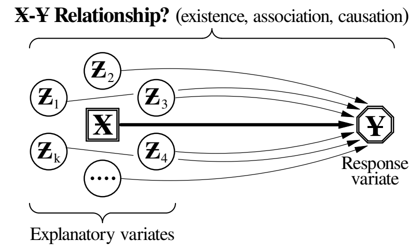
University of Waterloo W. H. Cherry

RELATIONSHIPS IN STATISTICS: An Introduction

1. Investigating Statistical Relationships: Changing and Comparing



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 (HL57.1) in which **Y**

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 z_{1j} + ... + \beta_{k+1} z_{kj} + R_j, \quad j = 1, 2, ..., n, \quad R_j \sim N(0, \sigma), \text{ pr. indep., EPS} \qquad \text{-----(HL57.1)}$$

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 (HL57.1) requires a similar caveat, of course.] A framework of terminology for discussing data-based investigating of statistical relationships is given in the schema below.

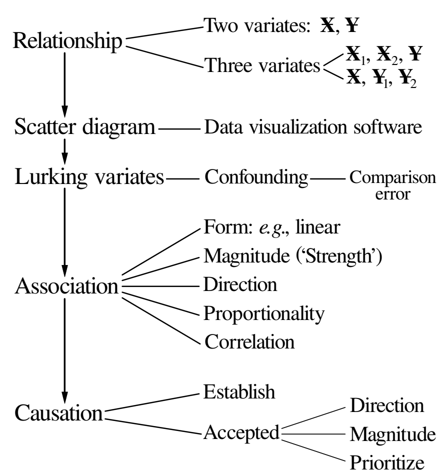
NOTE: 1. The method of investigating an **X**-**Y** relationship in statistics is by *changing* and *comparing* – we compare values of **Y** as the value of **X** changes, described above as **X** changing to make apparent its relationship to **Y**. This is why experimental and observational Plans are 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 (explanaroty) variate to each unit in the sample (or in each block);

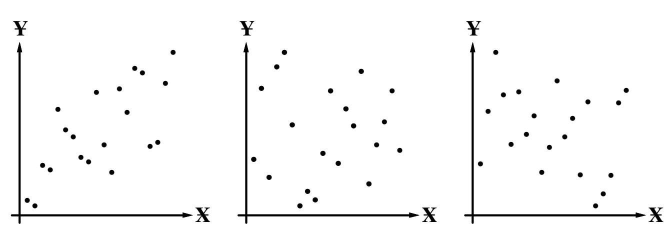
Observational Plan: a comparative Plan in which, for each unit selected for the sample, the focal (explanatory) variate (*passively*) takes on its 'natural' value ***un****influenced* by the investigator(s).

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 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.



* 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**.
    - \+ 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 coordinates determined by its **X** and **Y** values; three examples are shown at the right.



(*continued overleaf*)

○ 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

As background to an **X**-**Y** relationship, $\mathbf{Z}_1$, $\mathbf{Z}_2$, ..., $\mathbf{Z}_k$ in the schema at the upper right overleaf on page HL57.1 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 $\mathbf{Z}_i$ *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 (**Z**) whose differing distributions of values over groups of elements or units with different values of the focal variate, if taken into account, would meaningfully change an Answer about an **X**-**Y** relationship.

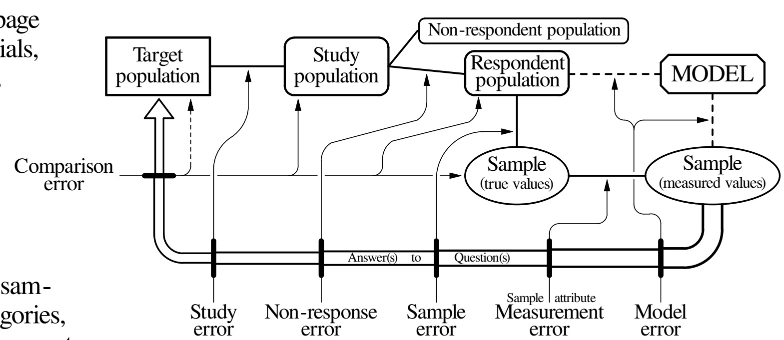
∗ **Confounding:** differing distributions of values of one or more *non*-focal explanatory variate(s) among two (or more) groups of elements or units [like (sub)populations or samples] with different values of the focal variate.
   – **Confounder (confounding variate):** a non-focal explanatory variate 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(s), 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 3 on pages HL63.1 to HL63.4 in Statistical Highlight #63), but the variety of usage of 'confounding' can be a source of difficulty – see Statistical Highlight #3.

The schema introduced at the centre right of page 5.20 in Figure 5.7 of the STAT 231 Course Materials, and progressively adapted on pages 5.25 and 5.27, has been further adapted as shown at the right to include *comparison* error; the schema now has all *six* error categories defined on the upper half of page HL6.4 in Statistical Highlight #6.



○ In the schema, the four arrows arising from comparison error point to *boxes* representing *groups* of elements or units (a population or a sample) rather than, as for the other five error categories, to *lines joining boxes*; the comparison error arrow at the right is to be taken as pointing to *both* sample ellipses.
   – *Multiple* comparison error arrows are a consequence of its different manifestations in different Question contexts, as summarized in Table HL60.2 on the lower half of page HL60.4 in Statistical Highlight #60.

Plan components to manage comparison error are summarized at the end of Table HL6.1 on page HL6.5 in Statistical Highlight #6.