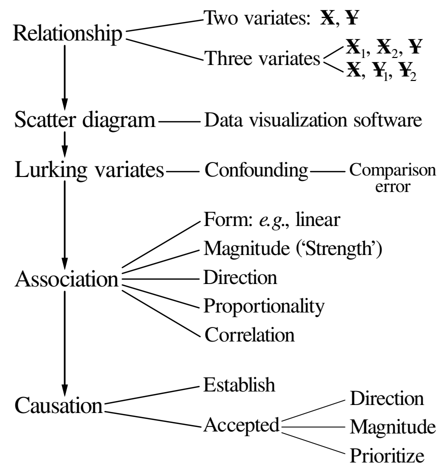
University of Waterloo W. H. Cherry

RELATIONSHIPS IN STATISTICS: Association – Statistical Issues



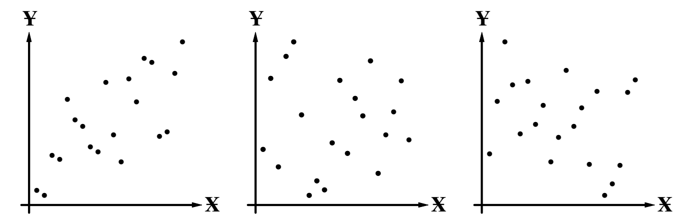
A **relationship** in statistics arises from the following sequence of happenings.

- 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** (but see Note 1 overleaf at the bottom of page HL58.2).
    - + We should recognize the distinction between the 'behavioural unconnectedness' of *independence* and the 'spatial separateness' captured by *disjoint*, as in 'disjoint events'.

A framework of terminology for discussing data-based investigating of statistical relationships is given in the schema above.

∗ 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 (or unit) appears as a dot (or other appropriate symbol) located at the coordinates determinted by its **X** and **Y** values; three examples are shown at the right.



The foregoing description of a relationship in statistics refers to the *association* of **Y** and **X**; this Statistical Highlight #58 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 Statistical Highlight #62.

∗ **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 right above. The right-hand diagram shows *weak negative* association and the middle diagram shows *little or no* association.

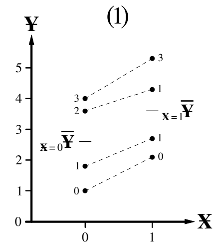
Questions of statistical interest about an association are:

- 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.
- **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 right above.
  - + 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.

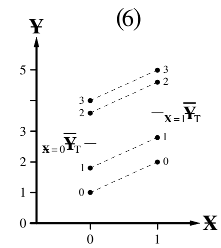
  Correlation is discussed and illustrated in detail in Statistical Highlight #66.

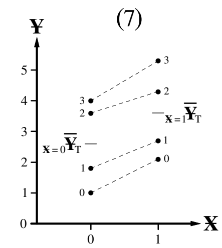
The background discussion for comparison error [*e.g.*, at the beginning of Section 2 on the overleaf side (page HL57.2) in Statistical Highlight #57] refers to a group of elements (or units) with a lurking variate (**Z**) whose distribution of values differs, over the elements (or units) of the group, 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 overleaf on page HL58.2, 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) and (5), is rare in real populations.]

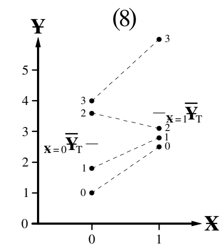
(*continued overleaf*)



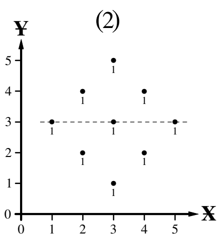
- In diagram (1) at the right, the element with **Z** = 2 when **X** = 0 has **Z** = 1 when **X** = 1; thus, the change in the average of **Y** (indicated by a short horizontal line) from 2.6 to 3.6, as **X** changes from 0 to 1, no longer reflects *only* the effect of changing **X**; a limitation is 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 Statistical Highlight #62 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).

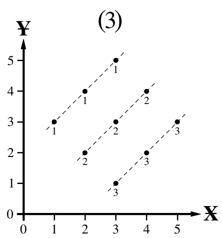


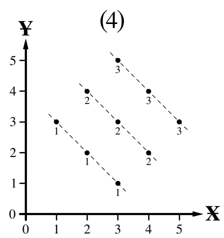




An extension of the illustration in diagram (1) is to the case of repeated values involving *more than two* **X** values.

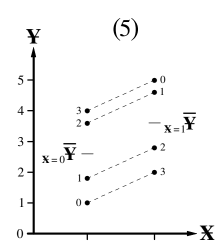






- In diagram (2), if **Z** has the *same* value (say 1) for all nine units whose **X** and **Y** values yield this scatter diagram, there is *no* **X**-**Y** relationship in the sense that the **X**-**Y** correlation is zero.
  + This *lack* of **X**-**Y** relationship is also reflected by the slope 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.

  [Experimental and observational Plans are discussed in Sections 4 and 5 on pages HL63.3 to HL63.6 in Statistical Highlight #63 (and also in Statistical Highlight #9).]
- Diagram (3) 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).
- In diagram (4), 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 (and also in Statistical Highlight #60).
  + 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).



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) above. 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 3 to 6 (starting at the botom of page HL68.1) in Statistical Highlight #68.

NOTE: 1. 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 – 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.

- 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**.