in the precept: *Use blocking to manage what is known, probability assigning to manage what is* un*known*].

Unfortunately, there may be practical or ethical constraints on investigators' freedom to implement best practice; for instance:

- A block which is an individual participant in an investigation may not practically be able to be assigned both values of the focal variate. For example, in the Physicians' Health Study (see Figure 10.2 of the STAT 231 Course Materials) of the effect of aspirin on heart disease in males, the investigation would have gone on for too long if each participant had been required to take aspirin for several years and *not* to take aspirin for another period of the same length. For this (and other) reasons, the experimental Plan for the Physicians' Health Study was *un*blocked [recall also the last bullet (●) of Note 8 on page HL89.9].
- It would be unethical to assign human participants to the smoking group when investigating health effects of cigarette smoking;
  - in addition to ethical considerations, it is *un*likely that many non-smokers would be able to take up smoking for the investigation or that most smokers would be prepared to quit if assigned to the non-smoking group.

  Ethical issues *can* be managed but considerable resources may be needed to achieve compliance among participants when the focal variate in medical investigations with an experimental Plan involves exercise levels or dietary practices.

NOTE: 17. A special class of comparative experimental investigation is a **clinical trial**, used in medical research to assess the efficacy of new forms of treatment (*e.g.*, drugs, surgery); because the elements are *humans*, a technique called **blinding** is used (where feasible) because of its statistical benefits. [To be *blind* means not to know, for any element, whether it is in the *treatment* group or the *control* group (which usually receives a dummy treatment known as a **placebo**)]. As shown in Table HL89.5 at the right, blinding is used to manage comparison error and/or measuring inaccuracy, depending on the degree to which it is (or can be) implemented – for instance, blinding of participants is often *not* feasible when the focal variate involves exercise level or diet.

Table HL89.5

Blinding of....	Short name	Statistical benefit
Participants	Single blind	Reduced risk of *comparison error*
Treatment administrators	Double blind	Reduced risk of *comparison error*
Treatment assessors	Triple blind	Reduced *measuring inaccuracy*

10. Observational Plans – Sample selecting, Matching and Subdividing [Statistical Highlight #63]

The comments in Section 9 (on the facing page HL89.12 and above) about the use of *judgement selecting* in experimental Plans are also generally applicable to observational Plans; similarly, *matching* reduces the limitation due to comparison error on Answer(s) from an observaltional Plan but, like blocking, matching may not be feasible in a particular Question context.

Subdividing samples from the respondent subpopulations with different values of the focal variate in an observational Plan, on the basis of a possible confounder $\mathbf{Z}_i$, is illustrated in Table HL89.6 at the right for the case of *two* subpopulations. These hypothetical data for two samples (selected from subpopulations of non-smokers and smokers) of 10,000 people involve a response variate $\mathbf{Y}$ which is lung cancer status, a focal variate $\mathbf{X}$ which is smoking status, and $\mathbf{Z}_i$ is whether a unit has a family history of lung cancer, as a possible indicator of genetic predisposition to the disease; for simplicity, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}_i$ are *binary* variates in this illustration. Each of the six sets of three table entries is the sample size ('Number') and the lung cancer 'Cases' as a number and a percentage of the sample size.

Table HL89.6	**Non-smokers ($\mathbf{X}=0$)**			**Smokers ($\mathbf{X}=1$)**		
	Number	**Cases**	**%**	**Number**	**Cases**	**%**
No family history ($\mathbf{Z}_i=0$)	9,000	63	0.7	8,900	712	8
Family history ($\mathbf{Z}_i=1$)	1,000	7	0.7	1,100	88	8
Both	10,000	70	0.7	10,000	800	8

The bottom line of Table HL89.6 shows a substantially higher proportion of lung cancer cases among the smokers; because this pattern *persists* in the upper two lines of the table when the data are subdivided by $\mathbf{Z}_i$ value, the association between smoking status and lung cancer status appears *not* to be due to (our type 2b) confounding by a genetic factor which determines *both* a unit's smoking status *and* its lung cancer status, at least in so far as family history is a measure of such a factor.

Unfortunately, such subdividing of sample data to manage the limitation imposed by comparison error on an Answer about an $\mathbf{X}$-$\mathbf{Y}$ relationship from an observational Plan encounters three potential difficulties.

- Investigators have no control over the sample sizes after subdividing; if one or more of the $\mathbf{X}$-$\mathbf{Z}_i$ combinations is rare, the resulting small sample size(s) *in*crease comparing imprecision and so increase(s) the limitation imposed by comparison error on an Answer about an $\mathbf{X}$-$\mathbf{Y}$ relationship (in even the 'best case' situation of probability selecting of the samples).

- Obtaining the $\mathbf{Z}_i$ value for each unit in the samples may be difficult (and, hence, expensive) and such resource-intensive data manage only *one* possible confounder.
  - If data for two (or more) $\mathbf{Z}_i$ are collected, the ensuing subdividing into more numerous subsamples is likely to increase the limitation imposed [under probability selecting] by small sample size(s).
- Subdividing data in the manner of Table HL89.6 raises the possibility (*not* realized here) of the phenomenon known as Simpson's Paradox (and its accompanying limitation imposed on an Answer) – see Statistical Highlight #51.

NOTES: 18. In an (observational) **Case-Control** Plan (used in medical research, for example), units with a response of interest (say, lung cancer) [the 'Cases'] are matched on relevant explanatory variates (like, sex, age, region of residence) with units with*out* the response of interest (the 'Controls'). The two groups are then compared on the basis of the value of a focal variate of interest (cigarette smoking, say); appreciably higher levels of smoking among the *cases* would show *association* of smoking and lung cancer, indicating smoking may be a *cause* of the disease.

- A Case-Control Plan is used commonly:
  - when an experimental Plan would require resources beyond those available, OR:
  - as a cheaper forerunner to a possible experimental Plan to assess a promising but unconfirmed treatment effect.
- A Case-Control Plan makes the response and focal variates *appear* to be interchanged.
  - An illustration is in the 1993 newspaper article EM9359 *Fats raise risk of lung cancer in non-smokers*, which describes an investigation that compared the diets of 429 non-smoking women who had lung cancer with the diets of 1,021 non-smoking women who did *not* have lung cancer. The women all lived in Missouri, were of about the same age and represented "a typical American female population". The women filled out forms that asked about their dietary habits and they were divided into five groups based on the amount of fat and other nutrients they said they consumed. The investigation found that those with diets with the lowest amount of saturated fat and the highest amount of fruits, vegetables, beans and peas were the least likely to develop lung cancer. At the other end of the scale, 20 per cent of the women with the highest consumption of fat and diets lowest in fruits, vegetables, beans and peas had about six times more lung cancer.

    The *actual* response variate (lung cancer) and focal variate (level of dietary fat) *appear* to be interchanged solely as an artifact of the Case-Control Plan.
- Probability selecting is commonly *not* used for the cases and/or the controls, which has consequences for the limitation imposed by comparison error on Answer(s).
  - Cases are often a **sample of convenience** – units with a response of interest conveniently *available* to the investigator(s), like people with a particular disease in a hospital or clinic nearby to the investigator(s).
    - + Consequences of non-probability selecting to answer Question(s) with a descriptive or a causative aspect are discussed in Statistical Highlight #83 – recall also the discussion on the lower half of page HL89.12.
  - Controls are often selected *non*-probabilistically to meet the matching criteria; this *in*creases the limitation imposed by comparison error due to the selecting method, to be set against the *de*creased limitation imposed by comparison error due to the confounding which is managed by the matching.
    - + A way of selecting controls probabilistically is to form *strata* (or groups) of controls where the units in one stratum match one case; controls for the investigation are then selected probabilistically from these strata.
      - ⊙ While decreasing the limitation imposed by *comparison* error, such stratifying *in*creases the limitation imposed by *study* error, because the matching criteria which define the strata *restrict* the elements (or units) which can make up the study (and respondent) population of controls.

        A Plan should carefuly consider whether error from one source should be managed in a way that *in*creases error from another source – that is, whether there *is* a net gain in reducing the limitation on an Answer by managing one category of error in a way that *in*creases limitation due to *another* category – recall the discussion of comparison error and study error in Note 14 on the upper half of page HL89.12.
    - + When controls are selected *non*-probabilistically, there is no theoretical basis for an inverse relationship between sampling imprecision and (the square root of) of the sample size – see Statistical Highlight #21 – so there is no *statistical* reason why a larger sample size for controls will decrease comparing imprecision.
    - + The blocks in a blocked experimental Plan are also often selected *non*-probabilistically but, as discussed in Statistical Highlight #83, judgement selecting *may* still allow an experimental Plan to have *acceptable* limitation imposed by comparison error on an Answer to a Question with a causative aspect.

19. As well as being key components of *selecting* and *assigning,* probabilistic processes can also be useful in *measuring* when questions with sensitive answers are involved. The simplest version of so-called **randomized response** involves an iterviewer asking a 'Yes/No' question about past 'sensitive' behaviour of the person being interviewed (the 'interviewee'), usually with the goal of estimating the population proportion of people who will admit they have engaged in the behaviour (*e.g*., abortion, illicit drug use, viewing child pornography, money laundering, terrorism); randomized response was developed to manage two difficulties such investigations encounter:
    - the interviewee may find the question too sensitive to give a truthful answer, AND/OR:
    - the interviewer may be under a legal obligation to report the behaviour of a respondent who answers 'Yes'.

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans (continued 7)

NOTES: (cont.) 19. Randomized response manages both matters by having a box containing a number of (say, 100) cards, each with one of two questions in known proportions (say, 20% of cards have the first question, 80% have the second):

– *Is your birthday in July?* – *Have you ever engaged in ... ?*,

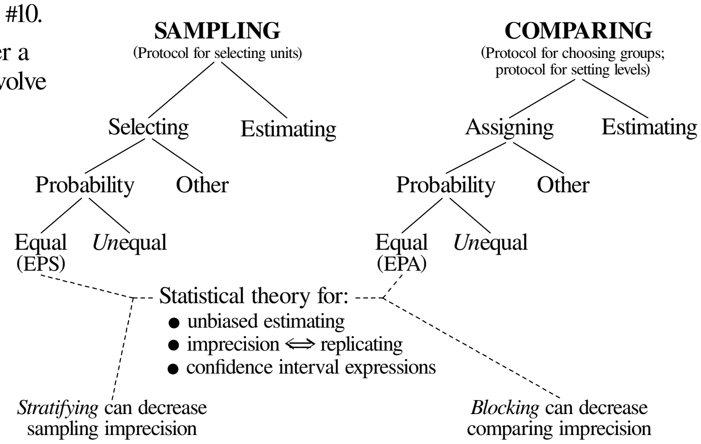
where the first question has a *known* distribution of *answers*, the second question names the behaviour of interest. The interviewee selects a card 'at random with replacement' from the (well-mixed) box and answers it.

○ If the person *has* engaged in the behaviour *and* has selected a card containing the question about it, (s)he is not necessarily divulging sensitive information by answering 'Yes' and so may be more likely to answer truthfully [provided the interviewer *has* convinced the interviewee of their protection under randomized response];

○ the interviewer is legally protected because (s)he does not know to which question a 'Yes' answer applies.

An introductory version of the probabilistic basis of estimating the population proportion under randomized response, from a sample selected from the population, is the topic of Question **A4-12** of the STAT 220 assignments.

20. Equiprobable *selecting* and equiprobable *assigning* are key components of the processes of sampling and (experimental) comparing, whose *similarities* are illustrated in the two tree diagrams in the schema at the right below – it is taken from page HL10.6 in Statistical Highlight #10.



● Investigations involving *comparing* (to answer a Question with a *causative* aspect) usually involve *sampling*; investigations involving *sampling* to answer a Question with a *descriptive* aspect need *not* involve *comparing*.

● **Probability selecting** means having *known* unit inclusion probabilities in the selecting process; introductory statistics courses emphasize *equi*probable selecting (EPS) as the basis of statistical theory for the behaviour of *sample* error under repetition.

– Here, *we* coin the term **probability assigning** (EPA) for having known *assigning* probabilities; *we* encounter mainly the special case of *equi*probable assigning, and we *hope* for a fortunate outcome of (roughly) equal numbers of units in the groups (*e.g.*, control and treatment) being compared.

+ Analogous to EPS, EPA is the basis of statistical theory for the behaviour of *comparison* error under repetition – see the discussion on pages HL10.3 and HL10.4 of Table HL10.4 in Statistical Highlight #10.

+ Surprisingly, 'probability assigning' is not currently used elsewhere, perhaps reflecting separate development of the two large statistical areas of survey sampling and design of experiments.

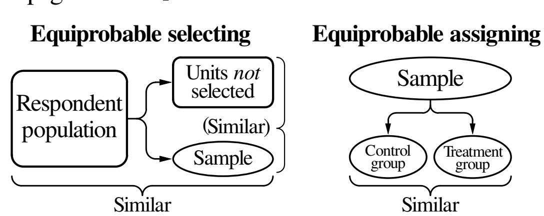
Our *equiprobable selecting* is usually *simple random selecting* or *random selecting* elsewhere; our *equiprobable assigning* is **random assigning** or **randomization** elsewhere.

– Statistical theory is *used* in the estimating branches of the two tree diagrams in the schema at the right above; these branches are part of the Analysis stage of the FDEAC cycle.

+ Selecting/assigning probabilities as the basis of the theory used for estimating is noteworthy.

– The schema at the right above reminds us of the analogous roles of stratifying and blocking in sampling and comparing [but recall the comment (–) near the middle of page HL89.11].

● As shown pictorially at the right, a common theme of EPS and EPA is dividing a group of elements into *sub*groups that are likely to be *similar* enough *under adequate replicating* for the respective limitations imposed on Answer(s) by sample error and comparison error to be acceptable in the investigation context.



– When *selecting* the sample, the group of elements is the respondent population, the subgroups are the units *not* selected and the sample.

21. Investigators may have the opportunity to trade study error and sample error; a Plan involving *smaller* study error and (likely) *larger* sample error is usually preferred because:

– study error requires *extra*-statistical knowledge to assess it and its behaviour can seldom be quantified; BUT:

+ under EPS, sampling theory describes the behaviour of sample (and, perhaps, measurement) error under repetition of selecting and estimating [see the second bullet (●) above Note 1 on page HL21.3 in Statistical Highlight #21].

(*continued overleaf*)

11. The Protocol for Setting Levels and Interaction [Statistical Highlight #68]

The protocol for setting levels specifies the *values* to be taken by relevant explanatory variate(s) in a comparative Plan; the simplest case is *two* values of *one* focal variate but there is terminology to deal with the complications of more than two values of more than one focal variate. This terminology is used mainly in the context of *experimental* Plans.

∗ A **factor** is an explanatory variate; we distinguish an explanatory variate that is:
- a *focal* variate; – a *non*-focal variate used as a **blocking factor**;
- a *non*-focal variate whose value is managed for other reasons – see Note 27 on the lower half of the facing page HL89.17.

Our concern in this Statistical Highlight #89 is with factor(s) that are *focal* variate(s).

∗ Factor **levels** are the set of value(s) assigned to a factor – that is, (usually) the set of values assigned to the (or a) focal variate. Choosing the *values* for levels in the context of a particular investigation may require extra-statistical knowledge.

∗ A **treatment** is a *combination* of the levels of the factor(s) applied to a unit [in the sample (or the blocks)].

∗ A **run** is part of the Execution stage of an experimental Plan in which all the data are collected for *one* treatment.

∗ A **factorial** treatment structure involves *all* combinations of the levels of the (two or more) factors.

∗ The **(treatment) effect** of **X** on **Y** (usually) refers to the change in the *average* of **Y** for *unit* change in **X** and:
- implies the **X**-**Y** relationship is (believed to be) *causal* – a change in **X** *causes* (brings about) a change in **Y**;
- includes both the *magnitude* and *direction* of the relationship – for example, the *slope* and its *sign* for a *linear* relationship;
- requires that all non-focal explanatory variates $\mathbf{Z}_i$ hold their (same) values when **X** changes;
- is defined (the 'true' effect) over the elements of the *respondent population.*

∗ **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.
- Interaction is a key concept also in the discussions of Statistical Highlights #60, #65, #69 and #83.

Illustrations of this terminology are:

○ Levels of sex as a factor are *female* and *male*;
the ranges used as levels of (human) age need careful consideration – ranges that are too *narrow* may consume unnecessary resources in attaining adequate replicating, while ranges that are too *broad* may obscure the effect(s) of age.

○ In a taste test of different brands of beer, the factor would be *brand of beer* and its levels would be the individual *brands.*

○ When there is only *one* focal variate, the treatments are its levels;
when there are *two* focal variates, $\mathbf{X}_1$ (say) with *two* levels (denoted 1 and 2) and $\mathbf{X}_2$ with *three* levels (denoted A, B, C), there are $2 \times 3 = 6$ treatments (1A, 1B, 1C, 2A, 2B, 2C) in a factorial treatment structure;
with four factors each at three levels, that are $4 \times 4 \times 4 = 4^3 = 64$ possible treatments.

NOTES: 22. Not all experimental Plans lead to an Execution stage in runs.
- Process improvement investigations often *do* – the Execution stage is then a set of runs, one for each treatment.
- A clinical trial of a drug usually does *not* involve runs – each participant takes the drug (or a placebo) [*i.e.*, the (two) treatments are applied to elements or units] for the *whole* period of the Execution stage.

When the Execution stage *does* involve runs, equiprobable assigning consists of equiprobable *ordering* of the *runs*, because unblocked, unknown and unmeasured non-focal explanatory variates are considered as being *time-dependent.*

23. Equiprobable assigning of treatments to units may not be feasible in an experimental Plan when one factor has hard-to-alter levels. For example, if pouring temperature (at two levels, say, of 1,450°F and 1,600°F) is a factor in an investigation to improve a process for making iron castings, the temperature of the furnace containing the molten iron cannot easily be altered; it may therefore be necessary to do *consecutively* all the runs at each temperature, instead of having the pouring temperature low or high under equiprobable assigning for each run. This *lack* of probability assigning *in*creases the limitation imposed on an Answer by comparison error.
- What is desirable statistically in data-based investigating may also be compromised in process improvement investigations by having to carry out the Execution stage under time pressure while the process continues normal operation; in addition to possible lack of equiprobable assigning, there may be limitations on Answers because:
  - there is not enough time to obtain adequate *replicating*;
  - the data reflect process operation only over a *limited* time period.

For a process with an *un*acceptably-high long-term scrap rate undergoing an investigation to try to reduce the rate, there have been instances of negative reaction from management to an investigation with an experimental Plan where some treatment(s) involve factor levels that would (temporarily) *in*crease the scrap rate.

24. In ordinary English, interaction customarily involves *two* entities; in statistics, *three* (or more) variates are involved – two (or more) focal variates and one response variate.
- *Confounding* also involves two explanatory variates and one response variate; it is compared and contrasted with interaction (and with other causal structures involving three variates) in Statistical Highlight #65.

Interaction is not limited to *two* factors – k focal variates have $\binom{k}{i}$ possible i-factor interactions; for example, four

(*continued*)

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – Experimental and Observational Plans (continued 8)

NOTES: (cont.) 24. focal variates have $\binom{4}{2} = 6$ two-factor interactions, $\binom{4}{3} = 4$ three-factor interactions, and $\binom{4}{4} = 1$ four-factor interaction. When i = 1, the k '1-factor interactions' are the k **main effects**, the effects of the k factors *individually*.

- Main effects and interaction effects are instances of **treatment effects**, and are represented by (response) *model parameters*. Any *linear combination* of such parameters where the coefficients sum to *zero* is called a **contrast**.
- For four focal variates, there are 4 + 6 + 4 + 1 = 15 treatment effects potentially of interest; these effects can *all* be estimated with a 16-run experimental Plan involving a factorial treatment structure.
- A *two*-factor interaction is the effect of one factor on the effect of another factor on a response variate; a *three*-factor interaction is the effect of one factor on the effect of another factor on the effect of a third factor on a response variate, and so on.

25. When there are two or more focal variates, 'lurking variates' criterion (1) [see the bottom of page HL89.6] entails all *non*-focal variates be kept the same but, to allow interaction effect(s) to be estimated, the *focal* variates must be changed *together* according to the balanced scheme of a factorial treatment structure. However, confounding *may* then arise as outlined in Note 26 below.

- A ***mis***understanding of criterion (1) is to extend the *ensuring everything stays the same* precept to the *focal* variates and to only change them ***one*** *at a time*. For example, for *two* factors each with *two* levels (denoted *Lo* and *Hi*), have one run with both $\mathbf{X}_1$ and $\mathbf{X}_2$ set 'Lo', another run with $\mathbf{X}_2$ set 'Hi' and another with $\mathbf{X}_2$ back at 'Lo' and $\mathbf{X}_1$ set 'Hi'; the resulting data, shown as three response variate averages in Table HL89.7 at the right, do *not* allow the $\mathbf{X}_1$-$\mathbf{X}_2$ interaction effect to be estimated, because there is no run with both factors set 'Hi'.

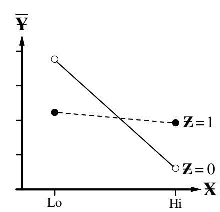
Table HL89.7	$\mathbf{X}_2$ Lo	$\mathbf{X}_2$ Hi
$\mathbf{X}_1$ Lo	$\overline{\mathbf{Y}}_{\text{Lo,Lo}}$	$\overline{\mathbf{Y}}_{\text{Lo,Hi}}$
$\mathbf{X}_1$ Hi	$\overline{\mathbf{Y}}_{\text{Hi,Lo}}$	No data

Such a Plan, if it required four replicates for each treatment, would involve 12 runs. With a *factorial* treatment structure, only *4* runs provide the *same* level of replicating *and* an estimate of the interaction effect.

26. The idea in Note 24 above of estimating 15 treatment effects from a 16-run experimental Plan can be adapted to *fewer* estimates (7, say) from *fewer* (say 8 of the 16) runs – this is called a **fractional factorial** treatment structure (here, a **half** fraction). Under such a Plan, it is only possible to estimate *combinations* of treatment effects, like the main effect of one factor *and* one three-factor interaction. Because we cannot separate such combinations into their individual effects without data for *all 16* runs, there is *confounding* within the combinations.

- Inability to separate *treatment* effects under a Plan involving a *fractional* factorial treatment structure would be better called *perfect* confounding, to distinguish it from *partial* confounding (see Section 2 in Statistical Highlight #3), where the association of $\mathbf{X}$ and $\mathbf{Z}$ typically has a correlation with magnitude *less* than 1. As discussed in Statistical Highlight #3, both cases are usually (unwisely) simply called 'confounding' without distinction.

27. 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 $\mathbf{X}$-$\mathbf{Y}$ relationship displays an interaction effect, because the *slope* of the (linear) relationship between $\mathbf{X}$ and the *average* of $\mathbf{Y}$ is *different* (here, smaller negative magnitude) when (non-focal) explanatory variate $\mathbf{Z} = 1$ ('Hi') than when $\mathbf{Z} = 0$ ('Lo'). In the tile-manufacturing process, if:



- $\mathbf{Y}$ is tile size *after* firing in an oven,
- $\mathbf{X}$ 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,
- $\mathbf{Z}$ 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 $\mathbf{Z} = 1$), 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 $\mathbf{Z} = 1$ and clay must not be too expensive an ingredient.

12. Postscript

To retain perspective on the many (statistical) issues in this Highlight #89, it is useful to keep in mind the key ideas, as introduced in the two schemas at the upper and lower right on page HL89.1.

- ∗ For investigating *relationships*, statistics presents methods intended to address the (many) sources of limitations on Answers, particularly when establishing *causation*; an important distinction is between *association* and causation.
- ∗ Appreciating the difference between *experimental* and *observational* Plans is key to understanding the likely severity of limitations on Answers, especially those involving causation.
  - ○ These matters take time to cover and are typically the focus of introductory courses.

 (*continued overleaf*)

∗ In later courses (like Design of Experiments), causation is often not in dispute and the emphasis is on *variation* [of response variate(s)] and on identifying the *explanatory* variates responsible for (*i.e.*, causing) it. The goal is then to manage the values of appropriate explanatory variate(s) in a way that *reduces* variation and so improves process output, usually in manufacturing or service contexts. [Recall Note 11 on page HL89.10 and the discussion preceding it.]

13. Appendix: *Probabilistic* Independence in Models, Independence in Relationships

Discussion of statistical investigations of *relationships*, as in this Highlight #89 and the related Statistical Highlights #57 to #70, raises issues arising from differences in the meaning of words used as technical terms and in ordinary converation.

As shown in the schema at the right, for (discrete) *probability* we start with the idea of a **process** that has two or more possible **outcomes** (but only one of which occurs when the process is executed); the set of *all* possible outcomes is the **sample space**, which is made up of **points** (although they could equally well be called outcomes as already defined). An **event** is any subset of these points, technically including the 'subset' consisting of *all* the points, so the sample space can be considered an event. [Common notation for the sample space is S – *we* use a Roman S to avoid confusion with the population (data) standard deviation $\mathbf{S}$ and the (model) random variable S representing the sample standard deviation s under equiprobable selecting.]

Process
↓
Outcome
↓
Sample space
↓
Points (Outcomes)
↓
Event

- Attentive readers will have noticed that the five **bolded** terms above have their own technical meanings – for instance, an event is different from describing the occurrence of a hurricane or a concert as an 'event' – and a **model** has already been mentioned.

A staple of introductory probability is *probabilistic independence*, often described as the probability of an intersection being the *product* of the probabilities of its component events – for example: $\Pr(A \cap B) = \Pr(A) \times \Pr(B)$. When learning this material, or even when it has become familiar, it is easy to forget that probabilistic independence:

- ○ is a *model* for a specific mathematical idealization;
- ○ is a statement about the relationship among the values of particular probabilities, *not* primarily about the events involved,
  - – if probability values change, events may change from probabilistic *in*dependence to probabilistic *de*pendence or *vice versa*;
- ○ has *nothing* to do with *causation* of events.

The concept of probabilistic independence carries over from events to models involving random variables: the joint probability function of two or more probabilistically independent random variables is the *product* of their individual probability functions.

- Probabilistic independence is extended to *conditional* probabilistic independence to describe models where the former occurs only under certain restrictions ('conditions'); the adjective 'conditional' is that of *conditional* probability.
  - ○ 'Conditional probabilistic independence' is commonly shortened to 'conditional independence', perhaps because it is thought that 'conditional' makes it clear that probability modelling is involved.

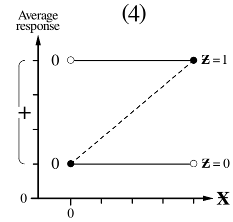
When we progress from discrete to *continuous* random variables and probability distributions, it is seldom possible to present the mathematically rigorous foundation provided by measure theory; instead, we learn mainly rules for manipulating these entities. For instance, we take it on trust that the probability a continuous random variable takes any particular value is zero.

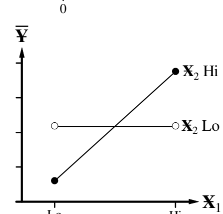
With this background of probabilistic independence in probability modelling, *we* now use **independence** to convey the *statistical* idea of behavioural unconnectedness, and its 'opposite', **dependence** to cover the seemingly limitless variety of connectedness we encounter in the real world. These two words are the language that arises naturally in *statistical* discussion of terminology like association and causation and their associated ideas in the schema at the lower right of page HL89.1.

- The price we pay for distinguishing (unqualified) 'independence' from *probabilistic* independence is that we must forego, in all probability modelling contexts, the (tempting) convenience of shortening the nine syllables of 'probabilistic independence' to just the four syllables of 'independence'.

With the independence-probabilistic independence distinction on hand, we can describe correctly statistical issues involved in the diagrams at the right from page HL60.3 in Statistical Highlight #60 and from Note 7 on page HL69.4 in Statistical Highlight #69. The feature of these diagrams relevant to the present discussion is their *horizontal* lines: for $\mathbf{Z} = 0$ and $\mathbf{Z} = 1$ in the upper diagram, for $\mathbf{X}_2$ Lo in the lower diagram. Referring to the upper diagram, *we* say:

(4)





+ the average response and $\mathbf{X}$ are *independent* when $\mathbf{Z} = 0$ *and* when $\mathbf{Z} = 1$, BUT **NOT**:

– the average response and $\mathbf{X}$ are *probabilistically* independent when $\mathbf{Z} = 0$ and when $\mathbf{Z} = 1$;

the latter statement (wrongly) implies that we have undertaken the difficult task of formulating a probability model for the context of the upper diagram. Similarly, for the lower diagram, *we* say:

+ $\overline{\mathbf{Y}}$ and $\mathbf{X}_1$ are *independent* when $\mathbf{X}_2$ is Lo but *not* when $\mathbf{X}_2$ is Hi, BUT **NOT**:

– $\overline{\mathbf{Y}}$ and $\mathbf{X}_1$ are *conditionally* independent when $\mathbf{X}_2$ is Lo but not when $\mathbf{X}_2$ is Hi.

Although our description of relationships in both diagrams involves independence, their $\mathbf{X}$-$\mathbf{Y}$ *causal* relationships (respectively absent and inherent) differ markedly – recall the bottom of page HL89.4.

More broadly, in simple linear regression, *we* describe a *horizontal* fitted line as showing that the average of $\mathbf{Y}$ [and hence, (likely) $\mathbf{Y}$] and $\mathbf{X}$ are independent – in the relevant theory, *only* $\mathbf{Y}$ (not $\mathbf{X}$) is modelled by a random variable.