Statistical Highlight #29

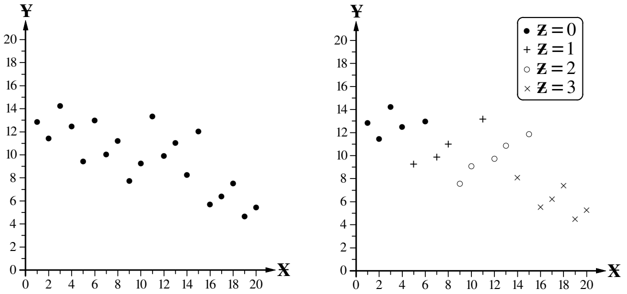
University of Waterloo

W. H. Cherry

GRAPHICAL ATTRIBUTES: Scatter Diagrams

To answer a Question about an X-Y relationship between *quantitative* variates, it is always useful to *look* at relevant data shown as a **scatter diagram** – Cartesian axes with dots (or other symbols), the coordinates of whose centre are the X and Y values of each bivariate observation. However, when examining such diagrams, it is easy to overlook the limitation on an Answer about the X-Y relationship imposed by different points on the scatter diagram having *differing* values of a lurking variate Z. This matter is illustrated by the two versions of the *same* scatter diagram at the right below:

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



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

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

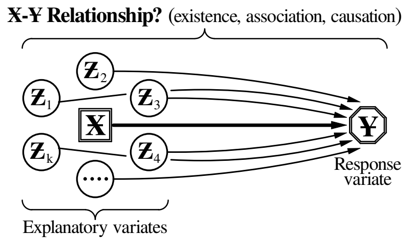
3. The foregoing discussion and scatter diagrams in this Statistical Highlight #29 draw attention to the distinction between *conditioning* on Z and *ignoring* Z when investigating relationships.
   - ∗ Conditioning is *subdividing*, as discussed at the upper left of page HL51.3 in Section 2 of Statistical Highlight #51.
   - ∗ A **marginal** (probability) distribution, referred to in Section 6 on page HL51.4 of Statistical Highlight #51 and illustrated in Tables HL51.10 to HL51.13, is an example of ‘ignoring’ the variate which is absent from the marginal distribution – for instance, in Table HL51.11, X_2 is absent, in Table HL51.12, X_1 is absent, and in Table HL51.13, Y is absent.

   The scatter diagram at the right above shows the marginal distribution of X and Y if we think of the Z direction as coming vertically up from the page. With the Z values as given at the upper right of the right-hand version of the diagram and thinking of the page as the plane Z = 0, the first five points of the cloud would lie *on* the page; the remaining 14 points would then lie progressively further *above* the page in groups as one moves to the right across the diagram. This discussion reminds us that a marginal distribution is a *projection* – we *see* the marginal distribution of X and Y if we look vertically *down* on the diagram (*i.e.*, we look along the Z axis) to project the three-dimensional point cloud on to the (*two*-dimensional) plane of the page.
   - ● It is interesting to speculate on the extent to which the ideas of conditioning and marginalizing (or projecting) provide a basis for understanding the ways in which mathematical models *approximate* reality (see the maxim quoted near the middle of page HL6.3 in Statistical Highlight #6 – *All models are wrong, some are useful*).

4. Other statistical issues involving scatter diagrams are discussed in Statistical Highlight #53.

Appendix – Lurking Variates and Confounding [optional reading]

As background to managing the limitations arising from lurking variates when answering Question(s) about an X-Y relationship between a focal variate X and a response variate Y, Z_1, Z_2, ..., Z_k in the schema at the right 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. As illustrated above, the importance of lurking variates is that, if the distributions of their values *differ* between groups of elements [like populations or samples] with different values of the focal variate,



(*continued overleaf*)

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.

Three relevant definitions are:

✻ **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 uints [like (sub)populations or samples] with different values of the focal variate.

✻ **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 of the definition of comparison error 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 (see Statistical Highlight #63), but the variety of usage of 'confounding' can be a source of difficulty (see Statistical Highlight #3).

2006-06-20