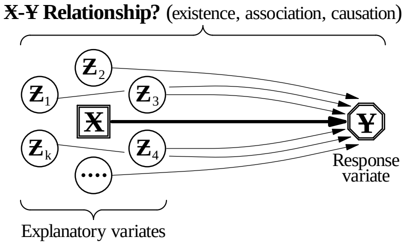
Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS

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 (Ӿ, say, which we call the **focal** variate) and one **response** variate ¥, 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 Ƶ$_1$, Ƶ$_2$, ..., Ƶ$_k$) when answering a Question about an Ӿ-¥ relationship, because the Answer is predicated on Ƶ$_1$, Ƶ$_2$, ..., Ƶ$_k$ remaining *fixed* when Ӿ *changes* to make apparent its relationship to ¥. 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 (9.2.1) in which

$$Y_j = \beta_0 + \beta_1 x_j + \beta_2 z_{1j} + ... + \beta_{k+1} z_{kj} + R_j, \quad j = 1, 2, ..., n, \quad R_j \sim N(0, \sigma), \text{ indep., EPS} \qquad \text{-----(9.2.1)}$$

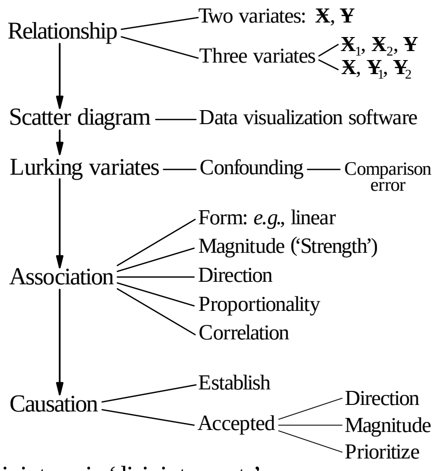
¥ has a first-power (or 'straight-line') relationship to each explanatory variate; the interpretation of β_1 (the coefficient of the *focal* variate in the model) is the change in the average of ¥ for unit change in Ӿ while Ƶ$_1$, Ƶ$_2$, ..., Ƶ$_k$ *all remain fixed in value.* [The interpretation of *any* of the k+2 coefficents in the structural component of (9.2.1) requires a similar caveat, of course.]

Terminology for describing data-based investigating of statistical relationships is given in the schema at the right below.

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

- Changes in the focal variate Ӿ may be those that occur naturally in the population or they may be changes imposed by the investigator(s) under an experimental Plan (see also Note 36 near the middle of page 9.28).
- After *two* variates, the next level of complication is relationships among *three* variates: *two* explanatory variates Ӿ$_1$ and Ӿ$_2$ and a response variate ¥ ('common response') [or two responses to *one* explanatory variate ('common cause')].

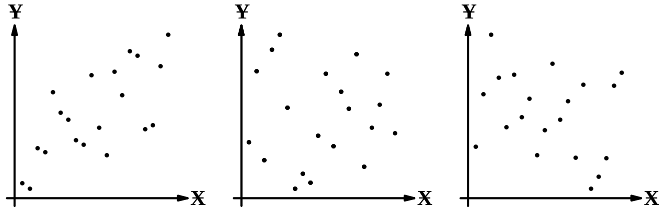
2. The notation in this Figure 9.2 is Ӿ for the *focal* variate and Ƶ for other *non*-focal explanatory variates, not *vice versa*.



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

- We observe that the value of a *response* variate ¥ *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 *causes* of (or 'reasons' for) these changes (*i.e.*, of this variation) in ¥.
- We wish to account for these changes (*i.e.*, for this variation) – we introduce the idea of an *explanatory* variate Ӿ (the **focal** variate).
- We look for **association** between the values of ¥ and Ӿ (*e.g.*, using a scatter diagram – see below) – a relationship is the *connection* (if any) between *changes* in Ӿ and *changes* in ¥ (or in the *average* of ¥).
  - ○ If (suitable data show that) ¥ remains *un*changed while Ӿ changes (or *vice versa*), there is *no* Ӿ-¥ 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 Ӿ-¥ relationship. Each element (or unit) appears as a dot (or other appropriate symbol) located at the coordinates determinted by its Ӿ and ¥ 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. Interpreting scatter diagrams and two classic examples of them are discussed in Appendices 1 and 2 on pages 9.28 to 9.33.

(*continued overleaf*)

2. Comparison Error – Lurking Variates and Confounding

As background to an **X**-**Y** relationship, $\mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_i, ..., \mathbf{Z}_k$ in the schema at the upper right overleaf on page 9.5 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:

- ● it is easy to overlook important **Z**s or their differing distributions for different values of the focal variate, 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 try to check if $\mathbf{Z}_i$ *is* an explanatory variate, AND:
- + comparing boxplots of z_i values for the different values of x to try to identify differences in $\mathbf{Z}_i$ for different **X** 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 groups of elements (or units)] differs for different values of the focal variate.

The following definitions summarize the foregoing discussion:

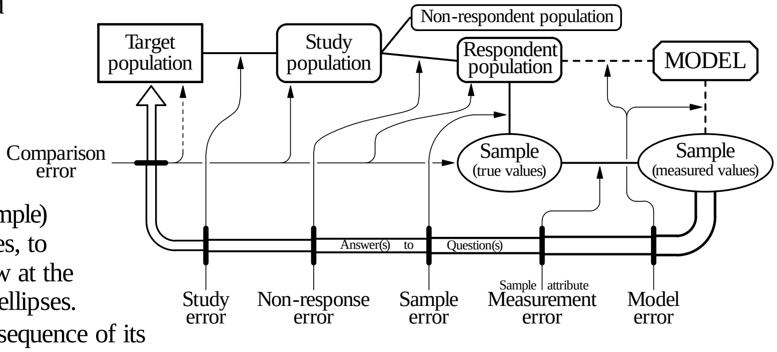
- ∗ **Lurking variate:** a non-focal explanatory variate whose differing distributions of values over groups of elements (or units) with different values of the focal variate, if taken into account, would meaningfully change an Answer about an **X**-**Y** relationship.
- ∗ **Confounding:** differing 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 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 – 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 Figure 9.8 on pages 9.57 to 9.64;
  - ● 'confounding' is more common in the context of comparative Plans, as in Section 15 which starts on page 9.26, but the variety of usage of 'confounding' can be a source of difficulty – see Figure 9.9 on pages 9.61 to 9.64).

Sections 3 to 12 (pages 9.6 to 9.21) which follow provide necessary background before we continue discussion of comparison error.



The schema which summarizes the data-based investigative process, using terminology of the FDEAC cycle, is given at the right; it shows all six error categories, although comparison error is the one of primary interest in the present context.

- ○ In the schema, the four arrows arising from comparison error point to *boxes* representing *groups* of elements or units (a population or 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 9.12.2 on page 9.76 in Figure 9.12.

Plan components to manage comparison error are summarized in Table 9.2.4 near the middle of page 9.14.

3. Association – Statistical Issues

The description of a relationship in statistics overleaf on page 9.5 refers to the *association* of **Y** and **X**; this Section 3 defines association in statistics and we then take up the issue of association between (or among) explanatory variates, and of association between them and the response variate, in Section 5 on pages 9.10 and 9.11.

- ∗ **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 scatter diagram at the lower right overleaf on page 9.5. The right-hand diagram shows *weak*

(*continued*)

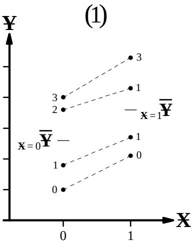
Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 1)

negative association and the middle diagram shows *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 it *linear*)?
- what is its **magnitude**? – for linear association, what is the magnitude of the *slope* (or the *correlation* – see below)?
- what is its **direction**? – for linear association, is the slope (or correlation) *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 on pages 9.34 and 9.35 in Figure 9.3.
- **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 lower right of page 9.5.
  - + 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.

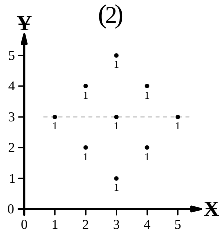
  Correlation is discussed in detail in Figure 9.3 of the Course Materials.

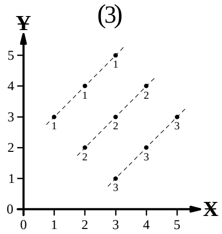
The discussion at the beginning of Section 2 at the top of page 9.6 refers to a group of elements with a lurking variate (Z) whose distribution of values differs, over the elements of the group, for different values of the focal variate X. A consequence of this behaviour of Z is that the values of X and Z are *associated*, as illustrated in the following scatter diagrams, 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 9.8, is rare in real populations.]

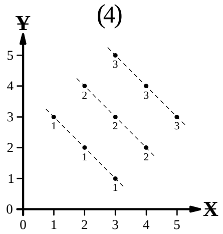


- ○ In diagram (1) at the right, the element with Z = 2 when X = 0 has Z = 1 when X = 1; thus, the change in the average of Y (indicated by a short horizontal line) from 2.6 to 3.6, as X changes from 0 to 1, no longer reflects *only* the effect of changing X; a limitation is therefore imposed on the 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 [as in diagrams (6), (7) and (8) overleaf on page 9.8], the X-Z correlation is *zero*.

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





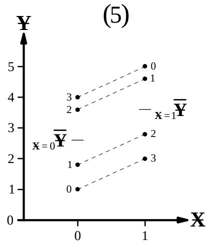


- ○ In diagram (2), if Z has the *same* value (say 1) for all nine elements whose X and Y values yield this scatter diagram, there is *no* X-Y relationship in the sense that the X-Y correlation is zero.
  - + This *lack* of X-Y relationship is also reflected by the slope of *zero* for the straight line (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 also Note 39 in Appendix 1 on pages 9.28 and 9.29.
  - + In diagrams (6) to (8) overleaf on page 9.8, 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 9 and 10 on pages 9.14 to 9.17.]
- ○ 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 Appendix 1 on pages 9.28 and 9.29.
  - + 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).

(*continued overleaf*)

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) below. 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 Section 12 and Notes 25 to 28 on pages 9.20 and 9.21.



NOTE: 3. The foregoing discussion 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 the diagram seldom (or never) arises.

4. Causation – Statistical Issues

To define ***formally*** in statistics what it means to say (a change in) **X** *causes* (a change in) **Y** in a ***target*** population, we state three criteria (useful in practice when establishing causation or quantifying the effect of **X** on **Y**):

(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* .....).

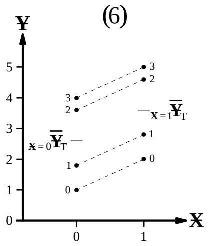
(2) **FOCAL VARIATE:** Observe the population **Y**-values, and calculate an appropriate attribute value, under *two* conditions:
⊙ with *every* element having $\mathbf{X} = 0$;
⊙ with *every* element having $\mathbf{X} = 1$.

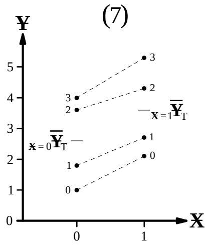
(3) **ATTRIBUTE:** $\text{Attribute}(\mathbf{Y}, \text{perhaps some of } \mathbf{Z}_1, \mathbf{Z}_2, \ldots\ldots, \mathbf{Z}_k | \mathbf{X} = 0) \neq \text{Attribute}(\mathbf{Y}, \text{perhaps some of } \mathbf{Z}_1, \mathbf{Z}_2, \ldots\ldots, \mathbf{Z}_k | \mathbf{X} = 1)$;
those of $\mathbf{Z}_1$,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

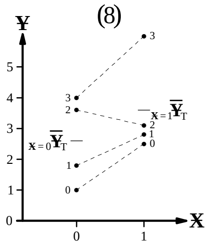
The notation $\mathbf{X} = 0$ and $\mathbf{X} = 1$ for values of the focal variate is *symbolic* – 0 and 1 *represent* two *actual* values of **X** in a particular context; actual values of the focal variate are set in the **protocol for setting levels**, discussed in Section 12 on pages 9.19 to 9.21.

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 the four diagrams overleaf on page 9.7 and diagram (5) above, 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: 4. The first two of the three criteria given above, 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 example, *alive* or *dead* instead of *lung cancer* or *no lung cancer* in our example.
- 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

(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 2)

NOTES: (cont.) 4.
- *sub*populations; this matter is pursued in Section 14 and Note 34 on pages 9.25 and 9.26.
  - 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 **X** need not bring about a change in **Y** for *every* element of the population for us to say **X** *causes* **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 **X** does *not* bring about a change in **Y**; we would normally think of these elements as being a *small* proportion of the population.
    - + For instance, there *may* be individuals for whom smoking would *never* cause 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 based on observed lung cancer *rates* among non-smokers and smokers.

There is further discussion of *statistical* issues involving causation in Figure 9.11 of these Course Materials.

5. The three criteria (on the facing page 9.8) 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 **Z**s) remain fixed; three approaches try to meet this criterion to manage comparison error in practice:
- hold *some* **Z**s fixed *physically* by blocking, matching or subdividing (see Section 7 on pages 9.12 to 9.14);
- under probability assigning of elements' focal variate values, use statistical theory to manage *under repetition* differences among unblocked, unmeasured and unknown **Z**s (see Section 13 on pages 9.21 to 9.25);
- use a *response model* in the Analysis stage of the FDEAC cycle to hold some **Z**s fixed *mathematically*, but even a quite elaborate model, like equation (9.2.1) on page 9.5 in Section 1, cannot involve *all* possible **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 (9.2.1) tries to manage mathematically the effects on **Y** of **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. It is inplicit in the three criteria that observed behaviour is *reproducible* among different investigations.

6. 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 Figure 9.8 on pages 9.57 to 9.64.
- 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 (or units) 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 35 on pages 9.27 and 9.28), 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.

7. '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 **X** *causes* **Y**, there is *association* of elements' **X** and **Y** values over the target population under the two values of **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 **X** over the elements of the target population under the two values of **X**; the values of these **Y**s with the common cause **X** will *also* be associated.

This causation-association connection under our definition of causation in statistics is used in Section 5 overleaf.

(*continued overleaf*)

NOTES: (cont.) 8. An example of the caveat in ‘attribute’ criterion (3) is: when using least squares estimates [equation (9.2.2) at the right] to *compare* simple linear regression *slopes*, the z values must be the *same* when **X** = 0 and **X** = 1.

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} y_j(z_j - \bar{z})}{\sum_{j=1}^{n} (z_j - \bar{z})^2} \qquad \text{-----(9.2.2)}$$

9. Ideas about investigating **X**-**Y** relationships are summarized at the right in Table 9.2.1.

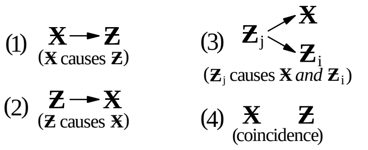
Table 9.2.1: Summary of Ideas About Investigating X-Y Relationships

Criterion (1): the ideal	Ensure all the **Z**$_i$ hold their (same) values for every population element when **X** = 0 and **X** = 1
Criterion (3)	For causation, a relevant *attribute* must differ in value when **X** = 0 and **X** = 1
Confounding	Confounding arises when one or more of the **Z**$_i$ change in value when **X** = 0 and **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.

5. Association Among Variates and Causation

Section 3 on pages 9.6 and 9.8 deals with *association* of two *explanatory* variates, like the *focal* variate **X** and a lurking variate (or confounder) **Z**; we now distinguish four reasons (‘cases’) for such associations, which are also shown symbolically at the right, where an arrow denotes causation.



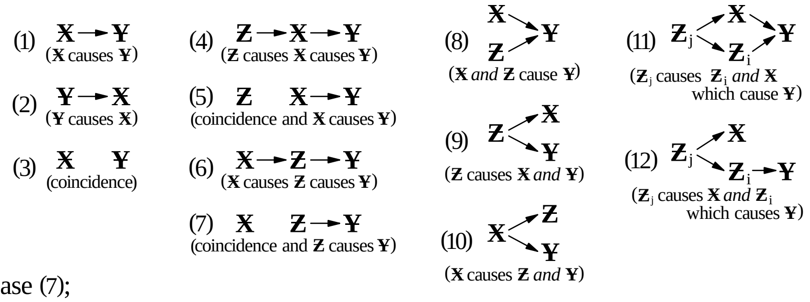
∗ **X** causes **Z**;
∗ **Z** causes **X**;
∗ **Z**$_j$ causes **X** *and* **Z**$_i$ – we say **Z**$_j$ is the **common cause** of **X** *and* **Z**$_i$;
∗ coincidence [which often means both **X** and **Z** are associated with *time* – *i.e.*, coincidence is often case (3) where **Z**$_j$ is time (whatever ‘causation’ by time means – recall Note 6 overleaf on page 9.9)].

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 response [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 **Z**$_i$) and **Y** are associated in the last *nine*.
∗ **Z** (or **Z**$_i$) and **X** are associated in the last *nine* [except perhaps in case (8)].



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:

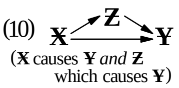
+ **X** and **Y** are *associated* **and X** *causes* **Y**: cases (1), (4), (6), (8), (10) and (11);
+ **X** and **Y** are *associated* **but X** does *not* cause **Y**: cases (9) and (12).

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)].

The twelve causal structures above 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 – recall the first bullet (•) in Note 4 on page 9.8 – 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)].
- 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 21 on page 9.19 and the discussion on page 9.22 in Section 13 to the left of Table 9.2.10].



(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 3)

- Case (12) is both: – case (9) with an intermediary variate shown in the Ƶ$_j$-¥ branch,
  – case (11) for the Question *Is X̶ a cause of ¥?* when the Answer is *No.*

This leaves cases (1), (9) and (11); we discuss cases (1) and (9) in Section 6 overleaf on page 9.12 and we pursue them and cases (8) and (11) in Section 11 on pages 9.18 and 9.19 – see also Figure 9.12 on pages 9.65 to 9.72.

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 X̶ and ¥ are *associated*, we can**not** say, without further investigating, that a change in X̶ will *bring about* (or *cause*) a change in ¥.

- The following Figure 9.3 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 X̶ and ¥ 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 X̶-¥ relationship, phrases used in statistics like *association is not (necessarily) causation* and *correlation is not (necessarily) causation* encompass *three* possibilities:
  – the X̶-¥ relationship is a *coincidence* – this may pique our curiosity but is seldom of practical importance;
  – X̶ and ¥ are *associated* but X̶ does not *cause* ¥;
  – X̶ *is* a (or possibly *the*) cause of ¥.

  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 ¥ and X̶, the variates *observed* to be associated.
  + *Lack* of association of X̶ and ¥ does *not* rule out causation of ¥ by X̶ – as X̶ changes, a confounder Ƶ may change in such a way that ¥ remains *un*changed – see diagrams (5) to (8) on the upper half of page 9.75 in Figure 9.12.

NOTES: 10. When (a change in) an explanatory variate Ʉ (a focal variate X̶ or a confounder Ƶ) *causes* (a change in) a variate V̶ (a response variate ¥ or a focal variate X̶), several matters determine the *strength* of the association (as quantified by the correlation, say, of Ʉ and V̶, if they are *quantitative* variates).

Ʉ → V̶
X̶ → ¥
Ƶ → X̶

- If Ʉ is the *only* cause of V̶ and acts on a time scale that is **short** relative to the period of observation, there is a *high* correlation of Ʉ and V̶; in the absence of measurement error, the magnitude of r would be 1.
  – An illustration is force X̶ causing acceleration ¥.
- *Weaker* association of Ʉ and V̶ can occur for several reasons, as illustrated by the data for the occurrence of lung cancer ¥ in relation to smoking status X̶ in three non-smokers and three smokers in Table 9.2.2 at the right. The *strong* ('perfect') association in case (A) can weaken because:
  – 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 X̶ does *not* cause ¥;

  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 start of Section 4 on page 9.8) in terms of an *attribute*.

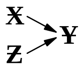
Table 9.2.2:

Element	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

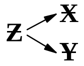
In these ways, we account for differing strengths of *association* observed in *causal* X̶-¥ relationships or, expressed another way, we account for why (a change in) X̶ *causes* (a change in) ¥ but, for *some* population elements:
- ¥ changes when X̶ does *not* change (*e.g.*, some *non*-smokers get lung cancer), OR:
- ¥ does *not* change when X̶ changes [*e.g.*, some smokers do *not* get lung cancer (before they die from another cause)].

11. Association is a straight-forward idea (we can *see* it), causation much less so; the two causal structures at the right [cases (8) and (9) from page 9.10] give insight into their difference. As discussed in Note 7 on page 9.9, under our definition of causation on the upper half of page 9.8:
- in the causal structure of case (8) [common *response*], there is *association* of X̶ and ¥ and of Ƶ and ¥ but *no* necessary association of the (unconnected) causes X̶ and Ƶ; BUT:
- in the causal structure of case (9) [common *cause*], there is *association* of Ƶ and ¥ and of Ƶ and X̶ so there is *necessarily* association of ¥ and X̶.

(8)



(9)



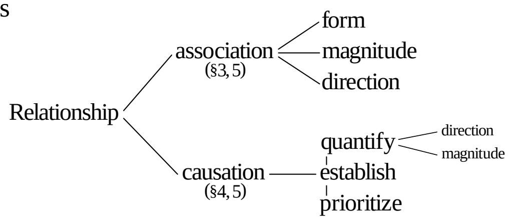
The *difference* between the two structures lies in the *direction* of the arrows denoting causation – if their direction is *reversed* in either diagram, they are the *same* causal structure, apart from the variate names. *Our* definition of causation thus suggests that causation is *directed association*, 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 on pages 9.63 and 9.64 in Figure 9.9.

(*continued overleaf*)

6. Investigating Statistical Relationships – Three Types of Causal Questions

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 (see also Figure 9.12 on pages 5.73 to 5.76). This Section 6 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 Figure 9.3) 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 page 9.10) is case (1) or case (9) [= case (12)].

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

    (9) $$\begin{tikzcd} & \mathbf{X} \\ \mathbf{Z} \arrow[ur] \arrow[dr] & \\ & \mathbf{Y} \end{tikzcd}$$
  - – **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.

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

    $(1)_2$ $$\begin{tikzcd} \mathbf{X}_1 \arrow[dr] & \\ & \mathbf{Y} \\ \mathbf{X}_2 \arrow[ur] & \end{tikzcd}$$
  - – **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 $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_q$.

    $(1)_q$ $$\begin{tikzcd} \mathbf{X}_1 \arrow[dr] & \\ \vdots \arrow[r] & \mathbf{Y} \\ \mathbf{X}_q \arrow[ur] & \end{tikzcd}$$

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 9.18 of the Course Materials) of the effect of aspirin on heart disease, *two* Questions, in the context of an appropriate target population, are:

- ● does aspirin reduce heart-attack risk?
- ● is the reduction in heart-attack risk due to 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: 12. In STAT 220, *establishing* causation was discussed in Part 9, starting in Figure 9.9, although the emphasis was on *quantifying* the relationship between *one* focal variate and a response variate; extension to more than one focal variate was taken up in STAT 322. *Prioritizing* causes is pursued in STAT 435.

7. Terminology for Comparative Plans – The Protocal for Choosing Groups

The three criteria defining what *we* mean by causation, in Section 4 on the upper half of page 9.8, 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:

- ∗ an **experimental** Plan – a comparative Plan in which the *investigator(s)* (*actively*) assign the value of the focal 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 18 on page 9.15).
- ● 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 – recall Note 6 on page 9.9.

(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 4)

– 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.

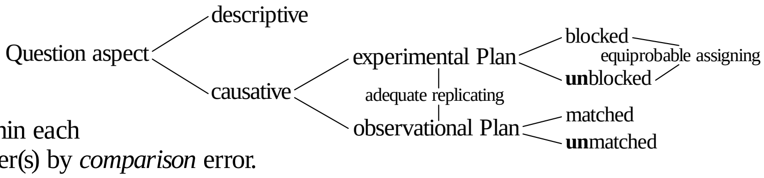
It is investigators' *in*ability to assign units' focal variate values that restricts choice of Plan type and so weakens ability to manage comparison error; this matter is pursued in Sections 13 to 15 on pages 9.21 to 9.28.

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 below at the right:

∗ **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) Ƶ$_i$ [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 Ƶ$_i$ made the same within each block, reducing the limitation imposed on Answer(s) by *comparison* error.



- By holding one or more Ƶs fixed within blocks in an experimental Plan, blocking reduces variation in ¥ 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 lower right of page 9.24 in Note 33. [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 elements (or 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 elements (or 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) [on the upper half of page 9.8] for those non-focal explanatory variate(s) Ƶ$_i$ 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 Ƶ$_i$ 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 the discussion on page 9.16 of Table 9.2.6.

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.

NOTES: 13. Where the definitions given above for blocking and matching refer to values of non-focal explanatory variates being the *same*, in practice the values may only be *similar*.

14. 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 9.2.3 at the right – but a *block* of two elements (or units) may also be referred to as a 'pair.'

Table 9.2.3
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.

15. In DOE, non-focal explanatory variate(s) made the same within blocks are called **blocking factor(s)**; in data-

(*continued overleaf*)

NOTES: (cont.) 15. 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).

16. 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 33 on pages 9.24 and 9.25.

17. *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 – recall Appendix 5 on page 6.12 in Figure 6.1.

Elsewhere, *both* processes may be called 'stratifying'. There is further discussion of subdividing in Section 10 near the bottom of page 9.15 and on page 9.16 and of stratifying near the bottom of page 9.16 near the end of Note 19.

8. Plan Components to Manage Comparison Error

Comparison error in comparative investigating, introduced on page 9.6 in Section 2, arises from confounding by non-focal explanatory variate(s); background information and Plan components to manage comparison error are then discussed in Sections 3 to 7 on pages 9.6 to 9.14. These Plan components are summarized in Table 9.2.4 below.

Table 9.2.4

Plan Component		Error category	Error Management Strategy
Question with a causative aspect	Experimental Plan	Comparison	● **Blocking:** forming groups of units with the *same* values of one or more non-focal explanatory variates; the units within a block are then assigned *different* values of the *focal* variate. ● **Equiprobable assigning:** a *probabilistic* mechanism used to assign the value of the focal explanatory variate to the units: – within each block in a blocked Plan; – in the sample in an *un*blocked Plan. ○ **Blinding participants and treatment administrators:** by withholding from participants and treatment administrators knowledge of which group a participant is in, these two blindings try (like *equiprobable assigning*) to manage factors which may promote differences in averages of unknown and unmeasured non-focal explanatory variates in the (treatment and control) groups whose (average) response variate is being compared. [Management of **comparison** error.] ⊙ **Blinding treatment assessors** tries (like making measurements *independent*) to prevent the assessors' other knowledge from improperly influencing their assessment of participants' health status. [Management of **measurement** error.]
	Observational Plan		● **Matching:** forming groups of units with the *same* values of one or more non-focal explanatory variates but *different* values of the *focal* variate. ○ **Subdividing:** a form of matching in which 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 9.2.6 at the bottom of the facing page 9.15 discussed on page 9.16.

9. Experimental Plans – Sample selecting and Blocking

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.

(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 5)

– 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 aspirin may be much more similar among doctors and all males. (Two newspaper reports of this investigation are reprinted in Figure 9.18 of these 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 when answering a Question with a *descriptive* aspect but is still far from the statistical ideal.

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:

∗ 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 – recall Section 9 (Appendix 3) on pages 6.7 to 6.11 in Figure 6.1;

∗ more generally, the theoretical basis is gone for interpreting formal methods of data analysis like confidence intervals and tests of significance. [Both limitations are commonly overlooked 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 9.18 of the 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 6 on page 9.9].

∗ 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: 18. 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 9.2.5 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 9.2.5

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

The comments in Section 9 (on the facing page 9.14 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 9.2.6 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

Table 9.2.6	**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

(*continued overleaf*)

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.

The bottom line of Table 9.2.6 overleaf 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 (common cause) confounding by a genetic factor which determines 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 9.2.6 raises the possibility (*not* realized here) of the phenomenon known as Simpson's Paradox (and its accompanying limitation imposed on an Answer) – see Figure 9.8 on pages 9.53 to 9.60.

NOTE: 19. 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 Figure 9.13 on pages 9.73 to 9.76 – recall also the discussion overleaf near the top of page 9.15.
  - 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 15 near the top of page 9.14.
    - + When controls are selected *non*-probabilistically, there is no theoretical basis for an inverse relationship

(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 6)

NOTE: 19. ● – + between sampling imprecision and (the square root of) of the sample size – see Appendix 3 on pages 9.35 to 9.37 – so there is no *statistical* reason why a larger sample size for controls will decrease comparing imprecision.
(cont.)

+ The blocks in a blocked experimental Plan are also often selected *non*-probabilistically but, as discussed in Figure 9.13 on pages 9.73 to 9.76, 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.

EM0424: The Globe and Mail, August 6, 2004, page A11

Death rate higher near busy roads

BY STEPHEN STRAUSS

Canadian scientists have found a startling rise in death rates associated with nothing more perilous than living within 50 metres of a major highway and 100 metres of a city road that carries a slew of polluting cars and trucks.

While there have been a number of studies tying surges in deaths to city air pollution in general, what the researchers at McMaster University uncovered was a roughly 18-percent spike in mortality in the Hamilton area among people who lived adjacent to streets carrying 35,000 to 75,000 vehicles daily.

The rise in the pollution death rate did not come from asthma, emphysema or lung cancer but from heart attacks and other heart conditions. "Basically air pollution does not affect your lungs but your heart," is how Murray Finkelstein of McMaster's program in occupational health and environmental medicine, and a co-author of the new study, describes what his group has found.

The reason for the large heart-disease hit is still uncertain, but Dr. Finkelstein points to research in animals that suggests air pollution particles can irritate arteries and lead to their general hardening and thickening.

Although the study, which was published in the July issue of the *American Journal of Epidemiology*, focused on the roads and highways of Hamilton, the researchers see no reason why the findings shouldn't apply to city dwellers perched above traffic surging along St. Lawrence Street in Montreal, Yonge Street in Toronto, or Hastings Street in Vancouver. Not to mention anyone whose dwelling is on the skirt of the traffic behemoth known as the Trans-Canada Highway, which, as it moves 400,000 people a day in some locations, is North America's second-busiest highway.

What the researchers also did is translate the increasing death rates, which have a relatively small impact on younger people, into something closer to an insurance company's life-expectancy table. They found there is a 2.5-year increase in age-related death levels for people whose dwellings are located cheek-to-jowl with heavy traffic.

But there is also a class-related confounding factor to the data. ..... more poor people may be more likely to live near busy streets than rich people, and poor people have other behaviours – smoking in particular – that might kill them in larger numbers.

"Basically, that means your mortality pattern if you are 50 years old is the same as someone 52.5 years old who doesn't live on a busy road," said Dr. Finkelstein. What is even more sobering is the fact that the deadliness of living near major thoroughfares is not far off the life-shortening effects of such known killers as diabetes or chronic lung disease.

The McMaster scientists say their research leads to a very simple bit of advice for a health-conscious individual. "If you have a heart condition, I would advise not buying a place very close to major roadways or highways," said Michael Jerrett, a McMaster University geography professor who is another co-author on the study.

He also suggests that susceptible people who live close to the busy thoroughfares consider air purification systems in their homes as a preventative act.

There some some caveats to the new study, which replicates Dutch research published two years ago. There was no direct measure of how much higher the motor-vehicle related pollution was near major roads. This omission should be remedied next month when the McMaster group tracks road-pollution level themselves.

But there is also a class-related confounding factor to the data. Because of existing concerns over noise and pollution, Dr. Finkelstein says more poor people may be more likely to live near busy streets than rich people, and poor people have other behaviours – smoking in particular – that might kill them in larger numbers.

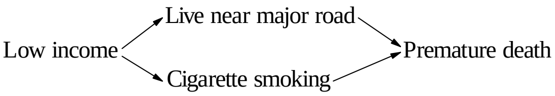
REFERENCE: Finkelstein, M.M., Jerrett, M. and M.R. Sears: Traffic Air Pollution and Mortality Rate Advancement Periods. *American Journal of Epidemiolgy* **160**(#2): 173-177, July 15 (2004). [UW Library E-journal]

The abstract given in the original article is:

Chronic exposure to air pollution is associated with increased mortality rates. The impact of air pollution relative to other causes of death in a population is of public health importance and has not been well established. In this study, the rate advancement periods associated with traffic pollution exposures were estimated. Study subjects underwent pulmonary function testing at a clinic in Hamilton, Ontario, Canada, between 1985 and 1999. Cox regression was used to model mortality from all natural causes during 1992-2001 in relation to lung function, body mass index, a diagnosis of chronic pulmonary disease, chronic ischemic heart disease or diabetes mellitus, household income, and residence within 50 m of a major urban road or within 100 m of a highway. Subjects living close to a major road had an increased risk of mortality (relative risk = 1.18, 95% confidence interval: 1.02, 1.38). The mortality rate advancement period associated with residence near a major road was 2.5 years (95% confidence interval: 0.2, 4.8). By comparison, the rate advancement periods attributable to chronic pulmonary disease, chronic ischemic heart disease, and diabetes were 3.4 years, 3.1 years, and 4.4 years, respectively.

(*continued overleaf*)

11. Comparative Plans and Causal Structures



With the additional background given in Sections 6 to 10 on pages 9.12 to 9.17, the causal structure at the right provides a convenient context for discussing cases (1), (8), (9) and (11) from Section 5 on pages 9.10 and 9.11; this context comes from an investigation whose newspaper report is reprinted overleaf on page 9.17. This investigation found that death rates were higher for people who lived near a major road – in our terminology, living near a major road (focal variate $\mathbf{X}$ with two values: living near such a road and not doing so) is associated with a higher death rate [an attribute of response variate $\mathbf{Y}$ with two values (alive or dead), quantified in this investigation as a 'mortality rate advancement period' (MRAP)].

(11) $\mathbf{Z}_j\begin{smallmatrix}\nearrow \mathbf{X} \searrow\\ \searrow \mathbf{Z}_i \nearrow\end{smallmatrix}\mathbf{Y}$

Case (11): The causal structure at the right above is an instance of case (11) [shown again at the right], although the investigation (described overleaf on page 9.17) of the effect of living near a major road was not explicitly concerned with $\mathbf{Z}_j$ (low income). However, this causal structure does not raise Questions that are not also raised by the three other cases (1), (8) and (9) discussed below, because it is a composite of them (and other) cases as follows:

- the left-hand side has the so-called 'common cause' structure of case (9),
- the right-hand side has the so-called 'common response' structure of case (8),
- the top and bottom are the causal chains of cases (4) and (6) [which is which depends on variate assignment];

there are also four instances of case (1): $\mathbf{Z}_j \to \mathbf{X}$, $\mathbf{Z}_j \to \mathbf{Z}_i$, $\mathbf{X} \to \mathbf{Y}$, $\mathbf{Z}_i \to \mathbf{Y}$.

Cases (1), (8) and (9): These cases involve five Questions that could be investigated; they are numbered 1 to 5 for convenient reference and given with other information in Table 9.2.7 below – 'E' or 'O' in the Plan column denotes 'experimental' or 'observational'.

Table 9.2.7

No.	Case	Question	Plan
1	(1) $\mathbf{X} \to \mathbf{Y}$	Is low income associated with premature death?	O
2	(8) = (1) $\mathbf{X} \to \mathbf{Y}$	Is living near a major road associated with premature death?	O
3	(8) = (1) $\mathbf{X} \to \mathbf{Y}$	Is cigarette smoking a cause of premature death?	O
4	(9) $\mathbf{Z}\begin{smallmatrix}\nearrow \mathbf{X}\\ \searrow \mathbf{Y}\end{smallmatrix}$	Are living near a major road and cigarette smoking associated with low income?	O
5	(8) $\begin{smallmatrix}\mathbf{X}\searrow\\ \mathbf{Z}\nearrow\end{smallmatrix}\mathbf{Y}$	To what extent are living near a major road and cigarette smoking associated with premature death?	O

Question 1: It has long been known that the answer to this Question is *Yes* – for example, nineteenth century vital statistics in the U.K. showed an association between 'social class' and death rates. The inability of the investigator(s) to assign the value of the focal variate $\mathbf{X}$ (a person's income) is why the Plan can only be observational and the Question is phrased in terms of association (rather than causation).

Question 2: This Question is answered by the investigation whose newspaper report is given overleaf on page 9.17. As the report points out, possible confounding by a lurking variate $\mathbf{Z}$ (cigarette smoking) means that comparison error imposes a severe limitation on the Answer.

Question 3: The health consequences of cigarette smoking are now well documented as a result of tens of thousands of investigations, most of them from around 1950 and later. The Plans of investigations involving humans have been observational because investigators cannot ethically (or practically) assign elements' smoking habits; *experimental* Plans have been limited to investigations involving animals, but they are relatively few in number, in part because of the difficulty (and, hence, the cost) of getting animals to smoke. [Another factor is the limited lifespans of cheaper laboratory animals (like mice and rats) in relation to the time for some health effects of smoking to become apparent.]

The Question wording involves *causation* because of the requirement that manipulation of the focal variate (reducing the prevalence of cigarette smoking) will produce a desired change in the response variate (a reduction in smoking-induced disease, resulting in better public health and reduced healthcare costs). The decades of research and the number of investigations of the health consequences of smoking is a reminder of the difficulties of establishing causation using an observational Plan.

Question 4: This Question involves $\mathbf{Z}$ (low income) as a *common cause* of $\mathbf{X}$ (living near a major road) and $\mathbf{Y}$ (cigarette smoking), although the Question wording involves (the weaker) association rather than causation – more appropriate notation would be $\mathbf{X}$ instead of $\mathbf{Z}$ and $\mathbf{Y}_1$ and $\mathbf{Y}_2$ instead of $\mathbf{X}$ and $\mathbf{Y}$.

Question 5: This Question involves the effects of *two* focal variates on a response variate, which is case (8) but with $\mathbf{X}_1$ and $\mathbf{X}_2$ in place of $\mathbf{X}$ and $\mathbf{Z}$ – see also the discussion on the upper half of page 9.12 of *quantifying* the relationship of two (or more) focal variates and a response variate.

In summary, four causal structures are introduced in the discussion of statistical association of explanatory variates at the upper right of page 9.10 in Section 5; these become twelve structures at the middle right of page 9.10 with the inclusion of the response variate. The discussion on page 9.10 below these structures and in this Section 11 shows that only *four* of these twelve are relevant to comparative Plans, for which the *primary* concern is the structure of case (1); the overlapping structures of cases (8), (9) and (11) serve mainly to inform case (1) investigating.

- The discussion in this Section 11 reminds us that, to develop a comparative Plan to answer a Question with a causative aspect, sufficient *extra*-statistical knowledge is needed to:

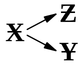
(*continued*)

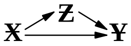
Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 7)

- frame a (clear) Question about the association being investigated;
- give a plausable causal structure that is appropriate for this Question;
- choose (and then develop) a feasible Plan type.

NOTES: 20. The *observational* nature of the Plans in Table 9.2.7 on the facing page 9.18 reflects their *context*; in contexts where the value of the focal variate(s) *could* be assigned by the investigator(s), the Plans could be experimental.

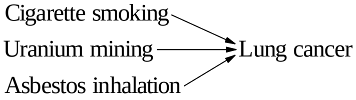
- ● It is also context-dependent whether the Questions involve *establishing* causation, *quantifying* (causal) relationships or *prioritizing* causes – recall Section 6 on page 9.12.

(10) 



21. Case (10) [shown again at the right] in its lower *real* form (because **Z** is an explanatory variate) is *not* a viable basis for a comparative Plan, which requires either:

- ● *direct* causation of **Y** by **X** [case (1)], OR:
- ● an *explicit* intermediate variate in *each* causal chain from **X** to **Y**, as in case (11) and illustrated at the start of Section 11 in the causal structure at the top right of the facing page 9.18.
  - There may be *more* than two causal chains from **X** to **Y**, as illustrated by a case of *three* possible intermediaries at the right. The may also be *interaction* (see page 9.20) among such explanatory variates; for example, the increased risk of lung cancer among uranium miners (presumably due to radioactive dust inhalation) might be mainly among smokers but *both* non-smokers and smokers may be at an increased risk of lung cancer from asbestos inhalation.
    - + Causes of 'lung cancer' are actually more complicated than implied by this example, because there are a *number* of such cancers involving different cell types (*e.g.*, mesothelioma from asbestos inhalation).



22. The newspaper article (reprinted on page 9.17) discussed in Section 11 on the facing page 9.18 is concerned with higher death rates among people who live near a major road (the focal variate in an investigation with an *observational* Plan). The abstract (on page 9.17 below the newspaper article) of the *journal* article mentions that six possible confounders – lung function, body mass index, household income, a diagnosis of chronic pulmonary disease, chronic ischemic heart disease and diabetes – were considered in the investigation as other possible factors in premature death, and larger mortality rate advancement periods (MRAPs) than for the focal variate were found for the last three of these variates using a model called Cox regression [analogous to the response model (9.2.1) on page 9.5 but differing in mathematical form]. Such modelling manages comparison error by trying to achieve the statistical benefit of *blocking* in an experimental Plan by *mathematically* (rather than physically) holding some (here, six) lurking variates 'fixed' as **X** changes. Such modelling encounters two difficulties.

- ● Like blocking, it manages comparison error *only* for confounder(s) which are identified *explicitly*:
  - for use as blocking factor(s) or inclusion in the model AND: – whose values it is feasible to measure.

  [This is true also for the confounder(s) used for matching and subdividing in an observational Plan.]
- ● It must be assumed that the model has the *correct* (or an *adequate*) *form* for each possible confounder in its structural component – for example, a first power, a second power, a square root, a logarithm; any **Z** for which this is not so will *not* be held 'fixed' by the model calculations and so can become a source of model error.
  - Holding the *same* (or *similar*) values *physically* (in an experimental Plan) for other explanatory variates as the focal variate changes manages comparison error more effectively than holding them fixed by means of the *model* (and using data from an observational Plan). Likewise, Answers which claim causation based on *physical* evidence have less severe limitation than those based on a *model*.

The similarity of intent between blocking in an experimental Plan and including possible confounders in the structural component of a response model for an observational Plan continues on by managing, *under repetition*, comparison error due to unblocked, unmeasured and unknown confounders:

- ● physically, by probability assigning (*e.g.*, EPA) in an experimental Plan – the greater the degree of *replicating*, the greater the reduction in comparing imprecision due to such confounders;
- ● mathematically, by the residuals [and their (sub)model] in the response model for an observational Plan (but at the cost of model error becoming one of the components of overall error).

It is *im*material in the *model* for the investigation described on page 9.17 whether we regard the seven explanatory variates as seven focal variates or as one focal variate and six possible confounders.

12. Comparative Plans – The Protocol for Setting Levels and Interaction

The protocol for setting levels specifies the *values* to be taken by relevant explanatory variate(s); 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.

(*continued overleaf*)

- ∗ 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 28 on the facing page 9.21.

  Our concern in this Section 12 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 (or element) [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 $\mathbf{X}$ on $\mathbf{Y}$ (usually) refers to the change in the *average* of $\mathbf{Y}$ for *unit* change in $\mathbf{X}$ and:
  - – implies the $\mathbf{X}$-$\mathbf{Y}$ relationship is (believed to be) *causal* – a change in $\mathbf{X}$ *causes* (brings about) a change in $\mathbf{Y}$;
  - – 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 $\mathbf{X}$ 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 $\mathbf{Y}$ 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 in the discussions of the Appendix (Section 8) on pages 9.62 to 9.64 in Figure 9.8 and in Figures 9.12 on pages 9.73 to 9.76 and 9.13 on pages 9.77 to 9.80.

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: 23. Not all experimental Plans lead to an Execution stage (of the FDEAC cycle) 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*.

24. Equiprobable assigning of treatments to elements (or 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.

25. 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) on pages 9.67 and 9.68 in Figure 9.9.

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

(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 8)

NOTES: (cont.) 25. *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.

26. When there are two or more focal variates, 'lurking variates' criterion (1) on the upper half of page 9.8 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 27 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 **X**$_1$ and **X**$_2$ set 'Lo', another run with **X**$_2$ set 'Hi' and another with **X**$_2$ back at 'Lo' and **X**$_1$ set 'Hi'; the resulting data, shown as three response variate averages in Table 9.2.8 at the right, do *not* allow the **X**$_1$-**X**$_2$ interaction effect to be estimated, because there is no run with both factors set 'Hi'.

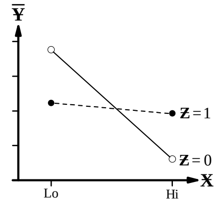
Table 9.2.8	$\mathbf{X}_2$ Lo	$\mathbf{X}_2$ Hi
$\mathbf{X}_1$ Lo	$\overline{\mathbf{Y}}_{\mathrm{Lo,Lo}}$	$\overline{\mathbf{Y}}_{\mathrm{Lo,Hi}}$
$\mathbf{X}_1$ Hi	$\overline{\mathbf{Y}}_{\mathrm{Hi,Lo}}$	No data

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.

27. The idea in Note 25 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 (introduced on page 9.6 in Section 2), where the association of **X** and **Z** typically has a correlation with magnitude *less* than 1. As discussed in Figure 9.9 on pages 9.61 to 9.64, both cases are usually (unwisely) simply called 'confounding' without distinction.

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



- ○ **Y** is tile size *after* firing in an oven,
- ○ **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,
- ○ **Z** is amount of clay in the ingredient mix used for the tiles,

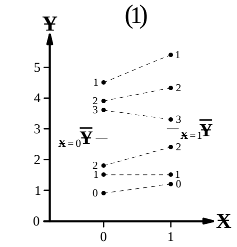
by managing the amount of clay in the ingredient mix (*i.e.*, setting **Z** = 1), the manufacturing process is improved by making variation in tile final size *less* sensitive to variation in firing temperature due to tile position within the oven. This *indirect* approach exploiting interaction avoids 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 **Z** = 1 and clay must not be too expensive an ingredient.

13. Experimental Plans – Quantifying a Treatment Effect Under EPA

To illustrate properties of *experimental* Plans (and then contrast them with those of observational Plans), hypothetical data for a response variate **Y** are given in Table 9.2.9 at the right for a respondent population of six elements under two values [assigned by the investigator(s)] of a focal variate **X** – the treatment effect (the change in the average of **Y** for unit change in **X** when all the **Z**s remain fixed) is 0.3 units, the average of (widely-varying) effects of changing **X** for the individual elements. The data in Table 9.2.9 are also shown in diagram (1) at the right; the value of a lurking variate **Z** given beside each dot reminds us that, for our initial discussion, changing **X** does *not* affect the value of **Z** [but see the comment in the second bullet (⊙) in the second paragraph overleaf on page 9.22]. The population averages when **X** = 0 and **X** = 1 are shown as

Table 9.2.9: Respondent Population Responses (N = 6)

Element no.	1	2	3	4	5	6	**Av.**
X = 0	0.9	1.5	1.8	3.6	3.9	4.5	2.7
X = 1	1.2	1.5	2.4	3.3	4.2	5.4	3.0
Treatment effect	0.3	0	0.6	−0.3	0.3	0.9	**0.3**



(*continued overleaf*)

short horizontal lines; the differing notation used for these averages between diagrams (1) [overleaf] and (2) [at the middle right of page 9.25] is to emphasize the distinction between experimental Plans [where the *investigator(s)* assign each element's **X** value (under EPA)] and observational Plans [where each element has its 'natural' **X** value *un*influenced by the investigator(s)].

Table 9.2.10 at the right below (which is *un*blocked – see Note 29 below) shows the twenty possible assignments of the six population elements whose data are given in Table 9.2.9 overleaf on page 9.21, together with their response variate averages, treatment effect and comparison error; for example, the first line of Table 9.2.10 shows:

- elements 1, 2 and 3 assigned **X** = 0 (often called the 'control group') with average response 1.4,
- elements 4, 5 and 6 assigned **X** = 1 (the 'treatment group') with average response 4.3,
- for this assignment, an estimated treatment effect $\overline{y}_1 - \overline{y}_0$ is of 4.3 – 1.4 = 2.9,
- for this assignment, comparison error of 2.6 – the difference between the estimated and true treatment effects, 2.9 and 0.3;

the five averages at the bottom of Table 9.2.10 have meaning only if all 20 assignments are *equi*probable (as they are under EPA).

Table 9.2.10: Data for the Set of All 20 Equiprobable Assignments of the 6 Elements in Table 9.2.9 (on page 9.21)

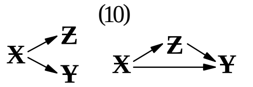
Element nubers		Averages		Treatment	Comparison	
X = 0	**X** = 1	**X** = 0	**X** = 1	effect	error	
(1, 2, 3)	(4, 5, 6)	1.4	4.3	2.9	2.6	*2.8*
(1, 2, 4)	(3, 5, 6)	2.0	4.0	2.0	1.7	
(1, 2, 5)	(3, 4, 6)	2.1	3.7	1.6	1.3	*1.5*
(1, 2, 6)	(3, 4, 5)	2.3	3.3	1.0	0.7	*0.9*
(1, 3, 4)	(2, 5, 6)	2.1	3.7	1.6	1.3	
(1, 3, 5)	(2, 4, 6)	2.2	3.4	1.2	0.9	*1.1*
(1, 3, 6)	(2, 4, 5)	2.4	3.0	0.6	0.3	*0.4*
(1, 4, 5)	(2, 3, 6)	2.8	3.1	0.3	0	
(1, 4, 6)	(2, 3, 5)	3.0	2.7	–0.3	–0.6	
(1, 5, 6)	(2, 3, 4)	3.1	2.4	–0.7	–1.0	*–0.8*
(2, 3, 4)	(1, 5, 6)	2.3	3.6	1.3	1.0	
(2, 3, 5)	(1, 4, 6)	2.4	3.3	0.9	0.6	*0.8*
(2, 3, 6)	(1, 4, 5)	2.6	2.9	0.3	0	*0.2*
(2, 4, 5)	(1, 3, 6)	3.0	3.0	0	–0.3	
(2, 4, 6)	(1, 3, 5)	3.2	2.6	–0.6	–0.9	
(2, 5, 6)	(1, 3, 4)	3.3	2.3	–1.0	–1.3	*–1.1*
(3, 4, 5)	(1, 2, 6)	3.1	2.7	–0.4	–0.7	
(3, 4, 6)	(1, 2, 5)	3.3	2.3	–1.0	–1.3	
(3, 5, 6)	(1, 2, 4)	3.4	2.0	–1.4	–1.7	*–1.5*
(4, 5, 6)	(1, 2, 3)	4.0	1.7	–2.3	–2.6	
	Av.	2.7	3.0	**0.3**	**0**	*0.1*

[The last column of ten values in *italics* at the right of Table 9.2.10 is discussed in the second bullet (⊙) below.]

The averages at the bottom of Table 9.2.10 illustrate several matters of statistical interest about EPA and (incidentally) about EPS.

- ⊙ under EPA, the average treatment effect over the 20 possible assignments is the *true* value, 0.3, SO THAT:
- ⊙ under EPA, the average comparison error over the 20 possible assignments is *zero* – that is, there is *un*biased estimating of the treatment effect; HOWEVER:
  - – if lurking variate **Z** and the response variate **Y** are a common response to **X** [recall case (10) of the causal structures on page 9.10 and Note 21 on page 9.19], this unbiasedness is lost, as the following illustration shows.

(10)



    - ○ Suppose the change in **Z** (resulting from the change in **X**) causes element 4 (with **Z** = 3) to have a response of 3.9 and an apparent effect of 0.3 instead of its 'true' value of –0.3 – for simplicity, we assume the other five elements (with **Z** values *other than* 3) still have the effects given for **X** = 1 in Table 9.2.9. The 10 assignments involving element 4 with **X** = 1 then have their averages increased by 0.2, as do the corresponding comparison error values (given in *italics* in the last column of Table 9.2.10); the *average* of the *twenty* comparison error values is then 0.1 instead of zero, indicating biased estimating of the treatment effect.
    - ○ If elements other than 4 were to *also* have their responses when **X** = 1 changed by the change in **Z**, some of these changes might be in *opposite* directions, resulting in some *cancellation* and a *smaller* (conceivably zero) magnitude for the comparison error of the *particular* assignment; however, the estimating bias (the *average* comparison error over the set of *all* possible assignments) is unlikely to be meaningfully changed by such (fortuitous) cancellation.
- The average of the set of 20 samples of three elements with a given **X** value is the relevant *population* average – see the right-hand column of Table 9.2.9 at the start of Section 13 overleaf on page 9.21 – this is unbiased estimating of a respondent population average under EPS (see also Appendix 3 on pages 9.35 and 9.36 and more detail on pages 6.7 and 6.8 in Appendix 3 in Figure 6.1 of these Course Materials).

Thus, experimental Plans provide unbiased estimating of the treatment effect unless one or more of the lurking variates $\mathbf{Z}_1$, ..., $\mathbf{Z}_k$ and the response variate **Y** are a common response to the focal variate **X**; *if* this state of affairs is *un*common in practice, an experimental Plan usually avoids such biased estimating. [Blocking factors are clearly *not* such a common response because they are held *fixed* when **X** is changed.]

NOTES: 29. If the responses in Table 9.2.9 at the start of Section 13 overleaf on page 9.21 were *real* data, the Answer about the value of the treatment effect could be made more useful by managing the substantial variation among the elements' effects – *e.g.*, by blocking to decrease comparing imprecision [recall the comment (–) near the middle of page 9.13].

30. To avoid confounding, 'lurking variates' criterion (1) on the upper half of page 9.8 requires the ideal of *all* non-focal explanatory variates $\mathbf{Z}_i$ holding their values for *every* population element when **X** = 0 and **X** = 1.
    - Blocking meets this criterion but *only* for the $\mathbf{Z}_i$ that are blocking factor(s).
    - Provided there is no common response of $\mathbf{Z}_1$, ..., $\mathbf{Z}_k$ and **Y** to **X**, EPA addresses criterion (1) for the *other* unblocked, unmeasured and unknown lurking **Z**s but it does so only *under repetition* – making their distributions (not their values *individually*) the same *on average* across elements when **X** = 0 and **X** = 1.

(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 9)

NOTES: (cont.) 30. ● – The *probabilistic* nature of equiprobable assigning means that, even in conjunction with adequate replicating, it cannot *guarantee* (roughly) the same distribution among groups for *every* unblocked, unmeasured and unknown non-focal explanatory variate – under a *particular* assignment, some such distribution(s) may differ substantially among the groups being compared; however, the degree of the resulting limitation imposed on Answer(s) by comparison error becomes:

+ *more* acceptable as the level of replicating (*i.e.*, the group sizes) increases;

⊙ *less* acceptable as the number of lurking variates (whose effects are to be 'balanced') increases.

– There may sometimes be data available on one or more $\mathbf{Z}_i$ that allow some assessment of the balance in the assignment obtained under EPA in a *particular* investigation. Two illustrations from clinical trials are:

+ in the usual situation where participants' sex is recorded, it is possible to check how close the female-male ratios are in the control and treatment groups (and how close both are to the ratio in the study population); when participants' age is recorded, the *average* age in the control and treatment groups can be compared.

Depending on how early in an investigation any (meaningful) imbalance is identified, investigator(s) may:

+ re-do the equiprobable assigning, OR:

+ redress the effect(s) of the imbalance.

Comparing *average* age (say) is a check for similar age *distributions* among the groups but a limitation on Answer(s) due to comparison error remains because distributions with *different* shapes or widths may have the *same* (or similar) averages.

– Other (*un*desirable) language sometimes used to describe how EPA addresses criterion (1) on page 9.8 is: EPA in conjunction with adequate replicating, tries to *remove association* (or *produce 'independence'*) between the focal variate and unblocked, unmeasured and unknown non-focal explanatory variates.

EPA epitomizes the *active* nature of experimental Plans and, in addressing criterion (1) for unblocked, unmeasured and unknown non-focal explanatory variates, confers (under repetition) a *unique* advantage on experimental Plans over observational Plans; probability assigning is what most clearly distinguishes the two Plan types.

31. Statisticians have argued about whether EPA is 'necessary and/or sufficient' in an experimental Plan to establish a *causal* relationship between **X** and **Y**. The disagreements are resolved when it is recognized that:

● EPA operates *probabilistically* and in *conjunction* with adequate replicating – as discussed in Note 30 on the facing page 9.22 and above, non-focal explanatory variates may differ in their values, among the groups being compared, to a degree that can meaningfully change the Answer under the assignment obtained in a *particular* investigation – for instance, in Table 9.2.10 on the facing page 9.22, the first and last assignments have comparison error of substantial magnitude in the context of the hypothetical data in Table 9.2.9 on page 9.21;

● the *mathematical* language of *necessity and sufficiency* is *in*appropriate in the context of investigative *uncertainty* and so a statement like *equiprobable (or 'random') assigning is neither necessary nor sufficient to establish causation* may be true but is unhelpful because it can obscure the following two matters:

– proper use of statistical methods does not *guarantee* a 'correct' Answer – it merely makes an Answer *likely* to be close enough to the actual state of affairs to be useful (*i.e.*, proper use of statistical methods yields an Answer with *acceptable* limitations);

– *im*proper use of statistical methods does not *guarantee* a 'wrong' answer – it may (occasionally) yield a 'correct' Answer; for instance, a response variate measured *in*accurately or *in*correctly on a sample of *one* unit may happen to be close (conceivably *equal*) to the value of the respondent population average.

It is difficult to develop a mind-set in which these matters are routinely recognized; the difficulty is compounded by that of framing in English clear and correct statements that deal with uncertainty in statistics.

● It is also challenging routinely to recognize and express the fact that, in statistics, we quantify uncertainty *only* in terms of behaviour under *repetition* – Answer(s) obtained in a *particular* investigation *remain* uncertain, as reflected by their limitations. Limitations on Answers are *unavoidable* when using *in*complete information, which arises most obviously in statistics from the processes of sampling and measuring.

– The idea of limitations also reminds us to avoid phrases like *the validity of a causal inference* instead of (disciplined) use of the terminology of comparison error to reflect one source of investigative uncertainty.

REFERENCE: Sprott, D.A., R.M. Royall in *Recent Concepts in Statistical Inference.* Proceedings of a Symposium in Honour of Professor V.P. Godambe, University of Waterloo, August 14-16, 1991, Randomization Discussion.

32. Two illustrations of the matters in Note 31 above in the context of *non*-probability assigning are:

○ Program 12 of *Against All Odds: Inside Statistics* describes (about 14 minutes into the video) a clinical trial of ribavirin as treatment for a pre-AIDS condition, swollen lymph nodes; the data for the three groups are shown in Table 9.2.11 at the right. The *de*creasing number of cases that progressed on to AIDS with increasing daily

Table 9.2.11: Ribavirin Trial Data

	RIBAVIRIN (mg/day)		
	0	600	800
Group size	52	55	56
Progress to AIDS	10	6	0

(*continued overleaf*)

NOTES: 32. ○ ribavirin dose indicated it was an effective treatment. Later, it transpired that ribavirin is *not* effective – the
(cont.) data were an artifact of the sickest patients being assigned to the control group and the healthiest to the group receiving the higher dose of ribavirin.

⊙ Scurvy is a disease caused by a deficiency of vitamin C in the diet; it is characterized by debility, blood changes, spongy gums and hemorrhages in bodily tissues. Up to the nineteenth century, it was common among sailors on long voyages, soldiers on campaign, inhabitants of beleagured cities and in other such situations where fresh fruit and/or vegetables in the diet were absent or insufficient. As illustrations:

- during Anson's circumnavigation voyage in 1742-1744 (a period *prior* to Lind's 1747 investigation described below), at least 380 of a crew of 510 on one of his six ships died of scurvy; BY CONTRAST:
- on his second voyage in 1772-1775, covering 70,000 miles over more than 1,000 days, Cook (who knew of Lind's investigation and acted on it) lost only 3 men to accidents and 1 to 'consumption' from a crew of 118.

⊙ Lind had direct experience of scurvy because he first went to sea with the British Navy in the late 1730s; he spent many years investigating its cause. *Our* interest in Lind's work is because, in 1747, he used an *experimental* Plan to investigate possible treatments; during a voyage which included a ten-week absence from shore and in which 80 of a crew of 350 sailors were struck down by scurvy, Lind used a sample of 12 sailors with scurvy, which he divided into groups of two for administering the following six daily treatments:

- two quarts of cider;
- 25 drops of elixir of vitriol;
- half a pint of sea water;
- six spoonfulls of vinegar;
- two oranges and one lemon;
- a garlic, mustard seed, balsam and myrrh gum electuary.

Parts of Lind's description of his investigation, from the reference below, are:

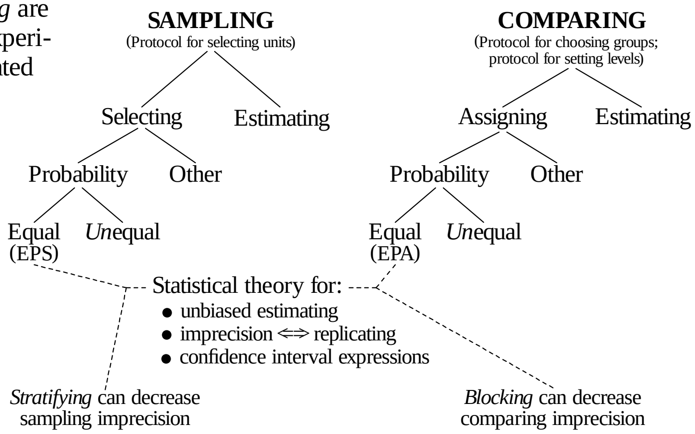
> On the 20*th* of May, 1747, I took twelve patients in the scurvy, on board the *Salisbury* at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakened knees. They lay together in one place, ... and had one diet common to all, ... Two of the worst patients, with tendons in the ham rigid, (a symptom none of the rest had), were put under a course of sea-water. .....
>
> The consequence was, that the most sudden and visible good effects were perceived for the use of the oranges and lemons; one of those who had taken them, being at the end of six days fit for duty. ..... The other was the best recovered of any in his condition; ..... .
>
> Next to the oranges, I thought the cyder had the best effects. ..... those who had taken it, were in a fairer way of recovery than the others at the end of a fortnight, which was the length of time all these different courses were continued, except the oranges. .....
>
> As to the elixir of vitriol, I observed that the mouths of those who had used it by way of gargling, were in a much cleaner and better condition than many of the rest, especially those who used the vinegar; but perceived otherwise no good effects from its internal use upon other symptoms. .....
>
> There was no remarkable alteration upon those who took the electuary, the sea-water, or vinegar, upon comparing their condition, at the end of the fortnight, with others who had taken nothing but a little lenative electuary and cream of tartar, ...
>
> It may be now proper to confirm the efficacy of these fruits (oranges and lemons) by the experience of others.

⊙ In the context of the Conclusion stage of the FDEAC cycle, because Lind obtained what is now known to be a *correct* Answer, it is easy to overlook the severe *limitations* on his Answer imposed by:

- the small sample size of 12 sailors;
- the non-probability selecting: likely *convenience* selecting of sailors who were on the ship and had scurvy;
- the non-probability assigning – not surprisingly, there is no mention by Lind of the 'modern' idea of probability assigning (*e.g.*, EPA) but some implication of *judgement* assigning in the description quoted above.

REFERENCE: Tröhler, U. (2003). James Lind and scurvy: 1747 to 1795. The James Lind Library (www.jameslindlibrary.org). Republished in the *J. Roy. Soc. Medicine* **98**: 51-522 (2005). [DC Library call number: PER R35.R7]

33. Equiprobable *selecting* and equiprobable *assigning* are components of the processes of sampling and (experimental) comparing, whose *similarities* are illustrated in the discussion on page 9.22 of Table 9.2.10 and are portrayed by the two tree diagrams in the schema at the right.


SAMPLING
(Protocol for selecting units)
COMPARING
(Protocol for choosing groups; protocol for setting levels)
Selecting
Estimating
Assigning
Estimating
Probability
Other
Probability
Other
Equal (EPS)
*Un*equal
Equal (EPA)
*Un*equal
Statistical theory for:
● unbiased estimating
● imprecision ⟺ replicating
● confidence interval expressions
Stratifying can decrease sampling imprecision
Blocking can decrease comparing imprecision

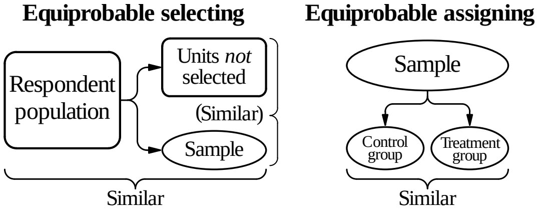
- Investigations involving *comparing* (to answer a Question with a *causative* aspect) usually involve *sampling*; investigations involving *sampling* to answer a Question with a *descriptive* aspect need *not* involve *comparing*.
- **Probability selecting** means having *known* element (or unit) inclusion probabilities in the selecting process;

(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 10)

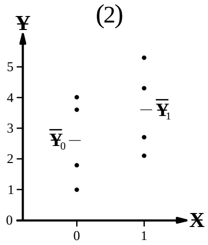
NOTES: (cont.) 33.
- introductory statistics courses emphasize *equi*probable selecting as the basis of statistical theory for the behaviour of *sample* error under repetition (see also Appendix 3 on pages 9.35 to 9.37).
  - Here, we coin the term **probability assigning** for having known *assigning* probabilities; *we* encounter mainly the special case of *equi*probable assigning – (roughly) *equal* numbers of elements 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 – recall the discussion on page 9.22 of Table 9.2.10.
    - + 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 lower right of the facing page 9.24; 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 lower right on the facing page 9.24 reminds us of the analogous roles of stratifying and blocking in sampling and comparing [but recall the comment (**–**) near the middle of page 9.13].
- As shown pictorially at the right, a common theme of EPS and EPA is dividing a group of units (or 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 (or units) is the respondent population, the subgroups are the units (or elements) *not* selected and the sample.



14. Observational Plans – The Confounding Effect

In an *observational* Plan, for a focal variate with q values, we think of the respondent population as being made up of q *sub*populations; each subpopulation is those elements which have a particular value of the focal variate. Diagram (2) at the right shows an instance of q = 2 with the two subpopulations being of the *same* size (4 elements); two short horizontal lines show the two subpopulation average responses $\bar{\mathbf{Y}}_0$ and $\bar{\mathbf{Y}}_1$ [as they also do in diagram (1) at the lower right of page 9.21]. The difference between $\bar{\mathbf{Y}}_1$ and $\bar{\mathbf{Y}}_0$ for the two sub*populations* has two components:



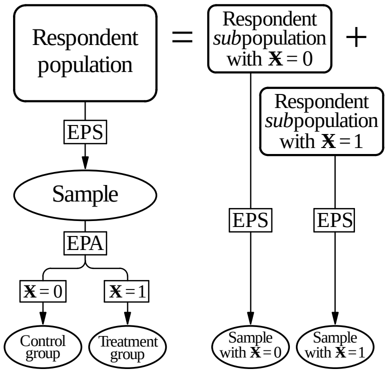
- ∗ the *treatment effect* arising from their different $\mathbf{X}$ values;
- ∗ an effect due to differences between the two subpopulations in the distributions of values (*e.g.*, in the averages) of one or more lurking variates – *we* call this the **confounding effect** and we write equation (9.2.3) below;

$$\bar{\mathbf{Y}}_1 - \bar{\mathbf{Y}}_0 = \text{effect of change in } \mathbf{X} + \text{effect of change in } \mathbf{Z}_1, \ldots\ldots, \mathbf{Z}_k = \text{treatment effect + confounding effect.} \qquad \text{-----(9.2.3)}$$

Explanatory variates are usually numerous and so, for each element, as these variates take their 'natural' values *un*influenced by the investigator(s), there is ample opportunity for different distributions of one or more $\mathbf{Z}_i$ among the q subpopulations of the respondent population. It is usually feasible to manage at most a *few* $\mathbf{Z}$s by matching and/or subdividing.

Assessing Answers from observational Plans must take account of the confounding effect because:
- it is a source of comparison error and the resulting limitation imposed on the Answer(s),
- the treatment effect and the confounding effect cannot be quantified *separately* – we can only know their sum;

thus, our efforts to manage an *inherent* limitation on Answers from observational Plans meet, at best, with only *partial* success. There is further discussion and illustration of the confounding effect in Section 15 overleaf – *e.g.*, in Schema O on page 9.27.


Respondent population
=
Respondent *sub*population with $\mathbf{X}$ = 0
+
Respondent *sub*population with $\mathbf{X}$ = 1
EPS
Sample
EPA
$\mathbf{X}$ = 0
$\mathbf{X}$ = 1
Control group
Treatment group
EPS
EPS
Sample with $\mathbf{X}$ = 0
Sample with $\mathbf{X}$ = 1

NOTE: 34. The schema at the right shows two ways we think about a respondent population in comparative investigating.
- On the left, we think of all elements (or units) having the focal variate value $\mathbf{X} = 0$ and, in an *experimental* Plan, a sample selected by EPS is divided in half by EPA, with one half retaining the value $\mathbf{X} = 0$ and the other being assigned $\mathbf{X} = 1$; the two (half) samples are then compared appropriately to answer the Question(s).
  - An illustration is a clinical trial of a drug – $\mathbf{X} = 0$ represents taking *no* drug (usually taking a placebo in practice) and $\mathbf{X} = 1$

(*continued overleaf*)

NOTES: (cont.) 34. ● – represents taking the drug.

[When *two* drugs are compared, *none* of the population elements (or units) may initially have $\mathbf{X} = 0$ or $\mathbf{X} = 1$, but this does not affect the point of this discussion.]

- On the right of the schema at the lower right overleaf on page 9.25, the 'natural' values of $\mathbf{X}$ define (two) subpopulations and, in an *observational* Plan, the samples to be compared are obtained by EPS from these subpopulations.

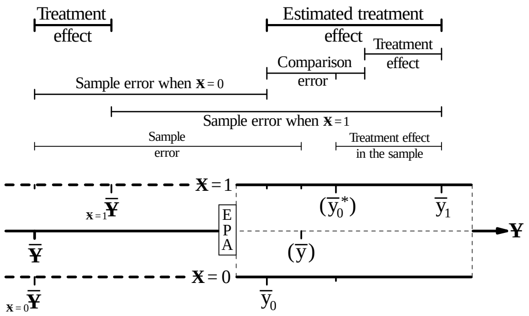
Dividing the sample in *half* by EPA is for simplicity in this discussion; in practice, the control and treatment groups may be made of *different* sizes to manage other sources of error; this can be accomplished by using *un*equal probabilities of assigning elements (or units) to the groups.

- It is also assumed for simplicity that the respondent population size is an exact multiple of the number of groups (*e.g.*, that $\mathbf{N}$ is *even* when there are *two* groups).

15. Comparison Error in Experimental and Observational Plans

Despite the *probabilistic* equivalence of EPS followed by EPA on the left and EPS of two samples on the right in the schema in Note 34 overleaf at the lower right of page 9.25, comparison error is involved in *different* ways in the two Plan types, as illustrated in the two schemas E at the right below and O and at the centre right of the facing page 9.27.

Schema E for an Experimental Plan


Treatment effect
Estimated treatment effect
Treatment effect
Comparison error
Sample error when $\mathbf{X}=0$
Sample error when $\mathbf{X}=1$
Sample error
Treatment effect in the sample
$\mathbf{X}=1$
${}_{\mathbf{X}=1}\overline{\mathbf{Y}}$
E P A
$(\bar{y}_0^*)$
$\bar{y}_1$
$\mathbf{Y}$
$\overline{\mathbf{Y}}$
$(\bar{y})$
$\mathbf{X}=0$
${}_{\mathbf{X}=0}\overline{\mathbf{Y}}$
$\bar{y}_0$

- In schema E representing an *experimental* Plan, the respondent population has (unknown) average $\overline{\mathbf{Y}}$ and the sample selected from it by EPS has (unobserved) average $\bar{y}$.
  - The *difference* in the values of $\overline{\mathbf{Y}}$ and $\bar{y}$ is *sample error*; its value remains *un*known in a particular investigation.
  - We also denote the respondent population average, when all elements have $\mathbf{X} = 0$, by ${}_{\mathbf{X}=0}\overline{\mathbf{Y}}$ and, when all elements have $\mathbf{X} = 1$, by ${}_{\mathbf{X}=1}\overline{\mathbf{Y}}$; the difference of these (unknown) averages is the (unknown) *treatment effect* – the change in the *average* of $\mathbf{Y}$ for *unit* change in $\mathbf{X}$ when all the $\mathbf{Z}$s remain fixed.
    + A treatment effect is an attribute describing a relationship.
- The sample is divided (roughly) in half by EPA.
  - One half yields an average $\bar{y}_0$ for the response variate when the focal variate takes assigned value $\mathbf{X} = 0$;
  - The other half yields an average $\bar{y}_1$ for the response variate when the focal variate takes assigned value $\mathbf{X} = 1$.
  - The (observed) *difference* $\bar{y}_1 - \bar{y}_0$ is the *estimated* treatment effect.
- The estimated and *true* treatment effects differ by comparison error arising from two sources.
  - The two half samples obtained under EPA would likely have *different* averages $\bar{y}_0$ and $\bar{y}_0^*$ when $\mathbf{X} = 0$, due to differences in their distributions for one or more lurking variates $\mathbf{Z}_i$.
  - The treatment effect in the (half) *sample* with $\mathbf{X} = 1$ is likely to differ from the *true* treatment effect;

    Solely to illustrate this discussion, the components of comparison error from the two sources are separated by the short vertical line on the lower side of the comparison error bar to the right of its centre.
    + The hypothetical (*un*observed) average $\bar{y}_0^*$ of the half sample with $\mathbf{X} = 1$, *if it were* to have been assigned $\mathbf{X} = 0$, is called a **counterfactual** and arises again in Note 35 at the bottom of the facing page 9.27 and page 9.28.
  - Comparison error (from both sources) is *eliminated* in the (unattainable) ideal of our three criteria (on the upper half of page 9.8) defining causation, which require a *census* of the respondent population *both* when $\mathbf{X} = 0$ and when $\mathbf{X} = 1$.
- By equating the relevant horizontal distances in schema E, we see that:

  sample error when $\mathbf{X}=0$ + comparison error + treatment effect = treatment effect + sample error when $\mathbf{X}=1$;

  ∴ comparison error = sample error when $\mathbf{X}=1$ – sample error when $\mathbf{X}=0$. -----(9.2.4)

  Because comparison error can be expressed as the difference of two *sample* errors, an experimental Plan which uses EPS and EPA provides the basis for statistical theory which yields:
  - an (inverse) relationship between comparing imprecision and the *group sizes* (or degree of *replicating*);
  - an expression for a *confidence interval* (CI) for the treatment effect (*i.e.*, for a respondent population average) – such an interval, under suitable modelling assumptions, *quantifies* comparing and measuring imprecision (as demonstrated *for EPS* in Appendix 3 on pages 6.7 to 6.11 in Figure 6.1 – see also Figure 8.11 of these Course Materials);
  - *unbiased* estimating (*i.e.*, *zero* comparing inaccuracy) of a treatment effect (a respondent population attribute commonly of interest in a comparative Plan) – recall the discussion of Table 9.2.10 on page 9.22.

  Hence, EPS and EPA in combination provide for quantifying comparing imprecision and so, *in conjunction with adequate replicating* (or *adequate group sizes*), allow an Answer to be obtained with acceptable limitation imposed by comparison error in the context of a particular investigation with a comparative Plan.
  - Experimental Plans which use EPA but cannot feasibly implement EPS (a common state of affairs in practice) have no basis for invoking the three benefits of statistical theory for EPA *and* EPS as these benefits are stated above. However, they *can* be retained in a restricted way if we think of the sample as a 'respondent *population*' which is then (under

(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 11)

EPA) divided into two ‘samples’ (the control and treatment groups).

+ The theoretical benefits are retained for the two (or more) groups (‘samples’) generated *probabilistically.* BUT:
+ Sample error of the original sample is now ‘study’ error with respect to the respondent population – its assessment would be based on *extra*-statistical knowledge and seldom *quantitative* like the provisions of sampling theory. HOWEVER:
+ The severity of the limitation imposed by this ‘study’ error may be alleviated because a *difference* is being estimated.

○ In schema O (at the right below) representing an *observational* Plan, the two respondent subpopulations with focal variate values $\mathbf{X} = 0$ and $\mathbf{X} = 1$ have respective (unknown) averages $\overline{\mathbf{Y}}_0$ and $\overline{\mathbf{Y}}_1$.

– The (unknown) respondent population average $\overline{\mathbf{Y}}$ is the weighted average of $\overline{\mathbf{Y}}_0$ and $\overline{\mathbf{Y}}_1$, the weights being determined by the sizes of the two subpopulations – schema O is drawn with *equal* weights.

– $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ is the *treatment effect* [due to the different values of $\mathbf{X}$ in the two subpopulations] plus a *confounding effect* [due to differences in the (average) values for one or more lurking variate(s) $\mathbf{Z}_i$ in these subpopulations].

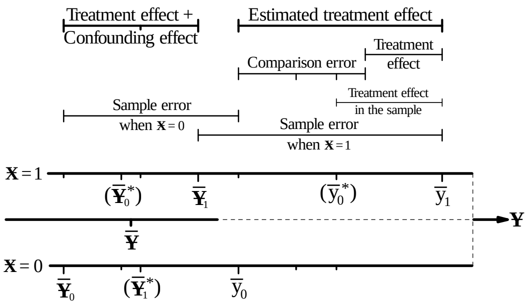
+ From Section 14 on page 9.25, the **confounding effect** in an *observational* Plan is the (unknown) difference between $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ and the treatment effect;

  i.e., $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ = treatment effect + confounding effect. -----(9.2.3)

+ The two components of $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ in the respondent population can*not* be separated in an observational Plan but, solely to illustrate the present discussion, one *possible* separation is indicated in schema O by the short vertical line on the lower side of the effect bar (at the upper left of schema O) a little to the right of its centre.

  The position of this separator involves the hypothetical respondent population averges $\overline{\mathbf{Y}}_0^*$ and $\overline{\mathbf{Y}}_1^*$, representing the hypothetical situation where all elements of each subpopulation have the *other* value of the focal variate.

Schema O for an Observational Plan



○ The samples obtained by EPS from the two respondent subpopulations with $\mathbf{X} = 0$ and $\mathbf{X} = 1$ yield averages of $\overline{y}_0$ and $\overline{y}_1$.

– As in an experimental Plan, the (observed) *difference* $\overline{y}_1 - \overline{y}_0$ is the *estimated* treatment effect.

– The two (unknown and likely different) sample errors when $\mathbf{X} = 0$ and $\mathbf{X} = 1$, the differences between the relevant respondent population and sample averages, are as shown in schema O.

○ The estimated and *true* treatment effects differ by comparison error arising from the confounding effect and two other sources.

– Due to differences in their distributions for one or more non-focal explanatory variates $\mathbf{Z}_i$, the two samples obtained by EPS likely have *different* averages in the hypothetical situation where the units in both had the *same* $\mathbf{X}$ value; for example, schema O shows a difference between $\overline{y}_0$ and $\overline{y}_0^*$ (the average for the sample with $\mathbf{X} = 1$, *if it were instead* to have $\mathbf{X} = 0$).

+ The hypothetical difference $\overline{y}_0^* - \overline{y}_0$ involves *both* the confounding effect *and* the effect of sampling and so differs from (in schema O, is larger than) $\overline{\mathbf{Y}}_0^* - \overline{\mathbf{Y}}_0$.

– The treatment effect in the *sample* with $\mathbf{X} = 1$ is likely to differ from the *true* treatment effect;

Again solely to illustrate this discussion, the components of comparison error from the confounding effect and the two sources are separated by the short vertical lines on the lower side of the comparison error bar.

○ By equating the relevant horizontal distances in schema O, we see that:

sample error when $\mathbf{X} = 0$ + comparison error + treatment effect = treatment effect + confounding effect + sample error when $\mathbf{X} = 1$;

∴ comparison error = confounding effect + sample error when $\mathbf{X} = 1$ – sample error when $\mathbf{X} = 0$. -----(9.2.5)

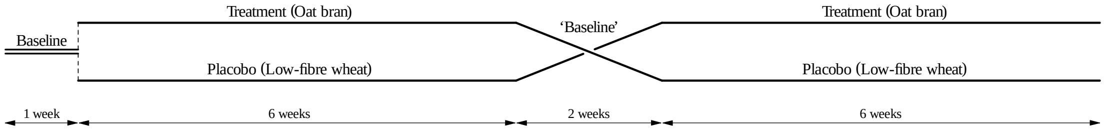
Comparing equations (9.2.5) above and (9.2.4) [on the facing page 9.26], we see why an Answer about a treatment effect from an *observational* Plan has more severe limitation imposed by comparison error than such an Answer from an *experimental* Plan – equation (9.2.5) has the additional confounding effect term arising from the respondent population.

– This additional term is *un*affected by the level of replicating – it persists in a *census* of both respondent subpopulations.

– Limitations on Answer(s) from observational Plans are discussed again in Appendix 4 on pages 9.37 and 9.38.

For clarity, schemas E and O are drawn with *positive* sample error, comparison error and treatment effect; in practice, there may be (some) *cancellation* within or between such entities when they have *opposite* signs.

NOTES: 35. The (half) sample average $\overline{y}_0^*$ in schemas E (on the facing page 9.26) and O (above) is commonly *un*observed but an exception occurs in a blocked experimental Plan called a **cross-over** design, represented pictorially below; an



(*continued overleaf*)

NOTES: (cont.) 35. example is a clinical trial of dietary oat bran as a way of reducing blood (serum) cholesterol levels (and, hence, heart disease).

- Twenty-four participants were divided under EPA into two groups of 12; serum cholesterol levels were monitored for all 24 participants for a baseline period of one week while they ate their normal diets.
- For the next six weeks, cholesterol levels were monitored while one group of 12 participants was assigned a dietary supplement of low-fibre wheat (the placebo), the other group was assigned oat bran (the treatment).
- This was followed for all participants by a two-week break during which no dietary supplement was consumed.
- In the final six weeks of the investigation, the two groups of 12 were assigned the *other* dietary supplement from the one they had consumed in the previous six-week period.

○ The final average serum cholesterol level of the group of 12 participants on placebo for the *second* six-week period can be regarded as $\bar{y}_0^*$, but this Plan really just yields values for $\bar{y}_0$ and $\bar{y}_1$ for *all* 24 participants.

○ The non-focal explanatory variates Ƶ$_{\text{i}}$ (the blocking factors) made the *same* when Ӿ = 0 and Ӿ = 1 are those personal characteristics (*e.g.*, genetic factors, level of exercise) that affect an individual's serum cholesterol level.

– The decreased comparing imprecision afforded by the blocking in this Plan must be set against the limitation imposed on the Answer by the possibility that *order* of being on treatment or placebo affects a participant's serum cholesterol level; *i.e.*, *no* time carry-over effect is *assumed* for being on treatment or placebo.

– Four participants in the investigation were lost due to missing data – the final sample size was 20; this small sample size (*i.e.*, this low level of replicating) means sample error imposes a severe limitation on the Answer.

○ As is common with comparative Plans, the sample was *not* obtained by *probability* selecting – the participants were *volunteers* from among dieticians and other employees of a hospital in Boston.

– The Plan included *double blinding* – see Table 9.2.5 in Note 18 on the lower half of page 9.15.

REFERENCE: Swain, J.F., Rouse, I.L. Rouse, Curley, C.B. and F.M. Sacks, Comparison of the Effects of Oat Bran and Low-Fibre Wheat on Serum Lipoprotein Levels and Blood Pressure. *New Engl. J. Med.* **322**(#3): 147-152 (1990). [DC Library call number: PER R11.B7]

36. The discussion of this Section 15 starting on page 9.26 makes it clear why the Plan for an investigation to answer a Question with a causative aspect will, in general, be experimental by choice, observational *only* by necessity; similarly, a comparative Plan will be blocked/matched by choice, *un*blocked/*un*matched *only* by necessity. The *importance* of observational Plans [or existing (and, hence, cheaper) data from them] is that:

- they are the only choice when it may be infeasible or is unethical for investigator(s) to *assign* values of the focal variate(s) – for instance, level of exercise, type of diet (when compliance is often equivocal) or cigarette smoking.
- they may suggest ('clue generation') how to improve a process *prior* to using an experimental Plan to confirm ('validate') that the sought-after improvement *does* occur when the relevant change is made.

An *experimental* Plan *must*, of course, be used when the relevant value of the focal variate would *not* occur naturally – for instance, an experimental Plan is needed to confirm that a change (like installating a *new* filtration system) *does* achieve the anticipated improvement in a process (like purifying drinking water more effectively).

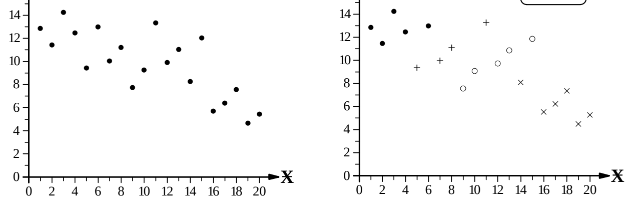
16. Appendix 1: Lurking Variates – Scatter Diagrams

To answer a Question about an Ӿ-¥ relationship between *quantitative* variates, it is 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 Ӿ and ¥ values of each bivariate observation. However, when examining such diagrams, it is easy to overlook the limitation on an Answer about the Ӿ-¥ relationship imposed by different points on the scatter diagram having *differing* values of a lurking variate Ƶ. This matter is illustrated by the two versions of the *same* scatter diagram at the right below:

∗ in the left-hand version in which Ƶ values are *ignored*, we see an Ӿ-¥ 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 Ƶ, the straight-line Ӿ-¥ relationship can have a slope which is (close to) zero (when Ƶ is 0), positive (when Ƶ is 1 or 2) or negative (when Ƶ is 3).

NOTES: 37. When looking at a scatter diagram of bivariate data to assess an Ӿ-¥ 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



(*continued*)

Figure 9.2. INVESTIGATING STATISTICAL RELATIONSHIPS (continued 12)

NOTES: (cont.) 37. on-going exposure in the media, and continuing up to post-secondary level courses, including calculus and algebra.

38. 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.*

39. The foregoing discussion and scatter diagrams in this Appendix 1 draw attention to the distinction between *conditioning* on $\mathbf{Z}$ and *ignoring* $\mathbf{Z}$ when investigating relationships.

∗ Conditioning is *subdividing*, as discussed at the upper left of page 9.59 in Section 2 in Figure 9.8.

∗ A **marginal** (probability) distribution, referred to in Section 6 on page 9.60 in Figure 9.8 and illustrated in Tables 9.8.10 to 9.8.13, is an example of 'ignoring' the variate which is absent from the marginal distribution – for instance, in Table 9.8.11, $\mathbf{X}_2$ is absent, in Table 9.8.12, $\mathbf{X}_1$ is absent, and in Table 9.8.13, $\mathbf{Y}$ is absent.

The scatter diagram at the bottom right of the facing page 9.28 shows the marginal distribution of $\mathbf{X}$ and $\mathbf{Y}$ if we think of the $\mathbf{Z}$ direction as coming vertically up from the page. With the $\mathbf{Z}$ values as given at the upper right of the right-hand version of the diagram and thinking of the page as the plane $\mathbf{Z} = 0$, 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 $\mathbf{X}$ and $\mathbf{Y}$ if we look vertically *down* on the diagram (*i.e.*, we look along the $\mathbf{Z}$ 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, bearing in mind a maxim of the late Dr. George E. P. Box, a respected U.S. statistician: *All model are wrong, some are useful.*

17. Appendix 2: Scatter Diagrams – Pearson's Parent-Child Data (reconstructed)

18. Appendix 3: Equiprobable Selecting

Stats introduction
Note 10, page 5.23
Note 14, page 5.24
Note 98, page 5.86
Unbiased estimating

18. Appendix 4: Limitations on Answers from Observational Plans

Appendix 15 of Figure 5.7, pages 5.82-5.84