Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources

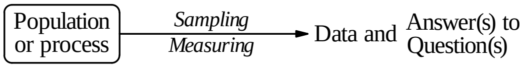
1. Error in Data-based Investigating

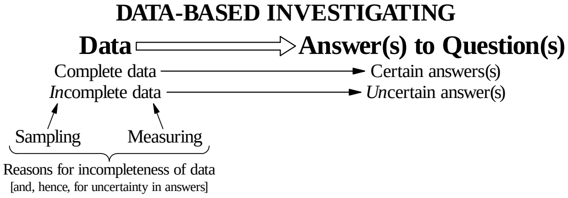
As summarized in the two schemas at the right, statistics is concerned with *data-based investigating* of some population or process to *answer* one or more *questions* of interest.

∗ If the investigating yields *complete* information, we can obtain a *certain* answer; that is, an answer we can *know* is correct.

∗ If the investigating yields *in*complete information, we can*not* know an answer is correct (an *un*certain answer) – in fact, we can be *fairly sure* a *numerical* answer is at least a little off;

– sampling and measuring yield data (and, hence, information) that are *inherently* incomplete.





The goal of statistical methods is to provide, for each Question of interest (in the usual situation of *in*complete information):

∗ an Answer, **AND:** ∗ an assessment of the *limitations* imposed by *error* on that (uncertain) Answer.

Overall error is the difference between the Answer provided by data-based investigating and the (unknown) answer that reflects the *actual* (or '*true*') state of affairs in the population or process. Two illustrations are:

∗ From a national poll which asked: *Do you think that the government should have the right to verify the information given out by welfare recipients or do you think that such a verification represents an invasion of privacy?*, Gallup Canada reported in January, 1993, that 71% of respondents supported verification, 24% saw it as an invasion of privacy and 4% said they did not know; the poll involved 1,011 telephone interviews with Canadian adults from December 19 to 23, 1992.

– Because the sample percentages are unlikely to be (exactly) equal to those in the respondent population, Gallup Canada described the limitation on their Answer by a confidence interval: *A national telephone sample of this size is accurate within a 3.1 percentage point margin of error, 19 in 20 times* (although *we* would refer to precision rather than accuracy).

∗ For the Question: *Is cigarette smoking a cause of lung cancer?*, the Answer is either *Yes* or *No* and error is giving the *wrong* one of these two categories. For a Question like this with a *categorical* Answer, likely error cannot easily be quantified conceptually as it can be for a numerical attribute like a proportion or an average.

– Assessing the limitation imposed by error on a categorical Answer is usually based on judging how well the investigation(s) which gave the Answer dealt with each of the six categories of error discussed starting in Section 2 below.

[Program 11 of *Against All Odds: Inside Statistics* states that the 1964 U.S. Surgeon General's Report considered information from over 6,000 investigations (predominantly with *observational* Plans) to answer Questions about the health consequences (including lung cancer) of cigarette smoking; these investigations would often have answered *their* Questions by comparing *numerical* attributes – for instance, the proportions of lung cancer cases among smokers and non-smokers.]

NOTE: 1. *Error* and *mistake* are often used synonymously in ordinary English but our (technical) meaning is distinct – our 'error' does *not* involve mistakes but is an *inherent* characteristic of an Answer based on incomplete information.

● *Mistakes* in data-based investigating impose additional (often unrecognized) limitations on Answers.

2. Six Categories of Error in Data-based Investigating

To pursue our discussion, we recognize that overall error in data-based investigating is usually the net result of several components, which we think of in terms of six categories of error (see also the diagram at the bottom right of page 5.25).

∗ study error; ∗ measurement error; ∗ model error;

∗ sample error; ∗ non-response error; ∗ comparison error.

These categories are useful because, contingent on proper Question formulation in terms of the target population/process, they help us identify *sources* of error; in a particular investigation, we then incorporate Plan components which we expect will manage inaccuracy and manage imprecision (by managing variation) in ways that will reduce, to a level acceptable in the Question context, the limitations imposed on Answer(s) by (the likely size or chance of) each category of error.

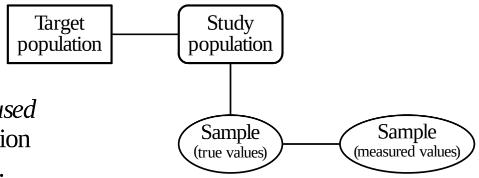
3. Study Error, Sample Error and Measurement Error

The schema at the right below (see also page 6 of Chapter 3 of the 2004 STAT 231 Course Notes) reminds us that, from an introductory perspective, data-based investigating is concerned with three groups of units:

∗ the **target population**: the group of units to which the investigator(s) want Answer(s) to the Question(s) to apply;

∗ the **study population**: a group of units *available* to an investigation;

∗ the **sample**: the group of units selected from the study population *actually used* in an investigation – the sample is a *sub*set of the study population [but see the comment under 'Measuring' overleaf on page 5.20].



(*continued overleaf*)

Associated with these three groups of units are:

∗ **attributes**: quantities defined as a function of response (and, perhaps, explanatory) variates over the group.

Familiar (simple) attributes are averages, proportions, medians and standard deviations. The importance of attributes is that:

- Answer(s) to Questions(s) are usually given in terms of attributes, often their values;
- five of the six categories of *error* are defined in terms of attributes.

The first three categories of error have the following definitions:

∗ **Study error:** the difference between [the (true) values of] the study population/process attribute and the target population/-process attribute. [The population-process distinction is discussed in Appendix 1 on page 5.55.]

∗ **Sample error:** the difference between [the (true) values of] the sample attribute and the study population/process attribute.

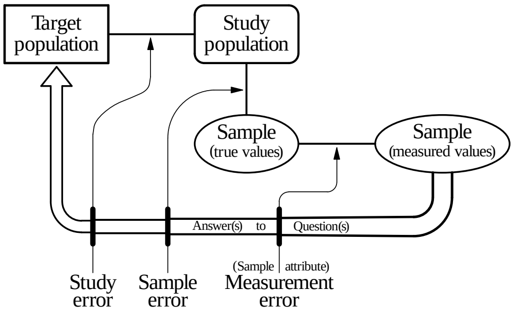
∗ **Measurement error:** the difference between a measured value and the true (or long-term average) value of a variate.

– **Attribute measurement error:** the difference between a measured value and the true (or long-term average) value of a [population/process or sample] attribute.

NOTES: 2. Study error and sample error are defined in terms of *attributes* of groups of units whereas measurement error involves *individual* measurements – this is why the additional (sub)category of *attribute* measurement error is needed.

3. We need *both* true values and long-term average values in the last two of these error definitions because:

- 'true' values for quantities like length, mass and time (and the many quantities derived from them) can be invoked because there are **standards** (*i.e.*, certified *known* values) for such quantities;
  - – Measuring a standard to quantify measuring inaccuracy is called **calibrating** the measuring process.
  - – W.J. Youden's classic discussion of measuring inaccuracy is summarized in Figure 6.4 of the Course Materials.
- long-term average values may be all we have available when, for instance, investigating the distribution of responses to a questionnaire with particular question wording and/or question order.



The impact of these three categories of error can conveniently be displayed as a development of the schema at the bottom right overleaf (page 5.19).

- The three error category names are given across the bottom of the schema at the right, although 'measurement error' is really 'sample attribute measurement error.'
- The arrow rising from each error category name shows the place of impact in the schema of that category of error.
- The broad arrow from the sample ellipse of measured values back to the target population represents Answers(s) to the Question(s) about the target population that are *inferred* from measured sample data on response (and, usually, explanatory) variates.
  - – The thick lines crossing this broad arrow at the arrows rising from each error category represent *limitations* imposed by error on Answer(s); the progressive *de*crease in width of the broad arrow after each error category reinforces this idea.

Other definitions needed now for this Figure are:

∗ **Uncertainty:** ignorance (incomplete knowledge) of error; for example:

– for a numerical Answer, ignorance of the magnitude and/or the sign/direction of error;

– for a categorical Answer (like *Yes* or *No*), ignorance of whether the Answer is the correct category.

∗ **Repetition:** repeating over and over (usually *hypothetically*) one or more of the processes of selecting, measuring, estimating.

∗ **Selecting:** the process by which the units/blocks of the sample are obtained from the study population – it is described in the **protocol for selecting units** (see Appendix 2 on pages 5.56 and 5.57). [*Equiprobable* selecting is abbreviated EPS.]

∗ **Sampling** is *two* processes – *selecting* and *estimating*.

∗ **Measuring:** the process used to determine the value of a variate; the components of a measuring process are the instrument (or gauge), the operator, the method (or instructions) and (sometimes) the unit being measured.

True and measured variate values are distinguished in the two schemas at the lower right overleaf (page 5.19) and above by having *two* sample ellipses; the vertical line in each schema is to remind us the sample is a subset of the study population. [More correctly, the sample is a subset of the *respondent* population, as we will see starting at the bottom of page 5.24.]

∗ **Inaccuracy:** *average* error (*i.e.*, its *systematic* component) under *repetition*.

– **Sampling inaccuracy:** average sample error under repetition of selecting and estimating.

– **Measuring inaccuracy:** average measurement error under repetition of measuring the *same* quantity.

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 1)

∗ **Imprecision:** *standard deviation* of error (*i.e.*, its *haphazard* component exhibited as *variation*) under *repetition*.
- **Sampling imprecision:** standard deviation of sample error under repetition of selecting and estimating.
- **Measuring imprecision:** standard deviation of measurement error under repetition of measuring the *same* quantity.

∗ **Accuracy:** the inverse of *in*accuracy. ∗ **Bias:** the *model* quantity representing *inaccuracy*.

∗ **Precision:** the inverse of *im*precision. ∗ **Variability:** the *model* quantity representing *imprecision*.

∗ **Variation:** differences in (variate or attribute) values across the individuals in a group (○) or arising under repetition (⊙); *e.g.*,
○ a target population/process, ○ a study population/process, ○ a sample, ○ repeated measurements;
⊙ error, ⊙ a sample average.

Variation can be *quantified* by (data or probabilistic) standard deviation. (See also Table 5.7.5 at the upper right of page 5.28.)

∗ **Estimating:** a process using statistical theory to derive the distribution of an *estimator* and data to evaluate an (interval) *estimate*.

∗ **Estimator:** a *random variable* whose distribution shows the possible *values* of the corresponding *estimate* under repetition of the selecting, measuring and estimating processes. [Random variables are introduced in Figure 5.6 of these STAT 231 Course Materials (*e.g.*, on pages 5.15 and 5.17) – see also Figure 5.3 on pages 5.5 to 5.9 of the Course Materials.]

∗ **Estimate:** *numerical value(s)* for a model parameter:
○ derived from the distribution of the corresponding *estimator*, AND: ○ calculated from *data*.
- **Point estimate:** a *single* value for an estimate.
- **Interval estimate:** an *interval* of values for an estimate, usually in a form that quantifies variability (representing imprecision).

NOTES: 4. The language of error is developed in relation to *particular* investigations but, in introductory statistics, we *quantify* error only in relation to behaviour under *repetition* because, for example, the quantifying is based on a *response model* which describes behaviour only under repetition. These individual investigation/repetition and real world/-model distinctions were introduced at the end of Figure 1.1 in Table 1.1.1; this table, with additional terms now defined, is given at the right as Table 5.7.1.

Table 5.7.1:

PARTICULAR INVESTIGATION	**(HYPOTHETICAL) REPETITION**	
Real World	**Real World**	**Model**
error	inaccuracy	bias
variation	imprecision	variability
uncertainty		probability
estimate		estimator

- The terms in the middle column of Table 5.7.1 refer to behaviour under repetition so we use such words in the context of *processes* like selecting, measuring and estimating.
  - ○ A statement like *the sampling imprecision in this investigation* ..... is a contradiction in terms.
- The terms in the left-hand column refer to an individual investigation; used in the context of repetition, they need a qualifier like *more likely to* or *increases the risk of*.
- As we will see later in the course, two main methods of estimating – confidence intervals and tests of significance – and the concept of (un)biased estimating are interpreted formally in terms of *repetition*.
- A disadvantage of English vocabulary is that words with greater inherent appeal – like *accuracy* and *precision* – are the *inverses* of the quantities – like ***in**accuracy* and ***im**precision* – we quantify directly in statistics.

5. Continuing the theme of Section 2 on page 5.19, the importance in statistics of the concept of error and its categorization is because:
- error leads to recognizing the concepts of uncertainty, inaccuracy and imprecision and to their succinct definitions, as given for uncertainty and inaccuracy on the facing page 5.20 and for imprecision at the top of this page 5.21;
  - we then see why statistical methods aim to *manage inaccuracy and imprecision (by managing variation)*.

Also, our meaning of 'error' enables us to define the statistically troublesome adjective 'representative'.
- **Representative sample:** a sample whose sample error [and corresponding limitation imposed on Answer(s)] is *acceptable* in the Question context. However, there are compelling reasons to avoid in statistics the terms 'representative' and 'representativeness' in the context of a sample:
  - a sample selected by EPS is unlikely to be 'representative' in the sense just given for *all* attributes of potential interest – for instance, a sample may have small [possibly (close to) zero] sample error for estimating the (respondent) population average $\overline{\cancel{\mathbf{Y}}}$ but large sample error for estimating its standard deviation $\cancel{\mathbf{S}}$;
  - the sample, of itself, provides no information about its 'representativeness';
  - there is no selecting process *known* to yield a 'representative' sample, except taking a census;
  - the terminology tends to obscure the distinction between the individual case (the *particular* sample) and behaviour under repetition (the properties of the selecting *process*);
  - *representative* has been used with a variety of (sometimes ill-defined) meanings in statistical contexts; in contrast to our 17-word definition above, Kruskal and Mosteller devote 50 pages to discussing the meanings of **representative sampling** in three articles in the *International Statistical Review*, **47**, 13-24, 111-127, 245-265 (1979). [UW Library call number HA 11.I505]

(*continued overleaf*)

NOTES: (cont.) 6. *Variation* generally has *negative* connotations in statistics – it is an *impediment* to:

- ● *estimating* an attribute (like an average) which is an Answer to a Question;
- ● *quality* in manufacturing or service processes – *improving* such processes means *reducing* variation.

By contrast, in the real world, variation can have *positive* connotations; it is:

- ○ viewed as an antidote to boredom – *variety is the spice of life*;
- ○ a requirement for the process of natural selection to operate.

In a similar vein, two contrary views of *incompleteness* are:

- ● incompleteness is the source of uncertainty whose management is the primary concern of statistical methods;
- ○ Answers with limitations acceptable in the Question context *seldom* require the commitment of resources needed to obtain *complete* information; *i.e.*, some uncertainty from conserving resources is usually an acceptable trade-off.

7. English prose in some contexts is considered 'better' if it uses synonyms instead of repeating the same word; with our statistical terminology, the *opposite* is true – it merely sows confusion to use what we think is a synonym in place of the word or phrase we have defined (see also Note 97 on page 5.85 in Appendix 17).

In particular, *we* characterize Answers separately in terms of their (likely) inaccuracy and their (likely) imprecision, together with the ensuing limitations; we ***avoid*** the words **untrustworthy**, **invalid**, **unreliable** and **weak** because:

- ● they too sow confusion if used as synonyms for inaccurate and/or imprecise;
- ● they are sometimes used carelessly to imply some undefined *combination* of inaccuracy and imprecision.

4. Plan Components to Manage Study Error, Sample Error and Measurement Error

This Section 4 can be omitted from a first reading of this Figure 5.7 (as can Sections 6, 8 and 16 on pages 5.26, 5.28 and 5.38). Plan components to manage study error, sample error and measurement error are summarized in Table 5.7.2 below.

Table 5.7.2

Plan Component			Error category	Error Management Strategy
Specifying the study population/process			Study	Specify the study population/process so its attribute(s) can be anticipated to be adequately close (in value) to those of the target population/process. ● *Restricting* values of explanatory variates can reduce variation in the study population/process – this may *de*crease sample error but *in*crease study error. ○ Sample error is preferred because statistical methods to manage it are better defined than the extra-statistical knowledge usually needed to manage study error.
Selecting units	Method of selecting	EPS	Sample	**EPS** is the basis of sampling theory which provides for: ● unbiased estimating of the respondent population average by the sample average; ● quantifying the likely size of sample error under repetition [*i.e.*, quantifying sampling imprecision, which we take here as 'quantifying uncertainty'].
		Judgement		**Judgement selecting** aims to make sample error as small as needed in the context of the *particular* investigation. ● It provides no basis for assessing if this aim has been achieved.
	Sample size (Replicating)	EPS		**EPS**: Sampling imprecision *de*creases with *in*creasing sample size (see Appendix 4 on page 5.59).
		Judgement		**Judgement selecting:** increasing sample size usually decreases the difficulty of making sample error as small as needed in the Question context. BUT: ● There is no theoretical basis which relates sample size to sampling imprecision.
	Stratifying the respondent population			Decreases sampling imprecision under EPS from the (properly-chosen) strata. ● Provides attribute estimates for the strata as well as for the respondent population.
Measuring variates	Inaccuracy		Measurement	Use a measuring process whose inaccuracy is acceptable in the Question context. ● Inccuracy of a measuring process does *not* necessarily decrease with its *cost*. ○ Inaccuracy is managed by using standards (where they exist – see Note 3 on page 5.20) to *calibrate* the measuring process.
	Imprecision			Use a measuring process whose imprecision is acceptable in the Question context. ● Decreased imprecision for a measuring process usually entails a more *costly* process but the converse is *not* necessarily true.
Estimating attribute values	Simple		Sample	Under EPS, the sample average and sample standard deviation provide estimates, with defined behaviour under repetition, of the corresponding respondent population attributes.
	Ratio			When estimating the respondent population average or total under EPS, ratio and regression estimating improve the (simple) estimate by using the respondent population average or total of an *explanatory* variate with a (strong) positive association with the response variate whose attribute is of interest. ● Ratio estimating decreases sampling imprecision when the standard deviation of the response variate increases linearly with the square root of the explanatory variate.
	Regression			● Regression estimating decreases sampling imprecision when the standard deviation of the response variate does not change with the value of the explanatory variate. Ratio and regression estimating introduce *estimating bias* but can have smaller rms error than $\overline{Y}$ or $\mathbf{N}\overline{Y}$ as the estimator of the respondent population average or total.

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 2)

NOTES: 8. Anticipating the broader view of error in Section 5 which starts overleaf near the bottom of page 5.24, 'study population' has been given (more correctly) as 'respondent population' in six places in Table 5.7.2 on the facing page 5.22 and in Notes 10 and 11 below and in Note 14 overleaf on page 5.24.

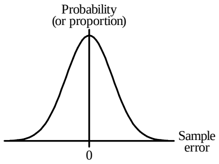
9. Illustrations of *restricting* (measured) explanatory variates when specifying the study population/process are:
   - ● When investigating a manufacturing process for components or parts, the study population might be specified as those parts still at the manufacturing site, which would usually be parts produced consecutively over a relatively short time; their variation is therefore likely to be *smaller* than the longer-term process variation.
   - ● A clinical trial (see Note 38 on page 5.39) of a drug or surgical procedure may restrict the study population of potential participants to one sex or a particular age group; this is again likely to *reduce* variation among participants.

   Managing study error may be assisted by constructing a **fishbone diagram** (see Appendix 3 on pages 5.57 to 5.59) for the *study* population/process and comparing it with the diagram for the *target* population/process, particularly when the units of the study population/process are a *subset* of those of the target population/process.

10. Statistics emphasizes EPS (or its equivalent), particulary in introductory courses, because it is the basis of statistical theory which provides (under repetition):
    - ● an (inverse) relationship between sampling imprecision and *sample size* (or degree of *replicating*);
    - ● an expression for a *confidence interval* (CI) for a population average – such an interval, under suitable modelling assumptions, *quantifies* sampling and measuring imprecision (see Figures 5.8 and 13.1 of these Course Materials);
    - ● *unbiased* estimating (*i.e.*, *zero* sampling inaccuracy) of a population *average* (an attribute commonly of interest).

    [Of course, an Answer obtained from a *particular* sample remains *uncertain*, as reflected by its limitations.]

    - ● Also, a result from probability theory (the Central Limit Theorem) makes approximately *Gaussian* (as illustrated at the right) the distribution of the averages of the set of all possible samples of a given size from the respondent population; as a consequence, *under EPS* there is a *higher* probability of selecting a sample with sample error of *smaller* magnitude, a *lower* probability of selecting one with *larger* sample error.
      - – The *centre* (or 'average') of the (symmetrical) Gaussian distribution in the diagram being at *zero* sample error is what is meant above by *unbiased* estimating – this matter is illustrated in more detail in Appendix 4 on page 5.59.



    EPS does *not*, of itself, *reduce* sample error or sampling imprecision, as implied in (wrong) statements such as:
    - ○ EPS generates a *representative* sample;
    - ○ EPS generates a sample which provides a proper basis for *generalization*; !

    as well as misrepresenting the statistical benefits from using EPS, such statements confuse repetition (the *process* of EPS) with a particular investigation (*a* sample). [Statements like these may arise from mistakenly interpreting language from statistical theory as referring to the sample obtained in an *individual* investigation when it actually refers to behaviour of the selecting *process* under *repetition*.] A *correct* statement is:

    EPS provides for quantifying sampling imprecision and so, *in conjunction with adequate replicating* (or an *adequate sample size*), allows an Answer to be obtained with acceptable limitation imposed by sample error in the context of a particular investigation.
    - ● What constitutes *acceptable* limitation imposed by sample error depends on the investigation requirements for Answer(s); such requirements are often quantified in terms of sampling imprecision.
      - – An example is a proportion – like the percentage of working Canadians who do not contribute to their RRSP – to be estimated to within 2 percentage points with 95% probability or at a 95% level of confidence.
        - + In this example, an Answer is to be obtained that is 'correct' (under repetition) about 95% of the time (*i.e.*, the CI *does* contain the population proportion) and 'wrong' about 5% of the time (the CI does *not* contain this proportion); such uncertainty, quantified (under repetition) in terms of probability or level of confidence, is *un*avoidable for an Answer from incomplete data (*i.e.*, an Answer obtained by *inductive* reasoning).

    Experience shows that EPS is *the* process for selecting the sample to answer a Question with a *descriptive* aspect, and that sample error under *judgement* selecting usually imposes an *un*acceptable limitation on an Answer, to the degree that such investigating is seldom a justifiable use of resources.
    - ● Judgement selecting is included in Table 5.7.2 on page 5.22 because, despite its lack of theoretical foundation, it is commonly used in investigations to answer a Question with a *causative* aspect, where EPS is often infeasible – see the discussion of experimental Plans on page 5.38 in Section 17 and in Appendix 14 on pages 5.79 to 5.82.

11. Sample size [which includes number of *blocks* (or *groups*) in a comparative experimental (or observational) Plan – see pages 5.36 and 5.37] is referred to as *replicating*; definitions of this term and of 'covering' are:

    Replicating: selecting more than one unit/block from the respondent population for the sample.

(*continued overleaf*)

NOTES: (cont.) 11. ∗ – **Adequate replicating:** selecting *just* enough units/blocks from the respondent population to make the likely magnitude of *sample* error [and, hence, the limitation imposed on Answer(s)] *acceptable* in the Question context.

∗ **Covering:** to try to manage sample error, the values of explanatory variates of the units of the sample are selected to cover the range of values that occur among (most of) the units of the respondent (or study) population.

Covering is a guiding principle for judgement selecting; its chance of (partial) success is *in*creased by:

– greater replicating (*i.e.*, a larger sample size), AND:

– greater knowledge about the values of explanatory variates among the units of the respondent population.

12. The result of statistical theory (given overleaf in Note 10 on page 5.23) which inversely relates (under repetition) sampling imprecision to (the square root of) sample size appears to be widely recognized, perhaps in part because it accords with intuition that an Answer from a 'large' sample is likely closer to the state of affairs in the population or process than an Answer from a 'small' sample. However, less widely appreciated (or more easily overlooked) is that the statistical theory is based on *EPS* of the sample. Hence, in proper statistical practice:

● investigator(s) must make clear, and readers (or users) take note of, *how* a sample was selected;

● with a *non*-probability selecting process (*e.g.*, judgement selecting), there is no theoretical basis to justify invoking the sampling imprecision-sample size relationship;

● we should recognize that sampling *inaccuracy* has *no* necessary relationship to sample size – inaccuracy in 'large' samples may thus be more dangerous statistically than in 'small' samples, regardless of the selecting process;

– *lack* of an 'inaccuracy-number of instances' relationship is also characteristic of *other* categories of error – for example, a ruler missing its first centimetre will yield length measurements one centimetre too high (that is, the ruler will have measuring inaccuracy) no matter how many times it is used.

Intuition about the likely smaller magnitude of sample error in an Answer from a 'large' sample *may* be correct in the rare case where the sample contains the majority (at least 80%, say) of the population units – the *statistical* issue is then the sample size *in relation to* the population size, *not* the sample size *per se*.

13. Statistical issues raised by *measuring* in data-based investigating are discussed in Appendix 5 on pages 5.59 to 5.62.

● Managing *sample attribute* measurement error is (usually) achieved by managing measurement error although, as discussed in Appendix 5, there are differences in the effect of measurement error on variates and attributes.

14. Limitation imposed by sample error on Answer(s) based on data from a sample selected by EPS can usually be *reduced* if there is prior information, exploited appropriately, about the respondent population – for example, measured values of relevant explanatory variate(s). Two statistical options are:

∗ **Stratifying:** subdividing the respondent population into groups (called **strata**) so that units with*in* a stratum have *similar* response variate values and units in different strata *differ* as much as feasible; the sample is obtained by selecting units (*e.g.*, by EPS) from *each* stratum. Although the process of stratifying refers to values of the *response* variate, it may be based in practice on values of a suitable *explanatory* variate.

Stratifying is used to achieve one or both of the following benefits:

– to make Answer(s) more useful by subdividing them by stratum; for instance, in Canada, the *national* unemployment rate needs also to be available by province and territory;

– to manage sample error (by decreasing sampling imprecision), provided the prior information about the respondent population allows 'homogeneity within strata, heterogeneity among strata' to be achieved – see Appendix 5 on the last side (page 5.96) of Figure 5.8 for an introductory discussion of stratifying and clustering.

∗ **Ratio or Regression Estimating:** using information about the values of an explanatory variate, over the units of the respondent population, to decrease imprecision of estimating a population attribute like an average or total; to accomplish this, the explanatory variate must have a (strong) positive association with the response variate whose attribute is of interest – the stronger the association, the greater the decrease in imprecision.

Ratio and regression estimating are discussed in later courses on survey sampling (*e.g.*, STAT 332).

● Rms error, an acronym for *root mean square* error, is discussed in Appendix 6 on page 5.63.

5. Non-Response Error

The schema in Section 3 at the bottom right of page 5.19 for our *introductory* discussion is more useful with two additions.

∗ The *respondent* and *non-respondent* populations, to allow for non-response error, discussed in this Section 5;

– definitions on page 5.20 are then restated at the top of page 5.26 with 'study population' replaced by 'respondent population'.

∗ The *model*, to allow for model error, discussed in Section 7 on pages 5.27 and 5.28.

Non-response error is of concern most obviously when study population units are humans and the Question has a descriptive aspect – so-called *sample surveys*. Non-response is an instance of the broader statistical topic of **missing data**. We define:

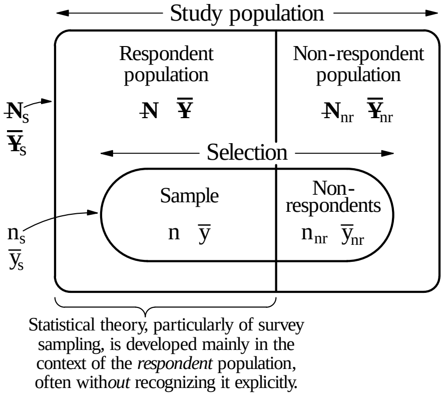
∗ **Respondent population:** those units of the study population that *would* provide the data requested under the incentives for response offered in the investigation;

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 3)

∗ **Non-respondent population:** those units of the study population that would *not* provide the data requested under the incentives for response offered in the investigation.

∗ **Non-reponse error:** the difference between [the (true) values of] the respondent population and study population attributes.


Study population
Respondent population
₦ $\overline{¥}$
Non-respondent population
$₦_{nr}$ $\overline{¥}_{nr}$
$₦_S$
$\overline{¥}_S$
Selection
Sample
n $\overline{y}$
Non-respondents
n_{nr} $\overline{y}_{nr}$
n_S
$\overline{y}_S$
Statistical theory, particularly of survey sampling, is developed mainly in the context of the *respondent* population, often with*out* recognizing it explicitly.

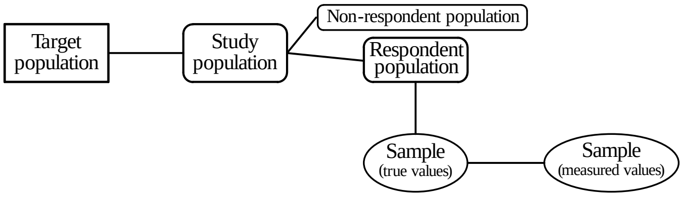
As indicated in the diagram at the right, we consider the *study* population to be made up of the *respondent* and *non-respondent* populations. The set of units selected from the study population is the *selection*, and comprises the *sample* (from the respondent population) and the *non-respondents* (from the *non*-respondent population). The diagram has *two* categories of symbols:

- the ₦s and ns refer to *numbers of units*;
- the $\overline{¥}$s and the $\overline{y}$s are *averages* of a response variate ¥ of the units.

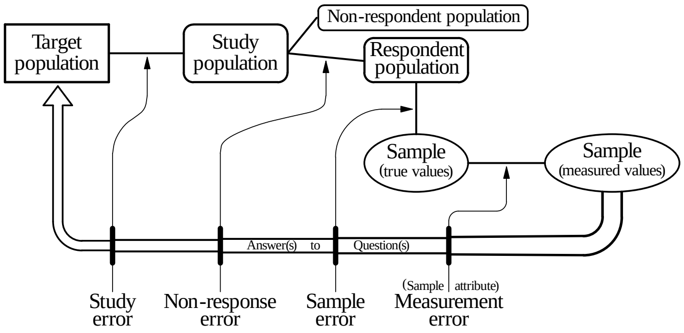
The relationships among the numbers of units are:

$$\text{Study population} = \text{Respondent population} + \text{Non-respondent population}$$
$$₦_S = ₦ + ₦_{nr}$$
$$\text{Selection} = \text{Sample} + \text{Non-respondents}$$
$$n_S = n + n_{nr}$$

The schema at the lower right of the first side of the Figure (page 5.19) is given at the right but with the respondent and non-respondent populations added. The vertical line now indicates the sample as a subset of the *respondent* population.



The schema is shown again, as a development of the one at the centre right of page 5.20, with the impact of *four* categories of error – study, non-response, sample and measurement – included.



For a Question with a **descriptive** aspect, the overall error is the sum of the four error categories:

$$\textit{overall error} = \textit{study error} + \textit{non-response error} + \textit{sample error} + \textit{sample attribute measurement error.} \quad \text{-----(5.7.1)}$$

The diagam at the bottom right shows pictorially, when estimating an *average* to answer a Question with a descriptive aspect, the breakdown of overall error given in equation (5.7.1); symbols are defined in Table 5.7.3 at the right below. Licence on two matters improves the clarity of the diagram:

○ all four error components are *positive* – in practice, overall error may involve some *cancellation* among error components of *opposite* sign;

○ the distribution of *measured* sample attribute values has been moved *down*.

Other matters about the diagram are:

⊙ true (T) and measured (M) values of a sample attribute (here, an *average*), under repetition of the selecting, measuring and estimating processes, have each been modelled by a *Gaussian* distribution;

⊙ the value of the *true* average of the sample *selected*, from among the set of all possible samples, is represented by the black filled circle (•);

⊙ the value of the *measured* sample average, from among the set of all possible such values, is represented by the black filled square (▪);

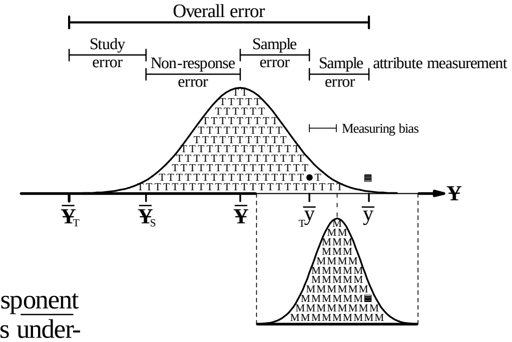
⊙ there is *no* sampling *bias* – the mean of the sampling distribution is $\overline{¥}$;

⊙ there *is* measuring *bias* – the horizontal distance between ${}_T\overline{y}$ and the long-term *average* (the *mean* of the distribution) of its *measured* values;

⊙ sampling variability is larger than measuring variability – the standard deviation of the distribution of the Ts is larger than that of the Ms.

Table 5.7.3: SYMBOL DEFINITIONS

¥	Response variate
$\overline{¥}_T$	(True) target population average
$\overline{¥}_S$	(True) study population average
$\overline{¥}$	(True) respondent population average
${}_T\overline{y}$	True average for sample selected
$\overline{y}$	Measured average for sample selected
T	True value of a sample average
M	Measured value of a sample average



Definitions from page 5.20, restated to take account of the study population-responent population distinction, are given overleaf at the top of page 5.26, with their changes underlined; four *un*changed definitions are also given again so five (correct) error definitions are

(*continued overleaf*)

together for convenient reference.

- ∗ **Study error:** the difference between [the (true) values of] the study population/process attribute and the target population/process.
- ∗ **Non-reponse error:** the difference between [the (true) values of] the respondent population and study population attributes.
- ∗ **Sample error:** the difference between [the (true) values of] the sample attribute and the respondent population attribute.
- ∗ **Measurement error:** the difference between a measured value and the true (or long-term average) value of a variate.
  - **Attribute measurement error:** the difference between a measured value and the true (or long-term average) value of a [population/process or sample] attribute.
- ∗ **Selecting:** the process by which the units/blocks of the sample are obtained from the respondent population – it is described in the **protocol for selecting units** (see Appendix 2 on pages 5.56 and 5.57).
- ∗ **Variation:** differences in (variate or attribute) values across the individuals in a group (○) or arising under repetition (⊙); *e.g.*,

  ○ a target population/process, ○ a study population/process, ○ repeated measurements; ⊙ error,
  ○ a respondent population, ○ a non-respondent population, ○ a sample; ⊙ a sample average.

  Variation can be *quantified* by (data or probabilistic) standard deviation. (See also Table 5.7.5 at the upper right of page 5.28.)

NOTE: 15. Which units of the study population fall in the respondent and non-respondent populations depends on the *incentives* offered for response – different incentives will presumably, in general, result in *different* sets of the study population units in the two populations. For *given* incentives for response (as specified in the protocol for selecting units in a *particular* investigation), statistical theory to manage non-response error can be based on:

- ● a **deterministic** model – a given unit will *always* make the *same* decision about whether or not to respond; OR:
- ● a **stochastic** model – a given unit's decision will involve uncertainty and so is modelled probabilistically.

Our respondent and non-respondent populations are *conceptual* in the sense that we only encounter *subsets* of them (as the sample and the non-respondents); if a unit is *not* included in the selection, we generally do *not* know (and do not *need* to know) to which of the two populations it belongs.

6. Plan Components to Manage Non-Response Error

When using the average, represented by the random variable $\overline{Y}$, of the sample selected by EPS as the estimator of the *study* population average, $\overline{\boldsymbol{\yen}}_{\text{s}}$, the *non-responding bias*, representing non-responding inaccuracy, is:

$$E(\overline{Y}) - \overline{\boldsymbol{\yen}}_{\text{s}} \equiv \overline{\boldsymbol{\yen}} - \overline{\boldsymbol{\yen}}_{\text{s}} = \overline{\boldsymbol{\yen}} - \frac{\text{₦}\cdot\overline{\boldsymbol{\yen}} + \text{₦}_{\text{nr}}\cdot\overline{\boldsymbol{\yen}}_{\text{nr}}}{\text{₦} + \text{₦}_{\text{nr}}} = \frac{\text{₦}_{\text{nr}}}{\text{₦} + \text{₦}_{\text{nr}}}(\overline{\boldsymbol{\yen}} - \overline{\boldsymbol{\yen}}_{\text{nr}}). \qquad \text{-----(5.7.2)}$$

Thus, to reduce non-responding inaccuracy, the Plan for an investigation needs to include components that are expected to reduce one or both of the terms on the right-hand side of equation (5.7.2):

● the non-response rate, AND/OR ● the difference in attribute values for the respondent and non-respondent populations.

These Plan components emphasize incentives for units to provide the information requested, as summarized in Table 5.7.4 below.

Table 5.7.4

Plan Component	**Error category**	**Error Management Strategy**
Obtaining responses — Incentives: Questionnaire Interviewer Call-backs Other	Non-response	Apart from a possible *legislated* requirement to respond (*e.g.*, to a population census), obtaining responses from units which are humans relies on *incentives* which include: ● a clear, answerable, succinct questionnaire; ○ when feasible, a questionnaire on *one* sheet of paper (or equivalent) is an advantage; ● properly trained interviewers; ● call back to units until those who are *un*available *are* contacted; ● appeal to altruism – respond to provide information that will benefit society; ● offer a material reward for response: ○ give *every* respondent a small item like a pen or a dollar coin; ○ offer respondents a chance to win a substantial prize like a trip. The skill and persistence of interviewers, developed by training, are a component of the incentives – see also Note 68 at the bottom of page 5.62. The clean separation of respondents and non-respondents is an idealization – partial (or 'item') non-response is also encountered in practice when sampled units provide some, but *not all*, of the information requested.
Imputing		**Imputing** is the process of assigning values for missing observations – *e.g.*, assigning a value for the reponse of a non-respondent on the basis of its values for known explanatory variates (like sex, age, location) that (it is hoped) are reasonable 'predictors' of the response variate. ● The purpose of imputing is to simplify the data analysis; it *rarely* meaningfully increases the completeness of the information in the data.

NOTE: 16. Illustrations of the requirement for clear and answerable Questions are:

- ● *How much have you spent on gasoline in the last decade?* is clear but unanswerable for many people.
- ● *How much do you spend per week on household items?* is *un*clear because 'household items' are not defined.

A question to quantify behaviour may ask about the behaviour over a time period (*e.g.*, last week or last month) and then ask if that time period was *typical*, rather than asking for the *average* behaviour over such a period.

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 4)

NOTE: (cont.) 16. About 14 minutes into the content of Program 14 of *Against All Odds: Inside Statistics*, entitled *Sampling and Sampling Distributions*, Roger Tourangeau comments that the response process is complicated and people have to:

- interpret the question correctly in the way the researcher wanted it interpreted,
- retrieve from memory whatever information is relevant to answering the question,
- often combine the information into some kind of overall judgement,
- report an answer,

and that things can go awry in each of these steps. The newspaper article *Hooked on your cell? You must be Canadian*, reprinted in Appendix 7 on page 5.64, is relevant to these matters.

7. Statistical Modelling and Model Error

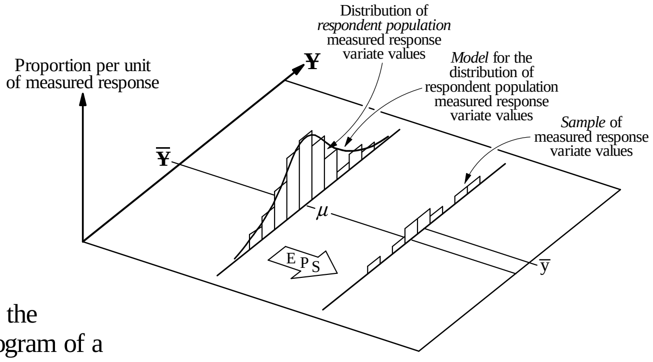
∗ **Response model:** a mathematical description, including modelling **assumptions**, of the relationship between a response variate and explanatory variate(s); the form of the relationship is contingent, in part, on the Plan.

– The **structural component** models the effect of specific explanatory variate(s) on the response variate.

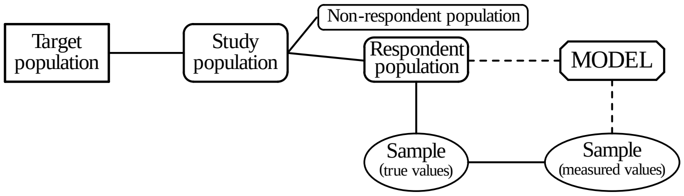
– The **stochastic component** models variation about the structural component.

∗ **Model parameter:** a constant (usually denoted by a *Greek* letter) in a response model that *represents* a respondent population *attribute* – for example, μ represents $\overline{\mathbf{Y}}$ in the response model (5.6.7) near the middle of the last side (page 5.18) of Figure 5.6.

The four main response models discussed in STAT 231 are summarized in Figure 7.1 of these Course Materials; model symbols are defined in Figure 5.9 and an overview of least square estimating of model parameters is given in Figure 8.1.


Distribution of *respondent population* measured response variate values
Proportion per unit of measured response
$\mathbf{Y}$
Model for the distribution of respondent population measured response variate values
Sample of measured response variate values
$\overline{\mathbf{Y}}$
μ
E P S
$\overline{y}$

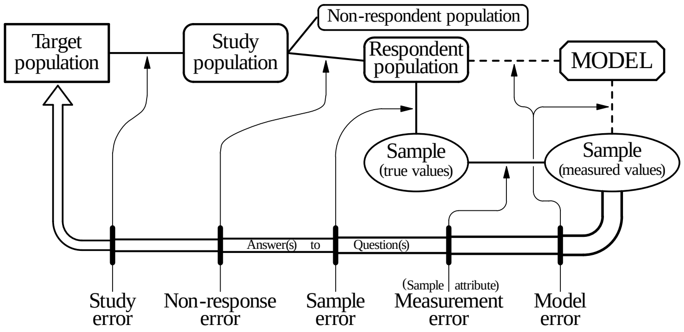
Model-based methods of analysis in statistics use data from a sample to *estimate* values of model parameters which then represent plausible values (in light of the data) for respondent population attributes and, hence, for Answer(s) to Question(s); we distinguish a *point* estimate from an *interval* estimate (as defined above Note 4 on page 5.21). When the Gaussian model is appropriate for the distribution of the response variate values, the model mean μ is estimated by the sample average $\overline{y}$ and σ is estimated by the sample standard deviation s – both *point* estimates. As illustrated at the right, we can think of the process of estimating μ by $\overline{y}$ and σ by s as approximating the histogram of a data set by the Gaussian p.d.f. with the same 'centre' and same 'width' as the histogram.



The schema from the centre right of page 5.25 can be adapted to include the model as shown at the right; this version now incorporates *two* additions over its introductory version at the bottom right of the first side (page 5.19) of the Figure – the respondent/non-respondent populations and the model.

∗ **Model error:** the difference between the model and its modelling assumptions and the actual state of affairs in the real world; modelling assumptions in introductory courses typically include:

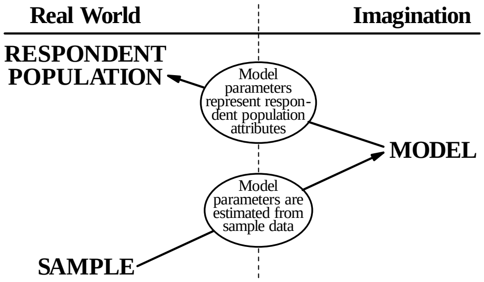
○ equiprobable selecting of units for the sample;
○ the form of the structural component of the response model;
○ the Gaussianicity of each residual;
○ probabilistic independence of the residuals;
○ equal standard deviations of (response) variate values among different groups of units.



Model error, with its mathematical and probabilistic focus, is a broad and complex topic and differs in its nature from the other five categories of error involving attributes; the discussion of Plan components to manage model error in Table 5.7.6 overleaf on page 5.28 is restricted to five assumptions (as given above) underlying our (four) response models (summarized in Figure 7.1). The lower schema at the right above is shown again at the right, as a development of the one at the centre right of page 5.25, with the impact of *five* categories of error – study, non-response, sample, measurement and model – included.

(*continued overleaf*)

All mathematical models are idealizations and all are products of the intellect and the imagination. As shown in the lower two schemas overleaf on page 5.27, we think of the model as a *link* between the *sample* and the *respondent population*; a more detailed pictorial representation of this idea is shown at the right.


Real World
Imagination
RESPONDENT POPULATION
Model parameters represent respondent population attributes
MODEL
Model parameters are estimated from sample data
SAMPLE

NOTES: 17. To maintain the distinction between the real world (represented by the data) and the model, we use different words – 'average' and 'mean' – for their measures of location; unfortunately, we do not have this option for the two measures of variation, which are both called 'standard deviation'. In the early stages of learning statistics, it is helpful to, at least mentally, add the respective adjectives 'data' and 'probabilistic' to distinguish the two uses of standard deviation. This terminology is summarized in Table 5.7.5 at the right.

Table 5.7.5

Attribute	Real World	Model
Location	Average	Mean
Variation	(Data) standard deviation	(Probabilistic) standard deviation

18. In Figure 5.8 of these Course Materials, we develop probability models for the investigative processes of sampling (selecting) and measuring; these models allow us, in Chapter 13, to quantify the likely size of sample and measurement error – that is, to quantify uncertainty from these two sources – for Answers to some types of Questions.

19. In Note 42 on page 5.43, there is further discussion of model error arising from differences between the mathematical form of the structural component of the model and the actual state of affairs in the real world – the context is using a response model to manage comparison error (due to possible confounders) in an observational Plan.
    - In *any* situation where an Answer is based, in whole or in part, on a mathematical model, we should bear in mind a maxim of the late Dr. George E. P. Box, a respected U. S. statistician: *All model are* f2wrong, some are useful.

8. Plan Components to Manage Model Error

Plan components to manage model error are summarized in Table 5.7.6 below.

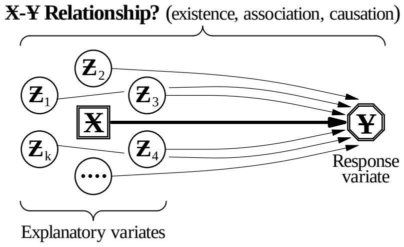
Table 5.7.6

Plan Component		Error category	Error Management Strategy
Assessing modelling assumptions	EPS Form of the structural component Gaussianicity Probabilistic independence Equal standard deviations	Model	Limitations imposed by model error from two modelling assumptions are managed by: ● ensuring the selecting process for units *is* (equivalent to) *EPS*; ● ensuring variate values are measured *independently*. Assessing how well modelling assumptions appear to be met usually involves graphical displays (*e.g.*, scatter diagrams) of the estimated residuals from the response model; ● use a Gaussian quantile plot (or, sometimes, a histogram) to assess Gaussianicity; ○ transforming (*e.g.*, taking logarithms of) the data can help meet this assumption; ● use a plot in the time order of data collecting to assess probabilistic independence. ● use side-by-side dot- or boxplots, or a plot with the explanatory variate from the structural component of the model on the horizontal axis, to assess equality among, or dependence on an explanatory variate of, standard deviation(s).

NOTE: 20. Measuring **independently** means the operator's knowledge of the value arising from one realization of the measuring process does *not influence* the value (s)he obtains from any other realization (see also page 5.60 in Appendix 5).

9. 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 units 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 (5.7.3) in which



$$¥_j = \beta_0 + \beta_1 x_j + \beta_2 z_{1j} + ... + \beta_{k+1} z_{kj} + R_j, \quad j = 1, 2, ..., n, \quad R_j \sim G(0, \sigma), \text{ indep., EPS} \qquad \text{-----(5.7.3)}$$

¥ 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 (5.7.3) requires a similar caveat, of course.]

(*continued*)

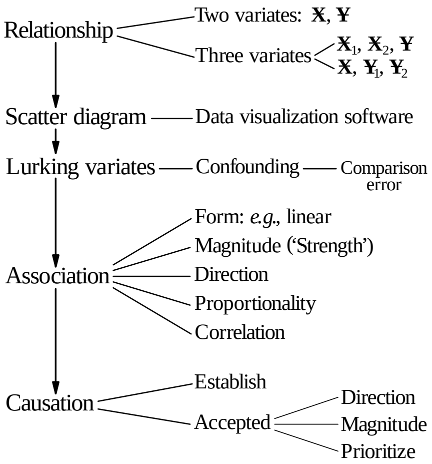
Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 5)

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

NOTES: 21. 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 near the bottom of the facing page 5.28 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 56 on the upper half of page 5.52).
- 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')].

22. The notation used in this Figure 5.7 is Ӿ for the *focal* variate and Ƶ for other *non*-focal explanatory variates; elsewhere, you may see the meanings of Ӿ and Ƶ interchanged.

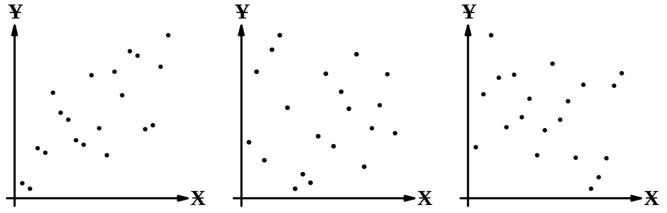


∗ A **relationship** in statistics arises from the following sequence of happenings.

- We observe that the value of a *response* variate ¥ *changes* (*i.e.*, shows *variation*) over the 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 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. [Taking account of lurking variates when interpreting a scatter diagram is discussed on pages 5.31 and 5.65 in Appendix 8.]

10. Comparison Error – Lurking Variates and Confounding

As background to an Ӿ-¥ relationship, Ƶ$_1$, Ƶ$_2$, ..., Ƶ$_i$, ..., Ƶ$_k$ in the schema at the lower right of the facing page 5.28 are called **lurking variates**, a phrase that means lurking *explanatory* variates in that each Ƶ accounts, at least in part, for changes from unit to unit in the value of the response variate. The importance of lurking variates is that if the distributions of their values *differ* between groups of units [like sub)populations or samples] with different values of the focal variate, an Answer about the Ӿ-¥ relationship may differ from the true state of affairs unless the differences in the values of the relevant Ƶs are taken into account.

A practical difficulty for data-based investigating of an Ӿ-¥ relationship is that lurking variates are often *numerous* and so:

- ● it is easy to overlook important Ƶs 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 Ƶs deemed to be important.

Variates other than Ӿ and ¥ 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 Ƶ$_i$ *is* an explanatory variate, AND:
- + comparing boxplots of z_i values for the different values of x to try to identify differences in Ƶ$_i$ for differemt Ӿ values.

(*continued overleaf*)

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 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 units with different values of the focal variate, if taken into account, would meaningfully change an Answer about an X̶-¥ relationship.

∗ **Confounding:** differing distributions of values of one or more *non*-focal explanatory variate(s) among two (or more) groups of units [like (sub)populations or samples] with different values of the focal variate.

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

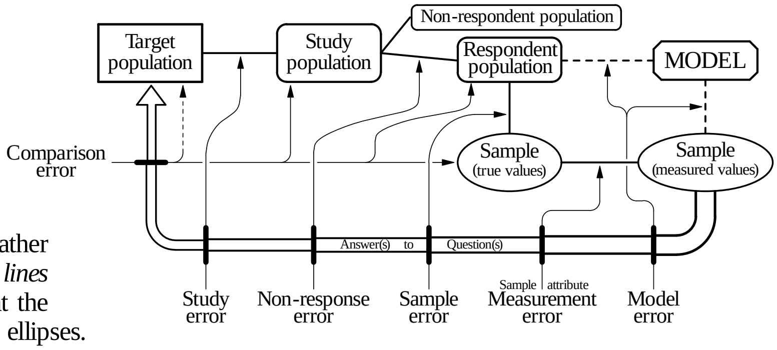
'Confounding' and 'confounder' have the convenience of being one-word terminology rather than the multi-word phrases involving 'lurking variates' which convey the same ideas.

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

The alternate wording of the last phrase accommodates the equivalent terminologies of lurking variates and confounding; in a particular context, we use the version of the definition appropriate to that context:

- 'lurking variates' can more readily accommodate phenomena like Simpson's Paradox – see Appendix 9 on pages 5.65 to 5.70;
- 'confounding' is more common in the context of comparative Plans, as in Section 15 which starts on page 5.36, but the variety of usage of 'confounding' can be a source of difficulty – see Appendix 10 on pages 5.70 to 5.73.

Sections 11 to 20 (pages 5.30 to 5.45) which follow provide necessary background before we continue discussion of comparison error.



The schema introduced at the centre right of page 5.20, and progressively adapted on pages 5.25 and 5.27, has been further adapted as shown at the right to include *comparison* error; the schema now has all *six* error categories introduced on page 5.19 in Section 2.

○ In the schema, the four arrows arising from comparison error point to *boxes* representing *groups* of units (a population or a sample) rather than, as for the other five error categories, to *lines joining boxes*; the comparison error arrow at the right is to be taken as pointing to *both* sample ellipses.

– *Multiple* comparison error arrows are a consequence of its different manifestations in different Question contexts, as summarized in Table 5.7.41 at the bottom of page 5.75 in Appendix 11.

Plan components to manage comparison error are summarized in Table 5.7.10 near the middle of page 5.38.

11. Association – Statistical Issues

The description of a relationship in statistics overleaf on page 5.29 refers to the *association* of ¥ and X̶; this Section 11 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 13 on pages 5.34 and 5.35.

∗ **Association:** if a scatter diagram shows a clustering of its points about, say, a line with positive slope (*i.e.*, we see that, as X̶ increases, ¥ also tends to increase), we say X̶ and ¥ show a (positive) *association*; there is *moderate* positive association of X̶ and ¥ in the left-hand scatter diagram at the centre right overleaf on page 5.29. The right-hand diagram shows *weak 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 X̶-¥ 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 4.14 and 4.15 of Figure 4.5.

– **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 overleaf on page 5.29.
+ If the points of a scatter diagram lie *on* a straight line with positive slope, r = +1;

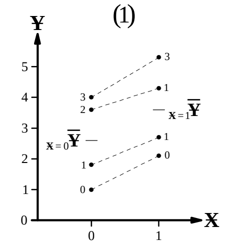
(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 6)

+ 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 4.5 of the Course Materials.

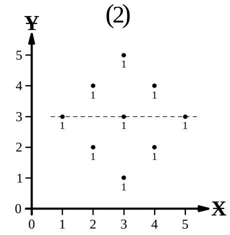
The discussion at the beginning of Section 10 near the bottom of page 5.29 refers to a group of units with a lurking variate (**Z**) whose distribution of values differs, over the units of the group, for different values of the focal variate **X**. A consequence of this behaviour of **Z** is that the values of **X** and **Z** are *associated*, as illustrated in the following scatter diagrams, for respondent populations with 4 or 9 units and **Z** values (shown beside the points) like 0, 1, 2 and 3. [*Distinct* **Z** values for *all* population units, as in diagrams (1) and (5), is rare in real populations.]

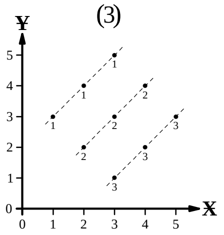


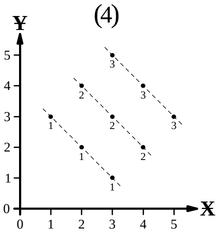
○ In diagram (1) at the right, the unit with **Z** = 2 when **X** = 0 has **Z** = 1 when **X** = 1; thus, the change in the average of **Y** (indicated by a short horizontal line) from 2.6 to 3.6, as **X** changes from 0 to 1, no longer reflects *only* the effect of changing **X**; a limitation is therefore 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 5.32], 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.







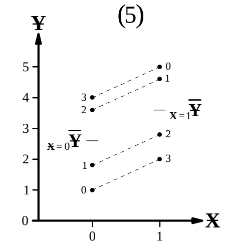
○ In diagram (2), if **Z** has the *same* value (say 1) for all nine units whose **X** and **Y** values yield this scatter diagram, there is *no* **X**-**Y** relationship in the sense that the **X**-**Y** correlation is zero.

+ This *lack* of **X**-**Y** relationship is also reflected by the slope of *zero* for the straight line (shown dashed) which summarizes the trend in the points of the scatter diagram.
+ When interpreting a scatter diagram like (2), it is easy to confuse *explicit* knowledge that there is the same **Z** value among the units, with *assuming* this to be the case by *ignoring* the units' **Z** value(s) – see also Appendix 8 on page 5.65.
+ In diagrams (6) to (8) overleaf on page 5.32, reminiscent of an *experimental* Plan with *two* values of the focal variate, we can accommodate *different* values of the potential confounder **Z** among the units; by contrast, in diagram (2) above, reminiscent of an *observational* Plan, the units 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 17 and 18 on pages 5.38 to 5.40.]

○ 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 8 on page 5.65.
+ 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 units' **Z** values do not remain fixed (are not the same) as **X** changes, and this behaviour is *not* taken into account (*e.g.*, when interpreting an **X**-**Y** scatter diagram).



A special case is when **Z** changes with **X** but in such a way that their values have *zero* correlation; an illustration is shown at the right in diagram (5), which is adapted from diagram (6) overleaf on page 5.32. 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 20 and Notes 45 to 48 on pages 5.44 and 5.45.

NOTE: 23. The discussion above shows that, when looking at a scatter diagram of bivariate data to assess an **X**-**Y** relationship, experience *out*side statistics with diagrams involving Cartesian axes provides poor preparation for statistics – in calculus and algebra courses, for example, the issue of another variate affecting the interpretation of what we see in the diagram seldom (or never) arises.

(*continued overleaf*)

12. Causation – Statistical Issues

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

(1) **LURKING VARIATES:** Ensure *all other* explanatory variates $\mathbf{Z}_1$, $\mathbf{Z}_2$, ..., $\mathbf{Z}_k$ hold their (same) values for *every* population unit when $\mathbf{X} = 0$ and $\mathbf{X} = 1$ (sometimes phrased as: *Hold all the* $\mathbf{Z}_i$ ***fixed*** *for* .....).

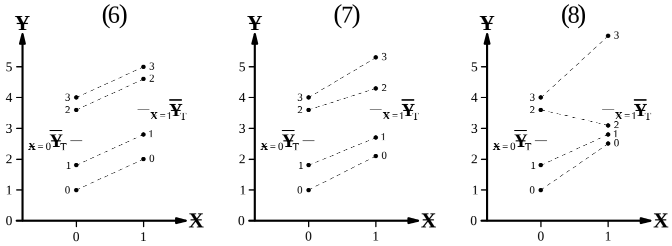
(2) **FOCAL VARIATE:** Observe the population $\mathbf{Y}$-values, and calculate an appropriate attribute value, under *two* conditions: ⊙ with *every* unit having $\mathbf{X} = 0$; ⊙ with *every* unit having $\mathbf{X} = 1$.

(3) **ATTRIBUTE:** $\text{Attribute}(\mathbf{Y}, \text{ perhaps some of } \mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k | \mathbf{X} = 0) \neq \text{Attribute}(\mathbf{Y}, \text{ perhaps some of } \mathbf{Z}_1, \mathbf{Z}_2, ..., \mathbf{Z}_k | \mathbf{X} = 1)$; 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 $\mathbf{X}$ in a particular context; actual values of the focal variate are set in the **protocol for setting levels**, discussed in Section 20 on pages 5.43 to 5.45.

Three illustrations, involving only *one* lurking variate $\mathbf{Z}$, of this formal definition are given at the right below for a target population of 4 units with respective $\mathbf{Z}$ values (shown beside the points) of 0, 1, 2 and 3.

- ○ In diagram (6), $\mathbf{Y}$ values increase by 1 as $\mathbf{X}$ changes from 0 to 1 and, correspondingly, the *average* of $\mathbf{Y}$ (indicated by a short horizontal line) increases by 1 from 2.6 to 3.6.
- ○ In diagram (7), the $\mathbf{Y}$ values again increase as $\mathbf{X}$ changes from 0 to 1 but by *differing* amounts.
- ○ In diagram (8), three $\mathbf{Y}$ values *in*crease but one *de*creases as $\mathbf{X}$ changes, although the *average* of $\mathbf{Y}$ again increases by 1 from 2.6 to 3.6.



In contrast to the five diagrams overleaf on page 5.31 where there *is* confounding, diagrams (6) to (8) illustrating our definition of causation have (of course) *no* confounding – the values of $\mathbf{Z}$ do *not* change as $\mathbf{X}$ changes, so there is *no* $\mathbf{X}$-$\mathbf{Z}$ association (*zero* $\mathbf{X}$-$\mathbf{Z}$ correlation). Also, the $\overline{\mathbf{Y}}$s have a subscript T denoting 'target population'.

NOTES: 24. 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 unit*, 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* units of the target population under each of *two* values of the focal variate is attained more closely in practice in an *experimental* Plan – the two samples to which the investigator(s) assign equiprobably the two values of the focal variate stand in for the respondent population (and, hence, at two stages removed, for the target population) under the two values.
  - In an *observational* Plan, the two values of the focal variate define *sub*populations of the respondent (and the study) population and the two samples with the two values of the focal variate stand in only for these *sub*populations; this matter is pursued in Section 22 and Note 54 on page 5.49.
  - 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 units. A consequence of criterion (3) is that $\mathbf{X}$ need not bring about a change in $\mathbf{Y}$ for *every* unit of the population for us to say $\mathbf{X}$ *causes* $\mathbf{Y}$.
  - ═ A rationalization of this departure from the intuitive idea that causation *always* produces an effect is [like criterion (1)] in terms of non-focal explanatory variates $\mathbf{Z}_i$ – there may be units with (some) such variate(s) whose value(s) have the consequence that a change in $\mathbf{X}$ does *not* bring about a change in $\mathbf{Y}$; we would normally think of these units 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

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 7)

NOTES: (cont.) 24. ● – + 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 10.6 of these Course Materials.

25. The three criteria (at the top of the facing page 5.32) defining causation are framed in terms of the (target) *population* and an appropriate *attribute*, **not** units 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 15 on pages 5.36 to 5.38);
- under probability assigning of units' focal variate values, use statistical theory to manage *under repetition* differences among unblocked, unmeasured and unknown **Z**s (see Section 21 on pages 5.45 to 5.49);
- use a *response model* in the Analysis stage of the PPDAC cycle to hold some **Z**s fixed *mathematically*, but even a quite elaborate model, like equation (5.7.3) on page 5.28 in Section 9, 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 (5.7.3) 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.

26. For 'focal variate' criterion (2), there are focal variates (like age and sex) whose values can*not* be *assigned* to units 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 units, which imposes limitations that arise from using animal units in an experimental Plan or human units in an observational Plan.

These matters are pursued in a discussion of Simpson's Paradox in Appendix 9 on pages 5.65 to 5.70.
- 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 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 55 on pages 5.51 and 5.52), there *is* a time difference between $\mathbf{X} = 0$ and $\mathbf{X} = 1$ for both half samples; any changes in units' *other* explanatory variates values over time may then be a source of comparison error.

27. '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 units' **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 units 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 13 overleaf.

28. An example of the caveat in 'attribute' criterion (3) is: when using least squares estimates [equation (5.7.4) at the right] to *compare* simple linear regression *slopes*, the z values must be the *same* when $\mathbf{X} = 0$ and $\mathbf{X} = 1$.

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} y_j(z_j - \overline{z})}{\sum_{j=1}^{n} (z_j - \overline{z})^2} \qquad \text{-----(5.7.4)}$$

29. Ideas about investigating **X**-**Y** relationships are summarized at the right in Table 5.7.7.

Table 5.7.7: Summary of Ideas About Investigating X-Y Relationships

Criterion (1): the ideal	Ensure all the $\mathbf{Z}_i$ hold their (same) values for every population unit when $\mathbf{X} = 0$ and $\mathbf{X} = 1$
Criterion (3)	For causation, a relevant *attribute* must differ in value when $\mathbf{X} = 0$ and $\mathbf{X} = 1$
Confounding	Confounding arises when one or more of the $\mathbf{Z}_i$ change in value when $\mathbf{X} = 0$ and $\mathbf{X} = 1$
Comparison error	A difference, due to confounding, from the *real* or *intended* value of an *attribute* of a relationship.

- The *difference* in attribute values in criterion (3) must be such as to be *practically important* in the Question context.
- A danger of appropriating 'confounding' as statistical terminology is that a word for *failure* to meet criterion (1) may shift the focus away from this overriding ideal.

(*continued overleaf*)

13. Association Among Variates and Causation

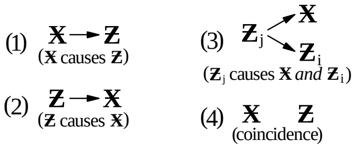
Section 11 on pages 5.30 and 5.31 deals with *association* of two *explanatory* variates, like the *focal* variate **X** and a lurking variate (or confounder) **Z**; we now distinguish four reasons ('cases') for such associations, which are also shown symbolically at the right, where an arrow denotes causation.

∗ **X** causes **Z**;

∗ **Z** causes **X**;

∗ $\mathbf{Z}_j$ causes **X** *and* $\mathbf{Z}_i$ – we say $\mathbf{Z}_j$ is the **common cause** of **X** *and* $\mathbf{Z}_i$;

∗ coincidence [which often means both **X** and **Z** are associated with *time* – *i.e.*, coincidence is often case (3) where $\mathbf{Z}_j$ is time (whatever 'causation' by time means – recall Note 26 overleaf on page 5.33)].



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*;

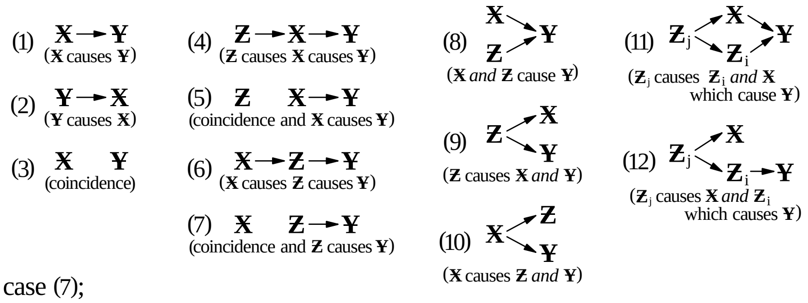
∗ **Z** (or $\mathbf{Z}_i$) and **Y** are associated in the last *nine*.

∗ **Z** (or $\mathbf{Z}_i$) and **X** are associated in the last *nine* [except perhaps in case (8)].

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).

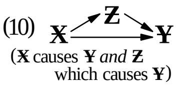
Thus, key statistical issues in association and causation are:

∗ 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 24 on page 5.32 – 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 41 on pages 5.42 and 5.43 and the discussion on page 5.46 in Section 21 to the left of Table 5.7.16].



- Case (12) is both: – case (9) with an intermediary variate shown in the $\mathbf{Z}_j$-**Y** branch,
  – case (11) for the Question *Is* **X** *a cause of* **Y**? when the Answer is *No*.

This leaves cases (1), (9) and (11); we discuss cases (1) and (9) in Section 14 starting on the facing page 5.35 and on page 5.36, and we pursue them and cases (8) and (11) in Section 19 on pages 5.40 to 5.43 – see also Appendix 11 on pages 5.73 to 5.76.

The foregoing discussion shows why, in statistics, we distinguish *association* from *causation*: to remind us that, just because we observe (for instance, in a scatter diagram) that **X** and **Y** are *associated*, we can**not** say, without further investigating, that a change in **X** will *bring about* (or *cause*) a change in **Y**.

- Figure 4.5 discusses *correlation* as a measure of the tightness of clustering of the points of a scatter diagram about a straight line; correlation is therefore one way of quantifying magnitude ('strength') of association between **X** and **Y** as seen in a scatter diagram. For this reason, the distinction between association and causation may also be referred to elsewhere as the distinction between correlation and causation, although this wording is better avoided.
- When referring to an **X**-**Y** relationship, phrases used in statistics like *association is not (necessarily) causation* and *corre-*

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 8)

- *lation is not (necessarily) causation* encompass *three* possibilities:
  - the **X**-**Y** relationship is a *coincidence* – this may pique our curiosity but is seldom of practical importance;
  - **X** and **Y** are *associated* but **X** does not *cause* **Y**;
  - **X** *is* a (or possibly *the*) cause of **Y**.

  Undue emphasis on the second possibility (*e.g.*, in introductory statistics teaching) can obscure three matters:
  + association *does* imply causation if coincidence can be ruled out; BUT:
  + the causation *may* be, but is not *necessarily*, between **Y** and **X**, the variates *observed* to be associated.
  + *Lack* of association of **X** and **Y** does *not* rule out causation of **Y** by **X** – as **X** changes, a confounder **Z** may change in such a way that **Y** remains *un*changed – see diagrams (5) to (8) on the upper half of page 5.75 in Appendix 11.

NOTES: 30. When (a change in) an explanatory variate **U** (a focal variate **X** or a confounder **Z**) *causes* (a change in) a variate **V** (a response variate **Y** or a focal variate **X**), several matters determine the *strength* of the association (as quantified by the correlation, say, of **U** and **V**, if they are *quantitative* variates).

$\mathbf{U} \rightarrow \mathbf{V}$

$\mathbf{X} \rightarrow \mathbf{Y}$

$\mathbf{Z} \rightarrow \mathbf{X}$

- If **U** is the *only* cause of **V** and acts on a time scale that is **short** relative to the period of observation, there is a *high* correlation of **U** and **V**; in the absence of measurement error, the magnitude of r would be 1.
  - An illustration is force **X** causing acceleration **Y**.

Table 5.7.8:

Unit	Smoking status	Lung cancer (A)	(B)	(C)
1	Non-smoker	No	No	No
2	Non-smoker	No	No	No
3	Non-smoker	No	Yes	No
4	Smoker	Yes	Yes	No
5	Smoker	Yes	Yes	Yes
6	Smoker	Yes	Yes	Yes

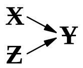
- *Weaker* association of **U** and **V** can occur for several reasons, as illustrated by the data for the occurrence of lung cancer **Y** in relation to smoking status **X** in three non-smokers and three smokers in Table 5.7.8 at the right. The *strong* ('perfect') association in case (A) can weaken 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 **X** does *not* cause **Y**;

  the first two possibilities have a time scale for causation that is **long** relative to the period of observation and the third involves our **definition of causation** (at the top of page 5.32) in terms of an *attribute*.

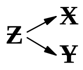
In these ways, we account for differing strengths of *association* observed in *causal* **X**-**Y** relationships or, expressed another way, we account for why (a change in) **X** *causes* (a change in) **Y** but, for *some* population units:
- **Y** changes when **X** does *not* change (*e.g.*, some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

31. 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 5.34] give insight into their difference. As discussed in Note 27 on page 5.33, under our definition of causation at the top of page 5.32:

(8)



(9)



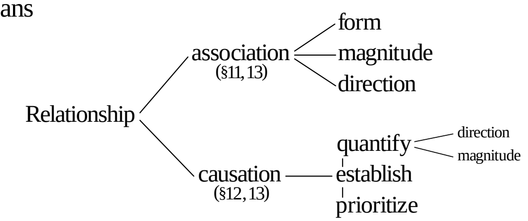
- in the causal structure of case (8) [common *response*], there is *association* of **X** and **Y** and of **Z** and **Y** but *no* necessary association of the (unconnected) causes **X** and **Z**; BUT:
- in the causal structure of case (9) [common *cause*], there is *association* of **Z** and **Y** and of **Z** and **X** so there is *necessarily* association of **Y** and **X**.

The *difference* between the two structures lies in the *direction* of the arrows denoting causation – if their direction is *reversed* in either diagram, they are the *same* causal structure, apart from the variate names. *Our* definition of causation thus suggests that causation is *directed association*, although it is questionable whether this (model) concept provides much insight into the *real world* difference between association and causation.

Cases (8) and (9) and three other similar causal structures are compared in Appendix 12 on pages 5.76 and 5.77.

14. Investigating Statistical Relationships – Three Types of Causal Questions

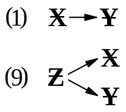
Relationships investigated in statistics, which *we* describe in terms of variates, are often encountered as *associations*; investigating associations includes identifying their characteristics and/or the reasons (causal or otherwise) for them (see also Appendix 11 on pages 5.73 to 5.76). This Section 14 is concerned with comparative Plans for investigating relationships where causation is to be established or *is* involved; the focus on the **X**-**Y** relationship being *causal* means that a *change* can (potentially) be induced in **Y** by *changing* **X**. These matters are summarized in the schema at the right, which reminds us that:



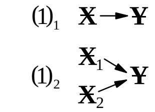
∗ association is usually characterized by its *form*, *magnitude* or *direction*;
  - correlation (see Figure 4.5) is one measure of magnitude ('strength') for a straight-line association; form can also be *non*-linear;

∗ it is useful to distinguish three types of Questions with a causative aspect:
  - **Establishing** whether **X** *is* a cause of **Y**, usually with a view to manipulating **X** to produce

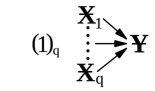
(*continued overleaf*)



- a (desired) change in 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 X and Y is causation of Y by X is answering the Question whether the relevant causal structure (shown again at the right from page 5.34) is case (1) or case (9) [= case (12)].



- **Quantifying** the relationship between X (or, more commonly, X_1, X_2, ..., X_q) and 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.



- **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, Y) from among many causes X_1, X_2, ..., X_q.

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

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

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

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

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

NOTE: 32. In STAT 231, *establishing* causation is discussed in Chapter 10 of the Course Notes but the emphasis is on *quantifying* the relationship between *one* focal variate and a response variate – for example, see Chapters 7, 10 and 15; extension to more than one focal variate is taken up in STAT 332. *Prioritizing* causes is pursued in STAT 435.

15. Terminology for Comparative Plans – The Protocol for Choosing Groups

The three criteria defining what *we* mean by causation, in Section 12 at the top of page 5.32, involve observing a *population* under two values of the focal variate: with *all* the units having $X = 0$ and with *all* the units having $X = 1$. We try to approach this ideal in a *sampling* context by having *two* samples, one with its units having $X = 0$ and the other with its units having $X = 1$; 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) Y corresponding to a change in 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) units 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 ($X = 1$) and withheld from others ($X = 0$) in a clinical trial (an *experimental* Plan – see Note 38 on page 5.39).
- A population in which each unit has one of *two (or more)* values ($X = 0, 1, ...$) of a focal variate of interest, whose value it is *not* feasible to change for any unit – recall Note 26 on page 5.33.
  - 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 21 to 23 on pages 5.45 to 5.52.

For comparative Plans to answer a Question with a causative aspect, the **protocol for choosing groups** specifies whether the units of the sample will be selected so they form groups that can be used to reduce the limitation imposed on an Answer(s) by comparison error – relevant Plan components are shown in the schema at the upper right of the facing page 5.37:

∗ **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:

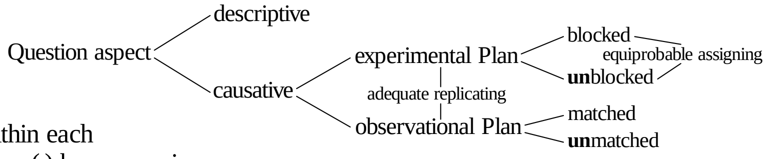
Blocking meets 'lurking variates' criterion (1) for those non-focal explanatory variate(s) Z_i [the **blocking factor(s)**] made

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 9)

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 **Ƶ**s 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 5.48 in Note 53. [This analogy is sometimes interpreted as showing that stratifying in survey sampling is merely an instance of blocking, but this interpretation (unhelpfully) downplays the different contexts and intents of blocking and stratifying.]

∗ **Equiprobable assigning (EPA) [random assigning** or **randomization]:** using a probabilistic mechanism (described in the protocol for choosing groups) in an *experimental* Plan to assign the values of the focal variate with *equal* probability:

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

Equiprobable assigning provides a basis for theory which relates comparing imprecision to level of replicating; thus, EPA, *in conjunction with EPS and adequate replicating*, provides for quantifying comparing imprecision arising from unblocked, unknown and unmeasured non-focal explanatory variates and so allows a particular investigation to set group sizes which are likely to yield an Answer(s) with limitation imposed by comparison error that is acceptable in the Question context.

∗ **Matching** in an *observational* Plan: forming groups of units with the *same* values of one or more non-focal explanatory variates but *different* values of the *focal* variate. THUS:

Matching meets 'lurking variates' criterion (1) [at the top of page 5.32] 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 Table 5.7.12 and its discussion on page 5.39.

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: 33. Where the definitions of blocking (on the facing page 5.36) and matching (above) refer to values of non-focal explanatory variates being the *same*, in practice the values may only be *similar*.

34. The groups of 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* units, they may be referred to as *matched pairs* – see Table 5.7.9 at the right – but a *block* of two units may also be referred to as a 'pair.'

Table 5.7.9
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.

35. In DOE, non-focal explanatory variate(s) made the same within blocks are called **blocking factor(s)**; in data-based investigating to improve industrial processes, typical blocking factors are days, shifts, batches of raw material, machine spindles or filler heads, moulding machines, moulds, or cavities within moulds.

- The values of a blocking factor among blocks should be chosen to make its sample attribute (*e.g.*, its average or distribution) similar to its respondent (or study) population attribute.
- An entity that is the same both within *and* among blocks (like the measuring process) is *not* a blocking factor but is part of what *defines the study population/process* – for example, data for an investigation collected on *one* day and *one* production shift. *If* such factors as day or shift have an appreciable effect on the response, the limitation imposed on the Answer by *study* error is more severe (comparison error is traded for study error).

36. 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

(*continued overleaf*)

NOTES: (cont.) 36. 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 53 on pages 5.48 and 5.49.

37. *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 Note 14 on page 5.24.

Elsewhere, *both* processes may be called 'stratifying'. There is further discussion of subdividing on page 5.39 in Section 18 and of stratifying on page 5.40 near the end of Note 39 – see also the top of page 5.59 in Note 64.

16. Plan Components to Manage Comparison Error

Comparison error in comparative investigating, introduced on page 5.30 in Section 10, arises from confounding by non-focal explanatory variate(s); background information and Plan components to manage comparison error are then discussed in Sections 11 to 15 on pages 5.30 to 5.38. These Plan components are summarized in Table 5.7.10 below.

Table 5.7.10

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 5.7.12 and its discussion on the lower half of the facing page 5.39.

17. Experimental Plans – Sample selecting and Blocking

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

∗ For example, in data-based investigating to improve a manufacturing process (*e.g.*, by identifying and removing causes of excessive variation in the process output), the items manufactured by the process are often shipped away from the manufacturing plant as they are made and investigators are then forced (quickly) to use recent production, or a subset of it, as the sample – a *sample of convenience.* Three factors alleviate this *statistically* unsatisfactory state of affairs:

- With *stable* processes [where the distribution(s) of the output response variate values remain (essentially) the same from one time period to another], a 'snapshot' of the process in time (like recent production) may often have attribute values that are *close* to those of the process in the long-term.
- Answers are derived from *differences* in sample attributes; such Answers may have less severe limitation imposed by sample error than Answers based on sample attribute values which do *not* involve taking a difference.
  + An illustration is the Physicians' Health Study (of the effect of aspirin on heart disease), which used about 22,000 male doctors as the sample – half the doctors took aspirin and half took a placebo. It is likely that the incidence of heart attacks among doctors differs appreciably from that for the target population of all males, but the *difference* in incidence of heart attacks caused by taking aspirin may be much more similar among doctors and all males. (Two newspaper reports of this investigation are reprinted in Figure 10.2 of 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

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 10)

– judgement seems to be better at sample selecting in such situations than it does when answering a Question with a *descriptive* aspect, but it is still far from the statistical ideal.

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 – see Section 9 (Appendix 3) on pages 5.91 to 5.95 of Figure 5.8;
- ∗ more generally, the theoretical basis is gone for interpreting formal methods of data analysis like confidence intervals and tests of significance. [Regrettably, 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 10.2 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 26 on page 5.33].
- ∗ 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: 38. 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 units are *humans*, a technique called **blinding** is used (where feasible) because of its statistical benefits. [To be *blind* means not to know, for any unit, 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 5.7.11 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 5.7.11

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*

18. Observational Plans – Sample selecting, Matching and Subdividing

The comments in Section 17 (on the facing page 5.38 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 5.7.12 at the right for the case of *two* subpopulations. These hypothetical data for two samples (selected from subpopulations of non-smokers and smokers) of 10,000 people involve a response variate $\mathbf{Y}$ which is lung cancer status, a focal variate $\mathbf{X}$ which is smoking status, and $\mathbf{Z}_i$ is whether a unit has a family history of lung cancer, as a possible indicator of genetic predisposition to the disease; for simplicity, $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{Z}_i$ are *binary* variates in this illustration. Each of the six sets of three table entries is the sample size ('Number') and the lung cancer 'Cases' as a number and a percentage of the sample size.

Table 5.7.12	**Non-smokers ($\mathbf{X} = 0$)**			**Smokers ($\mathbf{X} = 1$)**		
	Number	**Cases**	**%**	**Number**	**Cases**	**%**
No family history ($\mathbf{Z}_i = 0$)	9,000	63	0.7	8,900	712	8
Family history ($\mathbf{Z}_i = 1$)	1,000	7	0.7	1,100	88	8
Both	10,000	70	0.7	10,000	800	8

The bottom line of Table 5.7.12 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

(*continued overleaf*)

$\mathbf{X}$-$\mathbf{Y}$ relationship from an observational Plan encounters three potential difficulties.

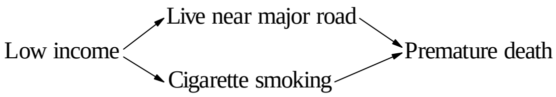
- Investigators have no control over the sample sizes after subdividing; if one or more of the $\mathbf{X}$-$\mathbf{Z}_i$ combinations is rare, the resulting small sample size(s) *in*crease comparing imprecision and so increase(s) the limitation imposed by comparison error on an Answer about an $\mathbf{X}$-$\mathbf{Y}$ 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 5.7.12 raises the possibility (*not* realized here) of the phenomenon known as Simpson's Paradox (and its accompanying limitation imposed on an Answer) – see Appendix 9 on pages 5.65 to 5.70.

NOTE: 39. 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 Appendix 14 on pages 5.79 to 5.82 – recall also the discussion on pages 5.38 and 5.39.
  - 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 units which can make up the study (and respondent) population of controls.
    - + When controls are selected *non*-probabilistically, there is no theoretical basis for an inverse relationship between sampling imprecision and (the square root of) of the sample size – recall Note 10 on page 5.23 – so there is no *statistical* reason why a larger sample size for controls will decrease comparing imprecision (see also Note 57 at the end of Section 24 at the top of page 5.55).
    - + The blocks in a blocked experimental Plan are also often selected *non*-probabilistically but, as discussed in Appendix 14 on pages 5.79 to 5.82, judgement selecting *may* still allow an experimental Plan to have *acceptable* limitation imposed by comparison error on an Answer to a Question with a causative aspect.

19. Comparative Plans and Causal Structures



With the additional background given in Sections 14 to 18 on pages 5.35 to 5.40, the causal structure at the right provides a convenient context for discussing cases (1), (8), (9) and (11) from Section 13 on pages 5.34 and 5.35; this context comes from an inves-

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 11)

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.

---

tigation whose newspaper report is reprinted above. 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)].

Case (11): The causal structure at the lower right of the facing page 5.40 is an instance of case (11) [shown again at the right], although the investigation (described above) of the effect of living near a

$$\begin{tikzcd}
 & & \mathbf{X} \arrow[dr] & \\
(11) & \mathbf{Z}_j \arrow[ur] \arrow[dr] & & \mathbf{Y} \\
 & & \mathbf{Z}_i \arrow[ur] &
\end{tikzcd}$$

(*continued overleaf*)

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): $\mathbf{Z}_j \rightarrow \mathbf{X}$, $\mathbf{Z}_j \rightarrow \mathbf{Z}_i$, $\mathbf{X} \rightarrow \mathbf{Y}$, $\mathbf{Z}_i \rightarrow \mathbf{Y}$.

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 5.7.13 below – 'E' or 'O' in the Plan column denotes 'experimental' or 'observational'.

Table 5.7.13

No.	Case	Question	Plan
1	(1) $\mathbf{X} \rightarrow \mathbf{Y}$	Is low income associated with premature death?	O
2	(8) = (1) $\mathbf{X} \rightarrow \mathbf{Y}$	Is living near a major road associated with premature death?	O
3	(8) = (1) $\mathbf{X} \rightarrow \mathbf{Y}$	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 5.41. 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 units' 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 are 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 near the top of page 5.36 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 5.34 in Section 13; these become twelve structures at the middle right of page 5.34 with the inclusion of the response variate. The discussion on page 5.34 below these structures and in this Section 19 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 19 reminds us that, to develop a comparative Plan to answer a Question with a causative aspect, sufficient *extra*-statistical knowledge is needed to:
  - 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: 40. The *observational* nature of the Plans in Table 5.7.13 above 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 14 on pages 5.35 and 5.36.

(10) $$\begin{tikzcd}[row sep=small] & \mathbf{Z} \\ \mathbf{X} \arrow[ur] \arrow[dr] & \\ & \mathbf{Y} \end{tikzcd}$$

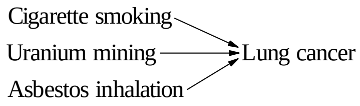
$$\begin{tikzcd}[row sep=small] & \mathbf{Z} \arrow[dr] & \\ \mathbf{X} \arrow[ur] \arrow[rr] & & \mathbf{Y} \end{tikzcd}$$

41. Case (10) [shown again at the right] in its lower *real* form (because $\mathbf{Z}$ is an explanatory variate) is *not* a viable basis for a comparative Plan, which requires either:
    - *direct* causation of $\mathbf{Y}$ by $\mathbf{X}$ [case (1)], OR:

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 12)

NOTES: (cont.) 41. ● an *explicit* intermediate variate in *each* causal chain from **X** to **Y**, as in case (11) and illustrated at the start of Section 19 in the causal structure at the bottom right of page 5.40.



– There may be *more* than two causal chains from **X** to **Y**, as illustrated by a case of *three* possible intermediaries at the right. There may also be *interaction* (see page 5.44) 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).

42. The newspaper article (reprinted on page 5.41) discussed in Section 19 on pages 5.40 to 5.43 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 5.41 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 (5.7.3) on page 5.28 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 5.41 whether we regard the seven explanatory variates as seven focal variates or as one focal variate and six possible confounders.

20. 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.

∗ 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 48 on page 5.45.
Our concern in this Section 20 is with factor(s) that are *focal* variate(s).

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

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

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

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

∗ The **(treatment) effect** of **X** on **Y** (usually) refers to the change in the *average* of **Y** for *unit* change in **X** and:
– implies the **X**-**Y** relationship is (believed to be) *causal* – a change in **X** *causes* (brings about) a change in **Y**;

(*continued overleaf*)

- 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 units 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 Appendices 13 and 14 on pages 5.77 to 5.79 and 5.79 to 5.82.

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: 43. Not all experimental Plans lead to a Data stage in runs.

- Process improvement investigations often *do* – the Data 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 units] for the *whole* period of the Data stage.

When the Data 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*.

44. Equiprobable assigning of treatments to units may not be feasible in an experimental Plan when one factor has hard-to-alter levels. For example, if pouring temperature (at two levels, say, of 1,450°F and 1,600°F) is a factor in an investigation to improve a process for making iron castings, the temperature of the furnace containing the molten iron cannot easily be altered; it may therefore be necessary to do *consecutively* all the runs at each temperature, instead of having the pouring temperature low or high under equiprobable assigning for each run. This *lack* of probability assigning *in*creases the limitation imposed on an Answer by comparison error.

- What is desirable statistically in data-based investigating may also be compromised in process improvement investigations by having to carry out the Data 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.

45. In ordinary English, interaction customarily involves *two* entities; in statistics, *three* (or more) variates are involved – two (or more) focal variates and one response variate.

- *Confounding* also involves two explanatory variates and one response variate; it is compared and contrasted with interaction (and with other causal structures involving three variates) in Appendix 12 on pages 5.76 and 5.77.

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 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.

46. When there are two or more focal variates, 'lurking variates' criterion (1) near the top of page 5.32 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 47 on the facing page 5.45.

- 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

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 13)

NOTES: (cont.) 46. ● *Hi*), have one run with both Ӿ$_1$ and Ӿ$_2$ set 'Lo', another run with Ӿ$_2$ set 'Hi' and another with Ӿ$_2$ back at 'Lo' and Ӿ$_1$ set 'Hi'; the resulting data, shown as three response variate averages in Table 5.7.14 at the right, do *not* allow the Ӿ$_1$-Ӿ$_2$ interaction effect to be estimated, because there is no run with both factors set 'Hi'.

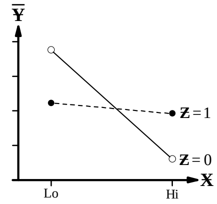
Table 5.7.14	Ӿ$_2$ Lo	Ӿ$_2$ Hi
Ӿ$_1$ Lo	Ɏ$_{\text{Lo,Lo}}$	Ɏ$_{\text{Lo,Hi}}$
Ӿ$_1$ Hi	Ɏ$_{\text{Hi,Lo}}$	No data

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

47. The idea in Note 45 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 5.30 in Section 10), where the association of Ӿ and Ƶ typically has a correlation with magnitude *less* than 1. As discussed in Appendix 10 on pages 5.70 to 5.73, both cases are usually (unwisely) simply called 'confounding' without distinction.

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



- ○ Ɏ is tile size *after* firing in an oven,
- ○ Ӿ 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,
- ○ Ƶ 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 Ƶ = 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 Ƶ = 1 and clay must not be too expensive an ingredient.

21. 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 Ɏ are given in Table 5.7.15 at the right for a respondent population of six units under two values [assigned by the investigator(s)] of a focal variate Ӿ – the treatment effect (the change in the average of Ɏ for unit change in Ӿ when all the Ƶs remain fixed) is 0.3 units, the average of (widely-varying) effects of changing Ӿ for the individual units. The data in Table 5.7.15 are also shown in diagram (1) at the right; the value of a lurking variate Ƶ given beside each dot reminds us that, for our initial discussion, changing Ӿ does *not* affect the value of Ƶ [but see the comment in the second bullet (⦿) in the second paragraph overleaf on page 5.46]. The population averages when Ӿ = 0 and Ӿ = 1 are shown as short horizontal lines; the differing notation used for these averages between diagrams (1) [above] and (2) [at the middle right of page 5.49] is to emphasize the distinction between experimental Plans [where the *investigator(s)* assign each unit's Ӿ value (under EPA)] and observational Plans [where each unit has its 'natural' Ӿ value *un*influenced by the investigator(s)].

Table 5.7.15: Respondent Population Responses (N = 6)

Unit no.	**1**	**2**	**3**	**4**	**5**	**6**	**Av.**
Ӿ = 0	0.9	1.5	1.8	3.6	3.9	4.5	2.7
Ӿ = 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**

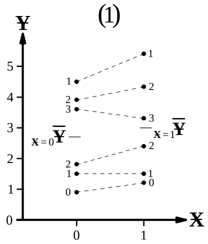


Table 5.7.16 at the upper right overleaf on page 5.46 (which is *un*blocked – see Note 49 overleaf on page 5.46) shows the twenty possible assignments of the six population units whose data are given in Table 5.7.15 above, together with their response variate averages, treatment effect and comparison error; for example, the first line of Table 5.7.16 shows:

- ● units 1, 2 and 3 assigned Ӿ = 0 (often called the 'control group') with average response 1.4,
- ● units 4, 5 and 6 assigned Ӿ = 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 5.7.16 have meaning only if all 20 assignments are *equi*probable (as they are under EPA).

(*continued overleaf*)

[The last column of ten values in *italics* at the right of Table 5.7.16 is discussed in the second bullet (⊙) below.]

Table 5.7.16: Data for the Set of All 20 Equiprobable Assignments of the 6 Units in Table 5.7.15 (on page 5.45)

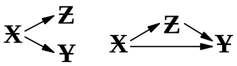
Unit numbers		**Averages**		**Treatment**	**Comparison**	
X = 0	**X** = 1	**X** = 0	**X** = 1	**effect**	**error**	
(1, 2, 3)	(4, 5, 6)	1.4	4.3	2.9	2.6	*2.8*
(1, 2, 4)	(3, 5, 6)	2.0	4.0	2.0	1.7	
(1, 2, 5)	(3, 4, 6)	2.1	3.7	1.6	1.3	*1.5*
(1, 2, 6)	(3, 4, 5)	2.3	3.3	1.0	0.7	*0.9*
(1, 3, 4)	(2, 5, 6)	2.1	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 averages at the bottom of Table 5.7.16 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 5.34 and Note 41 on pages 5.42 and 5.43], this unbiasedness is lost, as the following illustration shows.

(10) 

○ Suppose the change in **Z** (resulting from the change in **X**) causes unit 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 units (with **Z** values *other than* 3) still have the effects given for **X** = 1 in Table 5.7.15. The 10 assignments involving unit 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 5.7.16); the *average* of the *twenty* comparison error values is then 0.1 instead of zero, indicating biased estimating of the treatment effect.

○ If units 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 units with a given **X** value is the relevant *population* average – see the right-hand column of Table 5.7.15 at the start of Section 21 overleaf on page 5.45 – this is unbiased estimating of a respondent population average under EPS (see also Note 10 on page 5.23 and more detail on pages 5.91 and 5.92 in Appendix 3 in Figure 5.8 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: 49. If the responses in Table 5.7.15 at the start of Section 21 overleaf on page 5.45 were *real* data, the Answer about the value of the treatment effect could be made more useful by managing the substantial variation among the units' effects – *e.g.*, by blocking to decrease comparing imprecision [recall the comment (–) near the top of page 5.37].

50. To avoid confounding, 'lurking variates' criterion (1) near the top of page 5.32 requires the ideal of *all* non-focal explanatory variates $\mathbf{Z}_i$ holding their values for *every* population unit 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 units when **X** = 0 and **X** = 1.

– 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);

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 14)

NOTES: (cont.) 50. ● – + 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:
+ use the Analysis stage of the PPDAC cycle to try to 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 5.32 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.

51. 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 50 on the facing page 5.46 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 5.7.16 on the facing page 5.46, the first and last assignments have comparison error of substantial magnitude in the context of the hypothetical data in Table 5.7.15 on page 5.45;
- ● 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* – see also page 5.85).

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.

52. Two illustrations of the matters in Note 51 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 5.7.17 at the right. The *de*creasing number of cases that progressed on to AIDS with increasing daily ribavirin dose indicated it was an effective treatment. Later, it transpired that ribavirin is *not* effective – the 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.

Table 5.7.17: Ribavirin Trial Data

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

- ⊙ 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 overleaf on page 5.48), 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.

(*continued overleaf*)

NOTES: (cont.) 52. ⊙ 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 sailiors 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;
- half a pint of sea water;
- two oranges and one lemon;
- 25 drops of elixir of vitriol;
- six spoonfulls of vinegar;
- a garlic, mustard seed, balsam and myrrh gum electuary.

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

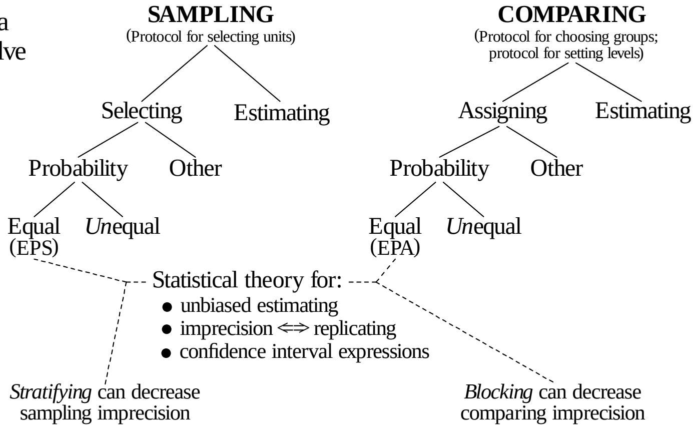
> On the 20*th* of May, 1747, I took twelve patients in the scurvy, on board the *Salisbury* at sea. Their cases were as similar as I could have them. They all in general had putrid gums, the spots and lassitude, with weakened knees. They lay together in one place, ... and had one diet common to all, ... 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 PPDAC 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]

53. Equiprobable *selecting* and equiprobable *assigning* are components of the processes of sampling and (experimental) comparing, whose *similarities* are illustrated in the discussion on page 5.46 of Table 5.7.16 and are portrayed by the two tree diagrams in the schema at the right below.


SAMPLING
(Protocol for selecting units)
Selecting
Estimating
Probability
Other
Equal (EPS)
*Un*equal
COMPARING
(Protocol for choosing groups; protocol for setting levels)
Assigning
Estimating
Probability
Other
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* unit inclusion probabilities in the selecting process; introductory statistics courses emphasize *equi*probable selecting as the basis of statistical theory for the behaviour of *sample* error under repetition.
  - 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 units in the groups (*e.g.*, control and treatment) being compared.
    - + Analogous to EPS, EPA is the basis of statistical theory for the behaviour of *comparison* error under repetition – recall the discussion on page 5.46 of Table 5.7.16.
    - + 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.

(*continued*)

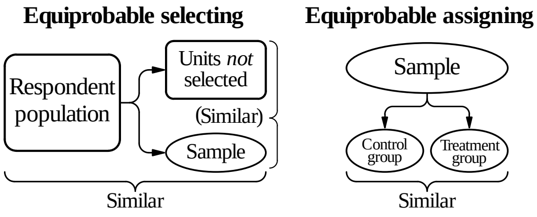
Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 15)

NOTES: (cont.) 53. ● – Statistical theory is *used* in the estimating branches of the two tree diagrams in the schema at the lower right of the facing page 5.48; these branches are part of the Analysis stage of the PPDAC cycle.

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

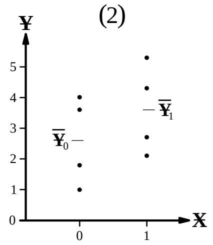
– The schema at the lower right on the facing page 5.48 reminds us of the analogous roles of stratifying and blocking in sampling and comparing [but recall the comment (–) near the top of page 5.37].

● As shown pictorially at the right, a common theme of EPS and EPA is dividing a group of units 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 units is the respondent population, the subgroups are the units *not* selected and the sample.

22. 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 units 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 units); two short horizontal lines show the two subpopulation average responses $\overline{Y}_0$ and $\overline{Y}_1$ [as they also do in diagram (1) at the lower right of page 5.45]. The difference between $\overline{Y}_1$ and $\overline{Y}_0$ for the two sub*populations* has two components:

∗ the *treatment effect* arising from their different 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 (5.7.5) below;

$\overline{Y}_1 - \overline{Y}_0$ = effect of change in X + effect of change in Z_1, ..., Z_k = treatment effect + confounding effect. -----(5.7.5)

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

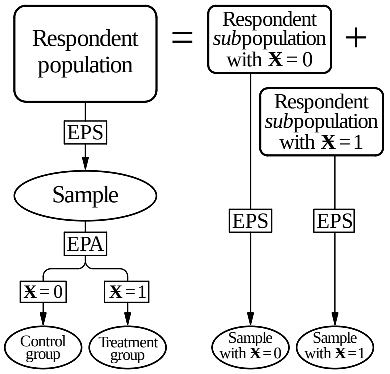
Assessing Answers from observational Plans must take account of the confounding effect because:

– 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 23 overleaf – *e.g.*, in Schema O on page 5.51.

NOTE: 54. The schema at the right below shows two ways we think about a respondent population in comparative investigating.



● On the left, we think of all units having the focal variate value $X = 0$ and, in an *experimental* Plan, a sample selected by EPS is divided in half by EPA, with one half retaining the value $X = 0$ and the other being assigned $X = 1$; the two (half) samples are then compared appropriately to answer the Question(s).

– An illustration is a clinical trial of a drug – $X = 0$ represents taking *no* drug (usually taking a placebo in practice) and $X = 1$ represents taking the drug.

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

● On the right, the 'natural' values of 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 units to the groups.

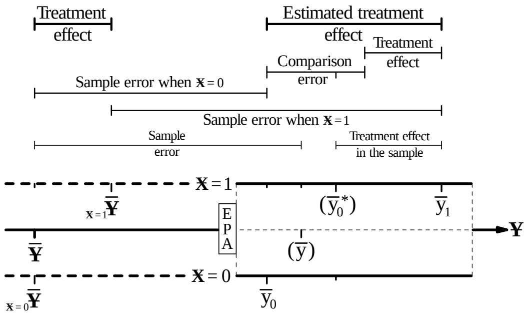
● It is also assumed for simplicity that the respondent population size is an exact multiple of the number of groups (*e.g.*, that N is *even* when there are *two* groups).

(*continued overleaf*)

23. 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 54 overleaf at the lower right of page 5.49, 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 at the upper right of the facing page 5.51.

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
$\mathbf{X} = 1$
$_{\mathbf{X}=1}\bar{\mathbf{Y}}$
E P A
$(\bar{y}_0^*)$
$\bar{y}_1$
$\mathbf{Y}$
$\bar{\mathbf{Y}}$
$(\bar{y})$
$\mathbf{X} = 0$
$_{\mathbf{X}=0}\bar{\mathbf{Y}}$
$\bar{y}_0$

- In schema E representing an *experimental* Plan, the respondent population has (unknown) average $\bar{\mathbf{Y}}$ and the sample selected from it by EPS has (unobserved) average $\bar{y}$.
  - ⇒ The *difference* in the values of $\bar{\mathbf{Y}}$ and $\bar{y}$ is *sample error*; its value remains *un*known in a particular investigation.
  - – We also denote the respondent population average, when all units have $\mathbf{X} = 0$, by $_{\mathbf{X}=0}\bar{\mathbf{Y}}$ and, when all units have $\mathbf{X} = 1$, by $_{\mathbf{X}=1}\bar{\mathbf{Y}}$; 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 55 on the facing page 5.51 and page 5.52.
  - – Comparison error (from both sources) is *eliminated* in the (unattainable) ideal of our three criteria (at the top of page 5.32) 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 $\mathbf{X} = 0$. -----(5.7.6)

  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 5.91 to 5.94 in Figure 5.8 – see also Figure 13.1 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 5.7.16 on page 5.46.

  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 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 upper right of the facing page 5.51) representing an *observational* Plan, the two respondent subpopulations with focal variate values $\mathbf{X} = 0$ and $\mathbf{X} = 1$ have respective (unknown) averages $\bar{\mathbf{Y}}_0$ and $\bar{\mathbf{Y}}_1$.

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

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 16)

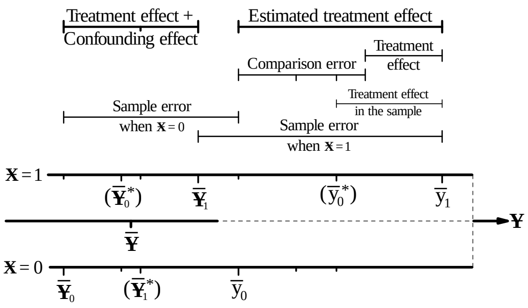
+ From Section 22 on page 5.49, the **confounding effect** in an *observational* Plan is the (unknown) difference between $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ and the treatment effect;

i.e., $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ = treatment effect + confounding effect. -----(5.7.5)

+ 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 units of each subpopulation have the *other* value of the focal variate.

Schema O for an Observational Plan


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

○ 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 $\mathbf{X} = 0$. -----(5.7.7)

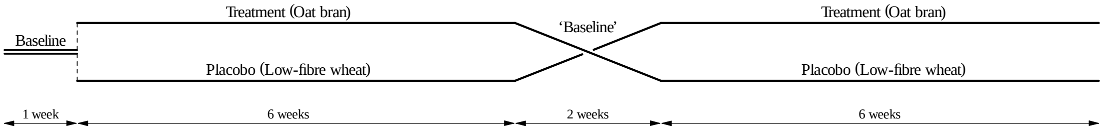
Comparing equations (5.7.7) and (5.7.6) [on the facing page 5.50], we see why an Answer about a treatment effect from an *observational* Plan has more severe limitation imposed by comparison error than such an Answer from an *experimental* Plan – equation (5.7.7) has the additional confounding effect 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 15 on pages 5.82 to 5.84.

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: 55. The (half) sample average $\overline{y}_0^*$ in schemas E (on the facing page 5.50) 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 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 $\overline{y}_0^*$, but this Plan really just yields values for $\overline{y}_0$ and $\overline{y}_1$ for *all* 24 participants.

(*continued overleaf*)

NOTES: (cont.) 55. ○ The non-focal explanatory variates $\mathbf{Z}_i$ (the blocking factors) made the *same* when $\mathbf{X} = 0$ and $\mathbf{X} = 1$ 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 5.7.11 in Note 38 near the middle of page 5.39.

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]

56. The discussion of this Section 23 starting on page 5.50 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).

24. Summary of Error Management Strategies

For convenient reference, Plan components to manage the six categories of error (listed again below), from Tables 5.7.2, 5.7.4, 5.7.6 and 5.7.10 (on pages 5.22, 5.26, 5.28 and 5.38), are given together in Table 5.7.18 below and on the facing page 5.53:

∗ study error;	∗ measurement error;	∗ model error;
∗ sample error;	∗ non-response error;	∗ comparison error.

Table 5.7.18

Plan Component			Error category	Error Management Strategy
Specifying the study population/process			Study	Specify the study population/process so its attribute(s) can be anticipated to be adequately close (in value) to those of the target population/process. ● *Restricting* values of explanatory variates can reduce variation in the study population/process – this may *de*crease sample error but *in*crease study error. ○ Sample error is preferred because statistical methods to manage it are better defined than the extra-statistical knowledge usually needed to manage study error.
Selecting units	Method of selecting	EPS	Sample	**EPS** is the basis of sampling theory which provides for: ● unbiased estimating of the respondent population average by the sample average; ● quantifying the likely size of sample error under repetition [*i.e.*, quantifying sampling imprecision, which we take here as 'quantifying uncertainty'].
		Judgement		**Judgement selecting** aims to make sample error as small as needed in the context of the *particular* investigation. ● It provides no basis for assessing if this aim has been achieved.
	Sample size (Replicating)	EPS		**EPS**: Sampling imprecision *de*creases with *in*creasing sample size (see Appendix 4 on page 5.59).
		Judgement		**Judgement selecting:** increasing sample size usually decreases the difficulty of making sample error as small as needed in the Question context. BUT: ● There is no theoretical basis which relates sample size to sampling imprecision.
	Stratifying the respondent population			Decreases sampling imprecision under EPS from the (properly-chosen) strata. ● Provides attribute estimates for the strata as well as for the respondent population.
Measuring variates	Imprecision		Measurement	Use a measuring process whose inaccuracy is acceptable in the Question context. ● Inccuracy of a measuring process does *not* necessarily decrease with its *cost*. ○ Inaccuracy is managed by using standards (where they exist – see Note 3 on page 5.20) to *calibrate* the measuring process.
	Inaccuracy			Use a measuring process whose imprecision is acceptable in the Question context. ● Decreased imprecision for a measuring process usually entails a more *costly* process but the converse is *not* always true.

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 17)

Table 5.7.18 (continued)

Plan Component	Error category	Error Management Strategy
Estimating attribute values Simple Ratio Regression	Sample	Under EPS, the sample average and sample standard deviation provide estimates, with defined behaviour under repetition, of the corresponding respondent population attributes. When estimating the respondent population average or total under EPS, ratio and regression estimating improve the (simple) estimate by using the respondent population average or total of an *explanatory* variate with a (strong) positive association with the response variate whose attribute is of interest. ● Ratio estimating decreases sampling imprecision when the standard deviation of the response variate increases linearly with the square root of the explanatory variate. ● Regression estimating decreases sampling imprecision when the standard deviation of the response variate does not change with the value of the explanatory variate. Ratio and regression estimating introduce *estimating bias* but can have smaller rms error than $\overline{Y}$ or $\mathbf{N}\overline{Y}$ as the estimator of the respondent population average or total.
Obtaining responses — Incentives Questionnaire Interviewer Call-backs Other	Non-response	Apart from a possible *legislated* requirement to respond (*e.g.*, to a population census), obtaining responses from units which are humans relies on *incentives* which include: ● a clear, answerable, succinct questionnaire; ○ when feasible, a questionnaire on *one* sheet of paper (or equivalent) is an advantage; ● properly trained interviewers; ● call back to units until those who are *un*available *are* contacted; ● appeal to altruism – respond to provide information that will benefit society; ● offer a material reward for response: ○ give *every* respondent a small item like a pen or a dollar coin; ○ offer respondents a chance to win a substantial prize like a trip. The skill and persistence of interviewers, developed by training, are a component of the incentives – see also Note 68 at the bottom of page 5.62. The clean separation of respondents and non-respondents is an idealization – partial (or 'item') non-response is also encountered in practice when sampled units provide some, but *not all*, of the information requested.
Imputing		**Imputing** is the process of assigning values for missing observations – *e.g.*, assigning a value for the reponse of a non-respondent on the basis of its values for known explanatory variates (like sex, age, location) that (it is hoped) are reasonable 'predictors' of the response variate. ● The purpose of imputing is to simplify the data analysis; it *rarely* meaningfully increases the completeness of the information in the data.
Assessing modelling assumptions EPS Form of the structural component Gaussianicity Probabilistic independence Equal standard deviations	Model	Limitations imposed by model error from two modelling assumptions are managed by: ● ensuring the selecting process for units *is* (equivalent to) *EPS*; ● ensuring variate values are measured *independently*. Assessing how well modelling assumptions appear to be met usually involves graphical displays (*e.g.*, scatter diagrams) of the estimated residuals from the response model; ● use a Gaussian quantile plot (or, sometimes, a histogram) to assess Gaussianicity; ○ transforming (*e.g.*, taking logarithms of) the data can help meet this assumption; ● use a plot in the time order of data collecting to assess probabilistic independence. ● use side-by-side dot- or boxplots, or a plot with the explanatory variate from the structural component of the model on the horizontal axis, to assess equality among, or dependence on an explanatory variate of, standard deviation(s).
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 5.7.12 and its discussion on the lower half of page 5.39.

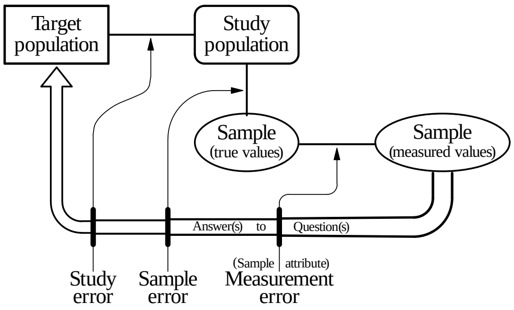
As an adjunct to Table 5.7.18, the 'error' schema (introduced at the centre right of page 5.20) is given overleaf on page 5.54 to show in one place the progressive developments of its four versions (from pages 5.20, 5.25, 5.27 and 5.30); version 2 (at the

(*continued overleaf*)

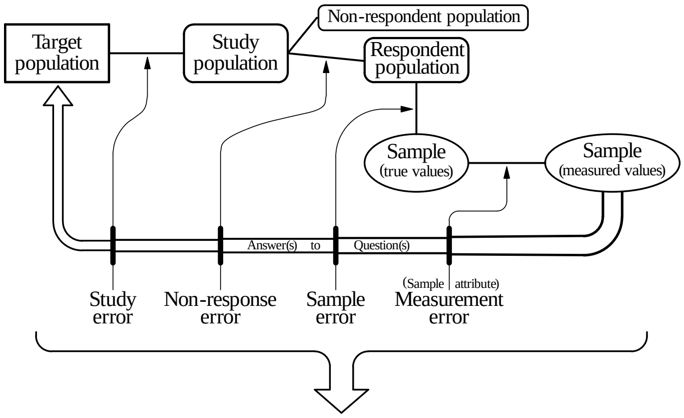
upper right) is supplemented by the diagram from the bottom right of page 5.25 (and its table of notation) which shows *overall* error as the (algebraic) sum of the first four error categories when answering a Question with a *descriptive* aspect. This diagram is, in turn, supplemented by schemas E and O from Section 23 on pages 5.50 and 5.51 and equations (5.7.6) and (5.7.7) as a reminder that, in comparative Plans for answering Questions (with a *causative* aspect) about (statistical) relationships:

- comparison error is an additional component of overall error;
- in *observational* Plans, comparison error imposes a more severe limitation on Answers than in experimental Plans because of the confounding effect (introduced in Section 22 on page 5.49).

Version 1 (page 5.20) **– 3 error categories**



Version 2 (page 5.25) **– 4 error categories**



Version 3 (page 5.27) **– 5 error categories**

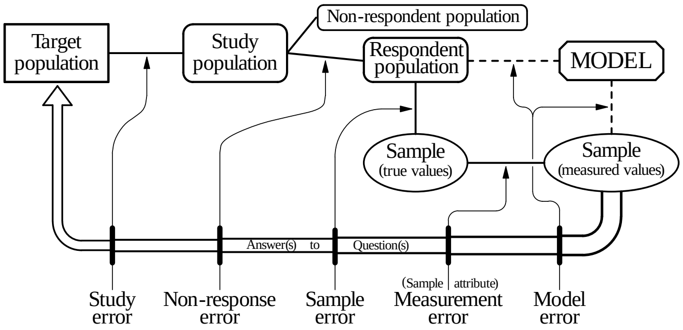
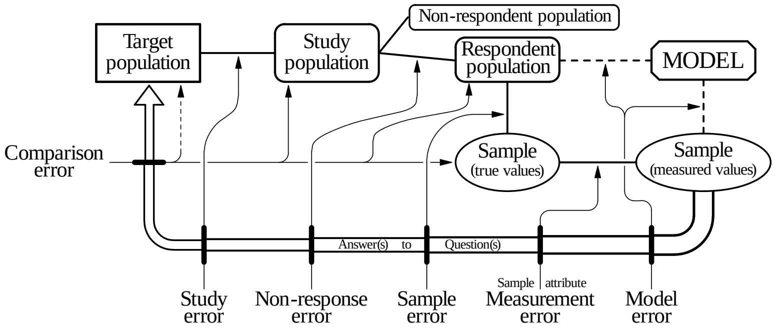
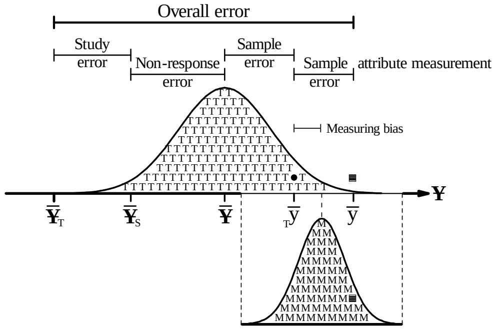


Table 5.7.3: SYMBOL DEFINITIONS

¥	Response variate
$\overline{¥}_T$	(True) target population average
$\overline{¥}_S$	(True) study population average
$\overline{¥}$	(True) respondent population average
$_T\overline{y}$	True average for sample selected
$\overline{y}$	Measured average for sample selected
T	True value of a sample average
M	Measured value of a sample average

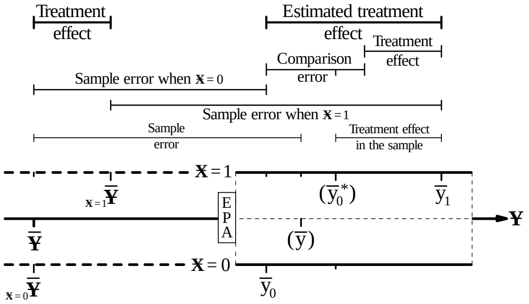
Version 4 (page 5.30) **– 6 error categories**





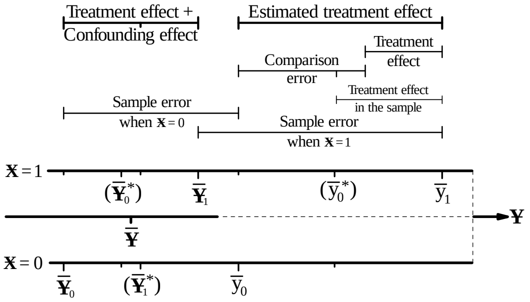
Discussion of the diagram above is at the left on the lower half of page 5.25.

Schema E for an Experimental Plan (page 5.50)



It is noteworthy that comparison error involves the difference of (two) *sample* errors.

Schema O for an Observational Plan (page 5.51)



In schema E: comparison error = sample error when X = 1 – sample error when X = 0. -----(5.7.6)

In schema O: comparison error = confounding effect + sample error when X = 1 – sample error when X = 0. -----(5.7.7)

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 18)

NOTES: 57. 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 35 near the bottom of page 5.37, near the end of Note 39 on page 5.40 and see the discussion of study error and sample error in Note 95 on page 5.84 in Appendix 16.

58. There is further discussion of error categories in Appendix 16 on page 5.84. Error categorization in the social sciences is compared with our terminology in Appendix 17 on pages 5.84 and 5.85.

Eighteen Appendices now follow in the order they are cited in the Figure, except for the citations of Appendix 14; there is *no* citation of Appendix 19 on page 5.86. Some Appendices involve idea(s) discussed *after* the Appendix is cited, because *strictly sequential* presentation of the ideas of introductory statistics is impossible – some terminology must be used *before* it is defined, a difficulty addressed here by having a Glossary (of 10 sides) as Figure 1.5 of these STAT 231 Course Materials.

25. Appendix 1: Populations and Processes (cited on pages 5.20 and 5.86)

Discussion at the bottom of the first side of the Figure (page 5.19), and its extension to include the respondent and non-respondent populations on pages 5.24 and 5.25 in Section 5, involve *populations* of *units* and the *sample*:

∗ **Population:** a well-defined group of units *other than* a sample.

∗ **Unit:** a basic entity for which variate values could be obtained; *target* units make up the *target* population.

∗ **Target population:** the group of units to which the investigator(s) want Answer(s) to the Question(s) to apply.

∗ **Study population:** the group of units *available* to an investigation (*study* units make up the study population).

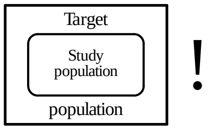
∗ **Respondent population:** those units of the study population that *would* provide the data requested under the incentives for response offered in the investigation;

∗ **Non-respondent population:** those units of the study population that would *not* provide the data requested under the incentives for response offered in the investigation.

∗ **Sample:** the group of units/blocks *selected* from the respondent population and *actually used* in an investigation; a sample is a *sub*set of the respondent population. – A **census** uses *all* the respondent population units/blocks.

[Blocking (introduced on pages 5.36 and 5.37 in Section 15) is included in these definitions to make them more general.]

Target units and study units are often the same but *may* differ; for instance, when assessing drug efficacy and side effects using laboratory animals, target units are humans but study units are laboratory animals (recall the newspaper article *Miracle cure that kills claims fifth victim* in Figure 3.1 of the Course Materials). The possibility of different target and study units may be obcured by a (misleading) diagram like the one at the right which shows the study population as a *subset* of the target population.



- Such a diagram may also show the *sample* as a subset of the study population [but see the four schemas from pages 5.20, 5.25, 5.27 and 5.30 (which are shown again together on the facing page 5.54) and the comment at the bottom of page 5.19].

To answer some types of Question, instead of a population, we start with a *process*; we distinguish two cases:

∗ **Process:** ● a set of *operations* that produce or affect units, OR:
● the *flow* of an entity (like water or electrons).

The first case arises when the Question is about improving a manufacturing or service-delivery process; we quantify the performance of the process by measuring variate values on the units it produces or affects.

– The target process is typically the process now and into the future for as long as the current (or improved) implementation of the process operates.

The second case arises when the Question is about an entity that flows, like water in a river or electrons in a circuit or network; we quantify characteristic(s) of such processes by measuring variate values on the entity that flows.

– The target process is typically the process over a defined period of time.

An investigation with a target *population* will have a study *population*; an investigation with a target *process* that is a set of operations will also have a study *population* – the available units produced or affected by the process. An investigation with a target *process* that flows will have a study *process*, usually the target process over a restricted time period (see Table 5.7.19 at the right).

Table 5.7.19
Populations and Processes

Target population	----------------►	Study population
Target process: operations	-------►	Study population
Target process: flow	-------------►	Study process

NOTE: 59. In STAT 231, populations and samples are both made up of units but, elsewhere, we distinguish the entities that make up a population (*elements*) from what is selected for the sample (*units*) – for example, in cluster sampling, a unit consists of a *group* of elements. The element-unit distinction is pursued briefly in Note 102 on page 5.86 in Appendix 18 and in more detail in Appendix 1 on pages 8.56 and 8.57 of Figure 8.11 of the STAT 220 Course Materials.

(*continued overleaf*)

26. Appendix 2: Equiprobable (Simple random) Selecting – The Protocol for Selecting Units (cited on pages 5.20, 5.26, 5.85, and 5.86)

∗ The **protocol for selecting units**, sometimes called the **sampling protocol**, is (a description of) the process (to be) used to select, from the respondent population, the units that comprise the sample.

There are many processes used in practice to select samples; three of them are discussed in this Appendix 2:

○ **equiprobable** selecting, ○ **systematic** selecting, ○ **judgement** selecting.

∗ **Equiprobable (simple random) selecting [EPS (SRS)]:** all samples of size n units from a (respondent) population of size **N** units have probability $1/\binom{N}{n}$ of being selected.

- What we call *equiprobable* selecting is likely to be called *simple random* (or *random*) selecting (or sampling) elsewhere.
- *Equiprobable* refers to a *process*; we should *not* refer to an equiprobable (or random) *sample*.
- The *definition* of EPS is in terms of *sample* selecting probabilities, not *unit* inclusion probabilities; consequences of this distinction for a sample of size n are:
  - + under EPS, the inclusion probability is n/**N** for each unit in the respondent population; BUT:
  - + even if the inclusion probability is n/**N** for each unit in the respondent population, the selecting process is *not necessarily* EPS – see Table 5.7.50 on page 5.85 in Appendix 18;
  - + the sample selecting process is *not* EPS if, for each respondent population unit, the inclusion probability is:

    ⊙ not equal to n/**N** OR: ⊙ not equal to that of all other unit(s).

Refinement of the usage of 'EPS' and 'unit' are discussed on page 5.86 in Notes 98 and 102 in Appendix 18.

∗ **Systematic selecting:** one unit is selected by EPS from the first k units of the respondent (or study) population (k ≪ **N**) and then every k*th* unit is selected.

- Referring to the *first* k units of the respondent (or study) population implies an ordered (*e.g.*, alphabetic or numeric) list of these units; such a list (called a **frame**) may be real or conceptual (*e.g.*, a rule that would, if implemented, generate the list).
- For convenience, it is usually assumed that **N** = nk so *all* 1-in-k samples selected systematically are of the *same* size n.

∗ **Judgement selecting:** human judgement is used to select n units from the **N** units of the respondent population.

Judgement selecting is discussed in Note 10 on page 5.23, on page 5.39 in Section 17 and in Appendix 14 on pages 5.79 to 5.82.

NOTES: 60. Other named methods of selecting units for the sample, which are largely omitted from this discussion, include:

- **accessibility selecting:** selecting units (easily) *accessible* to the investigator(s) – for instance, the *top* layer in a basket of fruit or a truckload of potatoes or the *front* pallets or cartons in a large stack in a warehouse;
- **convenience selecting:** selecting units that are *conveniently* available to the investigator(s) – for instance, people with a medical condition of interest who are at a hospital or clinic nearby to the investigator(s);
- **haphazard selecting:** selecting units with*out* (conscious) preference by the investigator(s) – shoppers who pass the location of an interviewer in a mall or rats in a cage which are more easily caught for a laboratory test;
- **quota selecting:** selecting units according to values of specified explanatory variates (like sex, age, income for human units) so the sample distribution of each variate will (approximately) match that of the study population;
- **volunteer selecting:** asking for (human) volunteers, usually after a brief explanation of what the investigating will entail for units in the sample.

These names do not necessarily specify a *unique* selecting method – the first two methods overlap and all five involve some degree of 'accessibility' and/or 'convenience'.

Haphazard selecting is sometimes ***wrongly*** equated with 'random' selecting; *i.e.*, with our *equiprobable* selecting.

Quota selecting is a similar idea to *covering*, defined near the top of page 5.24 in Note 11.

Volunteer selecting is *not* to be confused with **volunteer** (or **voluntary**) **response**, a phrase sometimes used to indicate that *human* units can (usually) *choose* whether to respond, *i.e.*, whether to provide the requested data; a separate (measuring) issue is whether these responses are correct or truthful (see also Note 68 on page 5.62).

61. A simple image of how EPS is implemented is to have, in a box, a slip of paper labelled for each unit in the respondent population; the **N** slips are thoroughly mixed and then n are selected without replacement – the labels of these n slips specify the units that comprise the sample (or, more correctly, the selection).

- It is seldom recognized how much effort is needed to *really* mix ('randomize') a collection of items like tickets or slips of paper, whose 'rough' surfaces do not readily slide over each other; in contrast, there are striking images of *two* sets of 'slippery' plastic capsules in rotating drums used in the 1971 U.S. draft lottery in Program 8 entitled *Describing Relationships* of the video series *Against All Odds: Inside Statistics*.
  - In a similar vein, the magician and statistician Persi Diaconis comments in Program 15 entitled *What is Probability?* of *Against All Odds: Inside Statistics* that most people do not realize it takes up to about seven *vigorous* shuffles to properly 'randomize' a deck of cards.
- In practice, EPS would usually be implemented with computer software that makes use of an *equiprobable digit* (or *random number*) *generator* – a source that is equally likely to generate any of the digits 0 to 9 at any position of a string of digits of specified length. Equiprobable digits are also available in printed tables – see

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 19)

NOTES: (cont.) 61. ● Table 6 in Appendix B (located after Chapter 18) of these Course Materials. To use this approach, the units of the respondent population are usually thought of as being numbered (labelled) from 1 to –**N**.

62. For telephone surveys – used for political polling and market research, for example – a **two-stage** selecting protocol for units is often employed:

- in the **first** stage, (listed) telephone numbers of a sample of households in the relevant geographic area(s) are generated equiprobably;
- in the **second** stage, the person who first answers the call to each household in the sample is asked to pass the call to the eligible household member (a Canadian citizen for a political poll, a homemaker for market research) who had the most recent birthday; this procedure implements (roughly) EPS of the eligible household members.

An advantage of this two-stage selecting process is that, when the units are *people* but there is a readily available (cheap) frame of *households* (*i.e.*, **clusters** of units), the frame of household members need be generated *only* for those households in the sample and each such frame exists only in the mind of the person who first answers the telephone call. How accurately this person follows the interviewer's instructions affects the degree to which EPS is achieved at the second stage.

○ Because households have differing numbers of members, unit inclusion probabilities are *un*equal at the second stage; these *two* stages of equiprobable selecting therefore do *not* achieve EPS overall (*cf.*, Table 5.7.50 on page 5.85

○ Because of non-response, many more (typically about *four* times as many) households need to be selected at the first stage as are required for the final sample size; for example, a national poll of 1,500 people may require around 6,000 telephone numbers to be generated, and some of these may have to be called multiple times to reach the eligible household member – recall the newspaper articles EM9342 (reprinted on the overleaf side of Figure 3.4 of these Course Materials), EM9330 and EM9337 (reprinted in Figure 3.9).

63. Systematic selecting is discussed in this Figure 5.7 because it is commonly used in practice; however, *we* think of it as being *equivalent* to EPS by applying the restrictive assumption that the frame (from which every k*th* unit is selected for the sample) has the units arranged so any value of the response variate is equally likely to be anywhere on the list (an **equiprobably ordered frame** for a given response variate). Three illustrations are:

- If a list of UW Faculty of Mathematics students, arranged in alphabetical order by family name, is used as a frame for 1-in-8 systematic selecting, the sample of about 500 students would most likely be essentially equivalent to selecting the students equiprobably from the list if the Question(s) involve the level of student debt but *not* necessarily equivalent to EPS if the Question(s) involve country of birth.
- If a list of family physicians licensed in Ontario, arranged in alphabetic order by family name, is used as a frame for 1-in-100 systematic selecting, the sample of about 300 physicians would most likely be essentially equivalent to selecting the physicians equiprobably from the list if the Question(s) involve drug prescribing characteristics.
- If a list of all school teachers in Ontario, arranged in order by year of graduation, is used as a frame for 1-in-500 systematic selecting, the sample of about 300 teachers would most likely *not* be equivalent to selecting the teachers equiprobably from the list if the Question(s) involve remunerations levels (which tend to *in*crease with time since graduation).

Thus, in this Figure, we consider two approaches to achieving equiprobability for the sample selecting process:

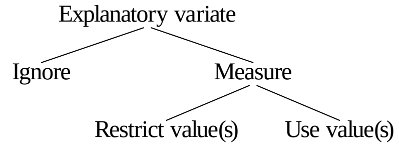
○ via an equiprobable **selecting process**, applied to a frame in *any* order;
○ via a *systematic* selecting process, applied to an equiprobably **ordered frame** (for a given response variate).

The second approach achieves (close to) equiprobability only under more restrictive conditions than the first approach.

27. Appendix 3: Fishbone Diagrams for Comparative Plans (cited on page 5.23)

Explanatory variates and their management is the central issue in investigating statistical relationships, as we see from the lengthy discussion of Sections 9 to 23 on pages 5.28 to 5.52. An aid to this management is what we call a **fishbone diagram**; properly constructing a fishbone diagram, as part of the process of developing a Plan for a comparative investigation, enables the investigator(s) to systematize their (statistical and *extra*-statistical) knowledge about explanatory variates. As summarized in the tree diagram at the right below, there are then three options for each (non-focal) variate in the fishbone diagram:

○ **ignore** it – that is, do *not* measure it;
○ measure it and **restrict** its value;
○ measure it and **use** its value [*e.g.*, to form *blocks* (recall page 5.36) or *strata*.]



An explanatory variate may be *ignored* for various reasons; for example, it may be:

– *unknown* to the investigator(s); OR:
– deemed *unimportant* in the investigation context;

a *poor* reason to ignore an explanatory variate is the cost or other difficulty of measuring it – it is debatable whether to undertake an investigation where resource constraints will only allow a Plan that may impose unacceptable limitation(s) on Answers.

(*continued overleaf*)

An explanatory variate may be *restricted* in value to reduce investigation cost – for example:

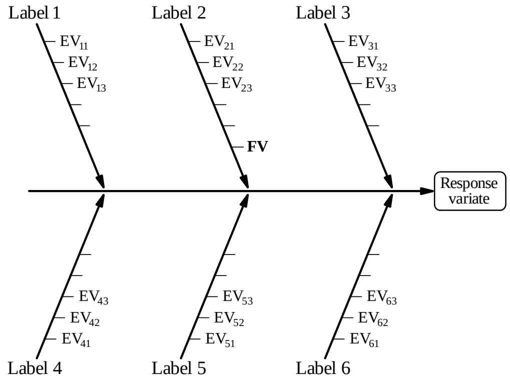
- a clinical trial of a new drug may decide to use participants of one sex and/or a restricted age range;
- when investigating a manufacturing process, the study population of parts might be specified as those parts still at the site – these would usually be parts produced consecutively over a relatively short time so their variation is likely to be *smaller* than the longer-term process variation.

The third option – *using* the values of an explanatory variate – is discussed earlier in this Figure 5.7; *e.g.*, in Note 14 on page 5.24.

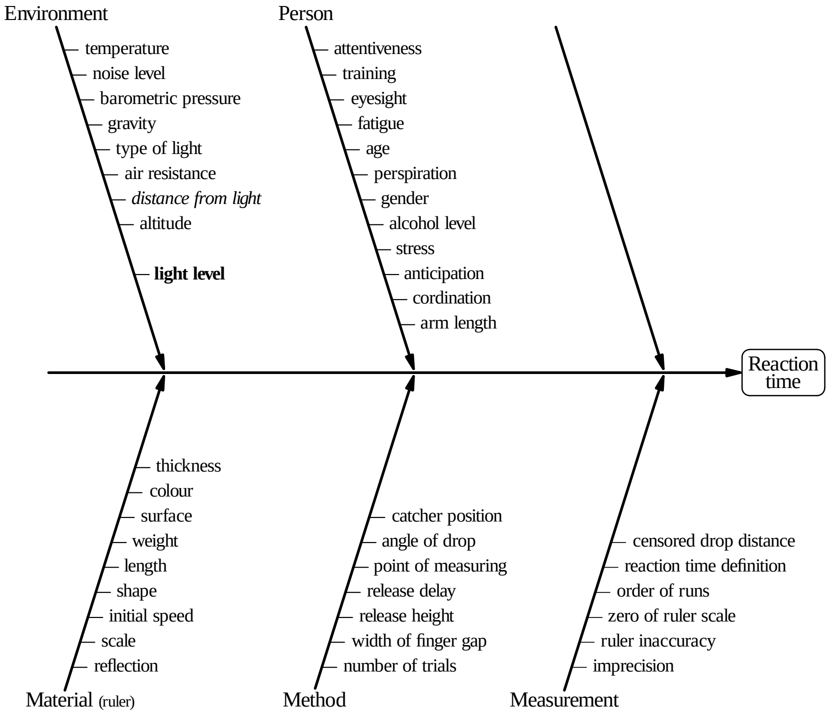
Choosing an appropriate option for each explanatory variate considered in an investigation usually requires extra-statistical knowledge and experience with data-based investigating; appropriate choice(s) can reduce limitation(s) on Answer(s), *in*appropriate choice(s) can impose unnecessary limitation(s). For example, restricting explanatory variate(s) may specify a study population which has unacceptably limited overlap with the target population, resulting in too severe a limitation imposed by study error.

As summarized in the diagram at right below, the components of a fishbone diagram are:

- a box on its right containing the name of the response variate,
- a central horizontal arrow pointing at this box,
- subarrows slanted from left to right and pointing either down or up at the central arrow;
  - ○ each subarrow has a label evocative (in the investigation context) of a category of explanatory variates useful in organizing them;
  - ○ names of explanatory variates (denoted 'EV' in the diagram) are associated with the slanted subarrows; some explanatory variates may themselves be broken down into components by small subsub-, subsubsub- (etc.) arrows;
    - + there can be more than one appropriate choice of subarrow for placing some explanatory variates;
    - + the focal variate (**FV**) is shown on the relevant subarrow.



An example of a fishbone diagram is given at the right below; its source is 'Laboratory 4,' one of five 2-hour practical exercises carried out by students in STAT 231 (lectures occupied the other eighteen 2-hour time slots scheduled for the course, one slot per chapter of the Course Notes). Laboratory 4 involved an experimental Plan to investigate the effect of light level on students' reaction time. Students worked in pairs – the 'dropper' held a 30-centimetre ruler above a gap between the 'catcher's' thumb and forefinger and reaction time was quantified by the distance the ruler fell between the catcher's fingers before it was stopped by closing them, after it was released by the dropper. There were two light levels; the high level had the usual classroom fluorescent lighting on, the low level had it off but an overhead projector on at the front of the classroom – the high light level was reasonably consistent across performances of the Laboratory, the low level was subject to the vagaries of the number of windows in a classroom but was usually low enough that some catchers did not catch the ruler as it dropped (an observation censored at 30 cm). Executing the Plan involved one run at each light level; half the student pairs (selected haphazardly) ran the high light level first, half ran the low level first. A measured non-focal variate was the distance of each student group from the overhead projector light source at the front of the classroom, quantified as 'floor tiles' (which were roughly 30 cm square). The fishbone diagram, produced as part of the Plan development, was facilitated by the course instructor from student input; only five of the six subarrows were used in this investigation context.

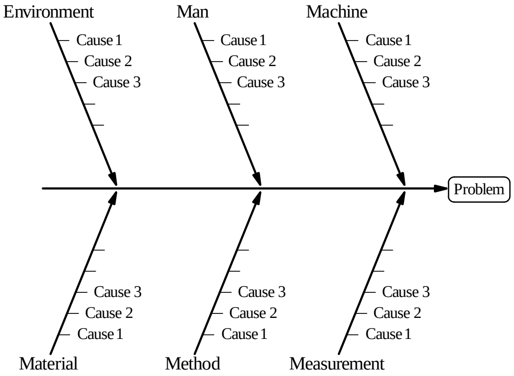


NOTE: 64. Our fishbone diagrams are an adaptation of cause-and-effect diagrams that are one of Ishikawa's seven industrial problem-solving tools:

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 20)

NOTE: (cont.) 64. 1. Check sheets
2. Pareto diagrams
3. Cause-and-effect diagrams
4. Histograms
5. Stratification charts
6. Scatter diagrams
7. Control charts;



comparing the cause-and-effect diagram at the right with the fishbone diagram at the centre right of the facing page 5.58, we see that differences are:

- the box at the right names the 'problem';
- the subarrows show possible *causes* of the 'problem'.

It is said that Ishikawa first used a cause-and-effect diagram in 1941 in a problem-solving session with engineers from a steel-making process.

The labels often used for the six subarrows are as shown and they reflect the industrial context of Ishikawa's problem-solving. However, in an automotive industry cause-and-effect diagram to address a problem of excessive transmission gear noise, for instance, the five subarrow labels were the components of the gear box: planet assembly, drum & sun gear, planet carrier, reverse gear, ring gear.

REFERENCES: 1. Ishikawa, K.: *Guide to Quality Control*. Asian Productivity Organization, 1982, and QR Quality Resources, White Plains, New York, ISBN 92-833-1035-7 (Casebound), 92-833-1036-5 (Limpbound).
2. Kane, V.E.: *Defect Prevention*. *Use of Simple Statistical Tools*. Marcel Dekker, Inc., New York and Basel, and ASQC Quality Press, Milwaukee, 1989, ISBN 0-8247-7887-1 (*e.g.*, pages 552 and 556).

Ishikawa's seven tools are also discussed in Figures 11.18 to 11.27 of the STAT 221 Course Materials.

28. Appendix 4: Sample Size and Sample Error under EPS (cited on pages 5.22, 5.23, 5.52 and 5.63)

This Appendix 4 illustrates properties of EPS introduced near the middle of page 5.23 in Note 10.

A respondent population of $\mathbf{N} = 4$ units has the following integer $\mathbf{Y}$-values for its response variate:

1, 2, 4, 5 [so that the population average and (data) standard deviation are: $\overline{\mathbf{Y}} = 3$, $\mathbf{S} \simeq 1.8257$];

we examine the behaviour of *sample error* under EPS as the sample size *in*creases from 1 to 2 to 3 to 4.

The number at the bottom of the four 'error' columns of Tables 5.7.20 at the right is the *average magnitude* of the sample error for that sample size.

Table 5.7.20a
EPS of n = 1 unit

Sample	$\bar{y}$	Error
(1)	1	–2
(2)	2	–1
(4)	4	1
(5)	5	2
		1½

Table 5.7.20b
EPS of n = 2 units

Sample	$\bar{y}$	Error
(1, 2)	1½	–1½
(1, 4)	2½	–½
(1, 5)	3	0
(2, 4)	3	0
(2, 5)	3½	½
(4, 5)	4½	1½
		⅔

Table 5.7.20c
EPS of n = 3 units

Sample	$\bar{y}$	Error
(1, 2, 4)	2⅓	–⅔
(1, 2, 5)	2⅔	–⅓
(1, 4, 5)	3⅓	⅓
(2, 4, 5)	3⅔	⅔
		½

Table 5.7.20d
EPS of n = 4 units

Sample	$\bar{y}$	Error
(1, 2, 4, 5)	3	0
		0

Tables 5.7.20 remind us of general results under EPS that follow from the theory in, for example, Section 5 on pages 8.52 and 8.53 of Figure 8.11 of the STAT 220 Course Materials.

- As the sample size *in*creases, the average magnitude (and, hence, the standard deviation) of sample error *de*creases – this is what we mean when we say that increasing sample size *de*creases sampling *imprecision* under EPS.
- Taking the *sign* of sample error into account, the average error is *zero* in each case – this behaviour is described as (the random variable representing) the sample average being an *unbiased* estimator of the respondent population average under EPS;
  - note that *both* the selecting method *and* the population attribute and its estimator are involved in this statement;
  - another statement with these components, which contrasts with the statement above about the random variable $\overline{Y}$, is that for the population attribute which is the *ratio* of the average of two response variates ($\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}}$), the sample ratio $r = \overline{y}/\overline{x}$ is *biased* [$E(R) \neq \mathbf{R}$] under EPS but *un*biased if the first sample unit is selected with probability proportional to its $\mathbf{X}$ value and the remainder selected equiprobably (see Cochran, page 175);
- There is no *sample* error when a **census** is taken – when *all* units of the respondent population are selected.

REFERENCE: Cochran, W.G.: *Sampling Techniquies*. Third edition, John Wiley & Sons, New York, 1977, ISBN 0-471-16240-X.

29. Appendix 5: Measuring Processes (cited on pages 5.24, 5.28 and 5.82)

Measuring processes are used to obtain *variate values* (*i.e.*, *data*); they exhibit *wide* variety and often involve technical matters from disciplines other than statistics. Some statisticians argue that measuring is therefore *not* part of Statistics, but these Course Materials take the position that:

○ Statistics answers Question(s) using *data*-based investigating; AND:

(*continued overleaf*)

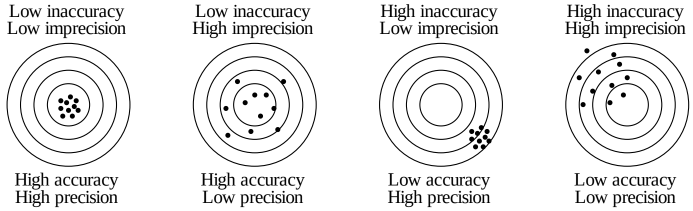
- ○ data are generated by measuring processes; SO THAT:
- ○ statisticians *must* be involved with the measuring process(es) used in an investigation to a degree that enables them to assess properly the limitation(s) imposed on Answer(s) by measurement (and, of course, other categories of) error.
  - – Assessing measurement error will usually be done in collaboration with other investigators who have relevant extra-statistical knowledge. (This may also be true of *other* categories of error – *e.g.*, study error).

Discussion (like this Appendix 5) of measuring processes is useful because it allows the *un*familiar idea of *error* (and ideas arising from it) to be presented in a context (measuring) with which readers have some familiarity.

- ○ An opportunity for *practical experience* with a measuring process is provided in Figures 3.6 and 3.7 of the Course Materials; there is other discussion of measuring in Figures 6.1 to 6.6.

When measuring on a unit a variate whose value does not change, we recognize that:

- ○ making *one* measurement provides a value for the variate but no information about **measurement error** – the difference between a measured value and the true (or long-term average) value of the variate;
- ○ making *more than one* measurement of the *same* variate on the *same* unit [the process of **repeated measuring** (or **repetition**)] and calculating their (data) standard deviation and average allows us to:
  - – *see* that repeated measurements of the same quantity usually do *not* agree (exactly) with each other;
  - – *quantify* **measuring imprecision** – the (data) *standard deviation* of the repeated measurements;
  - – *quantify* **measuring inaccuracy** – the *average* of the repeated measurements minus the *true* value being measured.
    - + Measuring a *known* value (*i.e.*, a **standard**) to quantify measuring inaccuracy is called **calibrating** the measuring process.
    - + A classic discussion of measuring inaccuracy by W.J. Youden is summarized in Figure 6.4 of these Course Materials.
  - – The *average* of repeated measurements is likely to be *closer* to the *long-term* average than an *individual* measurement – that is, the average has lower *im*precision (or higher precision) than individual measurements. Alternatively, we can say the *average* of repeated measurements is likely to have measurement error of *smaller magnitude* than an *individual* measurement.
    - + These (equivalent) statements are the meaning of the (familiar) idea that the *average* of repeated measurements is a 'better' value than just *one* measurement.
  - – We recognize that the sign and magnitude of error (which applies to a *particular* case) under *repetition* lead to the ideas of inaccuracy and imprecision, whose images are provided by patterns of shots on a target, as shown below at the right.
  - – Implicit in the idea of *repetition* is that the measurements are **independent**, meaning the operator's knowledge of the value arising from one execution of the measuring process does *not influence* the value from any other execution.
    - + This (more informal) meaning of 'independence' should not be confused with **probabilistic independence**, for which the probabilities of **events** *A* and *B* are such that Pr(*A*|*B*) = Pr(*A*) and Pr(*B*|*A*) = Pr(*B*) [see also Notes 87 and 88 on page 5.79 in Appendix 13 and Figure 7.8 of the STAT 220 Course Materials].



  - – When units are people and the measuring instrument is a questionnaire, repeated measuring is usually not feasible because respondents will likely recall their previous answers and so compromise the independence that is required statistically.
    - + When we distinguish measuring 'physical' variates (involving one or more of length, mass and time) from measuring that requires a questionnaire, the key *statistical* difference is compromised ability for repeated measuring of the same unit. There appear to be few (or no) *statistical* issues with the seemingly different nature of the *sources* of measurement error in the two situations [*e.g.*, imperfections in the components (discussed below) of the measuring process for 'physical' variates, ignorant and/or careless and/or untruthful responses to a questionnaire].
      - ⊙ In the same vein, for a questionnaire to measure self-esteem, for instance, it is a *subject-area* (extra-statistical) matter to define what is *actually* being measured (*e.g.*, see Figure 8.8e of the STAT 220 Course Materials).

  It is *un*clear how well the statistical issue of compromised repeated measuring is addressed by a questionnaire with *many* (perhaps hundreds of) questions, some of which are well-separated versions (or *in*versions) of the *same* question.

The foregoing discussion involves measuring variate values of *units* but the effect of a measuring process on an *attribute* value is more important statistically – recall equation (5.7.1) and the schema at the lower right of page 5.25. For example, under a model for measuring inaccuracy where bias is *constant* (*i.e.*, *not* dependent on the value measured):

- ● Inaccuracy will contaminate individual measured values and is *un*affected by *averaging* – this is the situation with the estimate of the *intercept* of a least-squares regression line.
  - – The average of measured values has lower *imprecision* than its individual measurements (*e.g.*, recall Note 10 on page 5.23).
- ● Inaccuracy is *zero* for a *difference*, which is typically involved in comparisons, in estimates of standard deviations, and in the *slope* of a least-squares regression line [recall equation (5.7.4) near the bottom of page 5.33].

We distinguish four *components* of a measuring process:

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 21)

∗ the **measuring instrument** or **gauge**; ∗ the **operator(s)**; ∗ the **measuring protocol**; ∗ the **unit measured**. Distinguishing these components makes it easier to identify *explanatory variates* which affect measured values and which may therefore be a source of measurement error; we consider in turn *statistical* matters associated with the four components.

Matters about the **measuring instrument** or **gauge** are:

○ Decreasing the *imprecision* of a measuring instrument usually involves increasing *cost*; for example, when measuring length:
- a *ruler* costs about $5 and can be read to 0.1 mm;
- a *pair of calipers* costs about $50 and can be read to 0.01 mm;
- a *micrometer* costs about $500 and can be read to 0.001 mm (1 micron);

in this instance, each decrease in imprecision by a factor of ten increases cost by about a factor of ten.

○ Higher *cost* of a measuring instrument does *not* necessarily mean higher *accuracy* (lower *in*accuracy).

○ In the context of a sample survey, the measuring *instrument* is the *questionnaire*; it is curious that investigators, who would *not* undertake assembly of the types of instrumentation used in a laboratory (*e.g.*, balances, spectrophotometers), often approach the task of developing the questionnaire with little recognition of the difficulty or importance of doing so successfully.

○ *Stability* of a measuring instrument – its ability to yield the *same* measured value in the same circumstances at points separated in time – is important but is not relevant to *all* measuring instruments; we usually distinguish:
- *short*-term stability;
- *long*-term stability.

What constitutes a 'short' or 'long' time scale for stability is context dependent.

Matters about the **operator** are:

○ In a clinical trial (used in medical research to assess, for example, the efficacy of a drug or surgical procedure), as indicated in Table 5.7.11 (given again at the right from page 5.39), blinding treatment *assessors* manages operator effect on *measuring inaccuracy* by trying to make assessment *independent* of the participant's treatment.

Table 5.7.11:

Blinding of ...	**Short name**	**Statistical purpose**
Participants	Single blind	Manage *comparison error*
Treatment administrators	Double blind	Manage *comparison error*
Treatment assessors	Triple blind	Manage *measuring inaccuracy*

- To be **blind** means not to know, for any unit, whether it is in the *treatment* group or the *control* group (which usually receives a dummy treatment known as a **placebo**).
- The short names in the second column of the table for the blinding are *not* recommended because they do not distinguish adequately among the eight possible combinations of which group(s) are blind.

Blinding of operator(s) as to the nature of the sample being measured may also be used in a medical diagnostic laboratory, where measuring *inaccuracy* is managed by analyzing **standards** at regular intervals concurrently with the primary task of analyzing biological materials.

○ In a self-administered questionnaire (received in the mail, for example), the *operator* is also the *respondent.*

The **measuring protocol** is the instructions for how to use the measuring instrument; one of its purposes is to promote *uniformity* in how different operators make measurements and so to try to make negligible, in the context of the investigation, any operator effect on the measured value obtained from the measuring instrument.

○ Clear measuring protocol(s) and adherence to them by operators on different shifts are vital in a multi-shift manufacturing operation if a consistent product is to come from the different shifts (see also Notes 66 and 67 overleaf on page 5.62).

Matters about the **unit measured** are concerned with the act of measuring *changing* the unit being measured or the value it yields. For example:

○ Maclean's ranking of Canadian universities might make universities change their operations in ways that would improve their ranking but make no substantive change to the quality of the educational experience they offer students.

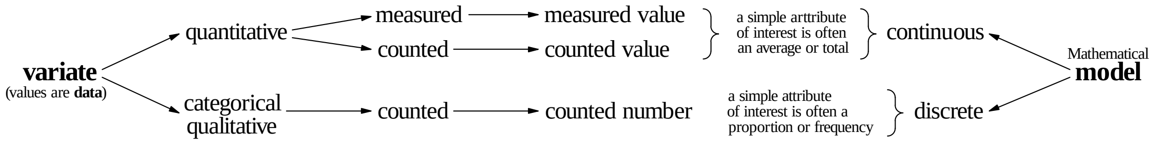
○ Households selected for a panel used to obtain Nielsen ratings of TV programs might change their TV viewing habits as a consequence of *knowing* their viewing habits are being monitored.

○ The interviewer administering a questionnaire (the 'operator') might (unintentionally) influence the person responding.

○ A *slanted* question on a questionnaire may have a different effect on different (types of) respondents.

An extreme case is when measuring *destroys* the unit (*e.g.*, in quality assurance, firing shotgun cartridges or measuring the bursting pressure of plastic bags and condoms); destructive measuring precludes the statistical benefits from repeated measuring on the same unit. (This is the same *statistical* issue as attempting repeated measuring when a questionnaire is involved).

NOTES: 65. The following schema illustrates distinctions among the terms *quantitative* and *categorical* (or *qualitative*), *measured* and *counted*, and *continuous* and *discrete*, when they are used as a qualifier of *variate.*



(*continued overleaf*)

NOTES: (cont.) 65. ● An illustration of the distinction between a counted *value* [a real number, modelled by a **continuous** variate] and a counted *number* [an integer, modelled by a **discrete** variate] is:

- to assess the effectiveness of an insecticide, the number of insects could be counted on a defined part of each of a sample of plants from *un*treated and treated crops – a measure of effectiveness would be the decrease in the *average* number of insects per plant;
- to assess gender balance in an area of employment, the number of men and women employed in the area could be counted – a measure of the balance would be the *proportion* of each sex in the area.

The quantitative *measured* and quantitative *counted* distinction blurs progressively as counted values become larger in magnitude; it is also affected by the limited resolving power (finite precision) of real measuring systems.

Continuous and discrete variates are *model* concepts because real measuring instruments with finite precision can yield only *discrete* values.

● *Quantitative* variate values can become (ordinal) *categorical* – *e.g.*, ages can be classified into age *groups*; *we* take *qualitative* to mean nominal (*non*-ordinal) categorical – *e.g.*, marital status or skin colour.

● A *binary* variate is a categorical variate in *two* categories.

66. When suppliers and assembly operations disagree about whether manufactured parts meet specifications, a common reason is measuring *inaccuracy* – the measuring processes used to check the parts at the supplier and assembly plants *dis*agree because they have not been calibrated to standards that *agree* with each other.

67. Assessing the imprecision and inaccuracy of the measuring process(es) to be used in an investigation, and the factors which affect them, is a common reason for one or more *sub*-PPDAC cycle(s) within the 'main' PPDAC cycle. An example is an industrial **gauge R&R** investigation.

∗ **Repeatability** of a gauge is the variation [expressed as an appropriate (data) standard deviation] of repeated measurements on each of a sample of (10, say) parts by *one* operator using the gauge;

∗ **Reproducibility** of a gauge is the between-operator variation [expressed as an appropriate (data) standard deviation] of two measurements, one by each operator using the gauge, on each of a sample of (10, say) parts.

- The two operators are usually assumed to have *equal* repeatability.

Repeatability quantifies the imprecision of a gauge under the most *favourable* conditions for operator effect. Reproducibility quantifies how this (lowest) imprecision is affected (increased) by having *two* operators.

● Investigators undertaking a measuring process assessment should take to heart the comments, made in 1966, by the U.S. National Bureau of Standards (now the National Institute of Standards and Technology), one of the world's premier measuring organizations:

> A major difficulty in the application of statistical methods to the analysis of measurement data is that of obtaining suitable collections of data. The problem is more often associated with conscious, or perhaps unconscious, attempts to make a particular process perform as one would like it to perform rather than accepting the actual performance ..... Rejection of data on the basis of arbitrary performance limits severely distorts the estimate of real process variation. Such procedures defeat the purpose of the ..... program. Realistic performance parameters require the acceptance of all data that cannot be rejected for cause.

SOURCE: Freedman, D., Pisani, R. and R. Purves: *Statistics*. First Edition, W. W. Norton & Company, New York, 1980, page 95.

68. A measuring process of statistical interest is so-called **randomized response**, whose simplest version involves an iterviewer asking a 'Yes/No' question about past 'sensitive' behaviour of the person being interviewed (the 'interviewee'), usually with the goal of estimating the population proportion of people who will admit they have engaged in the behaviour (*e.g*., abortion, illicit drug use, viewing child pornography, money laundering, terrorism); randomized response was developed to manage two difficulties such investigations encounter:

∗ the interviewee may find the question too sensitive to give a truthful answer, AND/OR:

∗ the interviewer may be under a legal obligation to report the behaviour of a respondent who answers 'Yes'.

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

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

where the first question has a *known* distribution of *answers*, the second question names the behaviour of interest.

The interviewee selects a card 'at random with replacement' from the (well-mixed) box and answers it.

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

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

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

● Question A4-16 (the 'three convicts' problem) raises a probabilistic issue with the warder's *response* to convict A.

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 22)

30. Appendix 6: Bias and Rms Error (cited on page 5.24)

For a random variable Y and some constant c (and where 'E' denotes probabilistic 'expectation'), we have:

$$\begin{aligned} E\{[Y-\mathrm{c}]^2\} &= E\{[E(Y)-\mathrm{c}+Y-E(Y)]^2\} = E\{[E(Y)-\mathrm{c}]^2+[Y-E(Y)]^2+2[E(Y)-\mathrm{c}][Y-E(Y)]\} \\ &= E\{[E(Y)-\mathrm{c}]^2\}+E\{[Y-E(Y)]^2\}+2E\{[E(Y)-\mathrm{c}][Y-E(Y)]\} \\ &= [E(Y-\mathrm{c})]^2+E\{[Y-E(Y)]^2\}+2[E(Y)-\mathrm{c}]E[Y-E(Y)] \end{aligned}$$

i.e., $E\{[Y-\mathrm{c}]^2\} = [E(Y-\mathrm{c})]^2+[s.d.(Y)]^2$ because $E[Y-E(Y)] \equiv 0.$ -----(5.7.8)

If we now think of Y as a random variable whose distribution represents the possible values of a *response variate* $\mathbf{Y}$ and c as a *true* value, the left-hand side of equation (5.7.8) is a **mean squared error** and $E(Y-\mathrm{c})$ in the first term on the right-hand side is a *bias*; we can therefore interpret equation (5.7.8) as:

mean squared error = bias^2 + standard deviation^2. -----(5.7.9)

Taking the square root so we are working on the *same* scale as the variate represented by Y, the **root mean squared error** is:

rms error = $\sqrt{\text{bias}^2 + \text{standard deviation}^2}$. -----(5.7.10)

Thus, the rms error is *one* concept that *combines* the two model quantities of bias and (probabilistic) standard deviation, or it can be used as a model for the two corresponding real-world entities of inaccuracy and imprecision.

Equation (5.7.10) provides useful insights about bias and variation in the context of survey sampling (to answer a Question with a *descriptive* aspect); different cases depend on how broad our focus is in terms of *which* true value c represents.

∗ The narrowest focus is *measuring* when c is the true value of the response variate $\mathbf{Y}$; equation (5.7.10) is then:

measuring rms error = $\sqrt{\text{measuring bias}^2 + \text{measuring standard deviation}^2}$. -----(5.7.11)

∗ For measuring *and* sampling, c is the true value of the *respondent* population attribute of $\mathbf{Y}$ and then:

measuring *and* sampling rms error = $\sqrt{\text{measuring} + \text{sampling bias}^2 + \text{measuring and sampling standard deviation}^2}$; -----(5.7.12)

NOTE: 69. Measuring and sampling standard deviation = $\sqrt{\text{measuring standard deviation}^2 + \text{sampling standard deviation}^2}$. -----(5.7.13)

∗ For measuring *and* sampling *and* non-responding, c is the true value of the *study* population attribute of $\mathbf{Y}$ and then, under our assumption that non-response is *deterministic* (*not* stochastic – recall Note 15 near the middle of page 5.26):

$$\begin{array}{r}\text{measuring and sampling and}\\ \text{non-responding rms error}\end{array} = \sqrt{\begin{array}{l}\text{measuring} + \text{sampling}\\ + \text{non-responding bias}^2\end{array} + \text{measuring and sampling standard deviation}^2}. \quad \text{-----(5.7.14)}$$

∗ For measuring *and* sampling *and* non-responding *and* specifying, c is the true value of the *target* population attribute of $\mathbf{Y}$ and then, under our assumption that specifying the study population also is *deterministic*:

$$\begin{array}{r}\text{measuring and sampling}\\ \text{and non-responding and}\\ \text{studying rms error}\end{array} = \sqrt{\begin{array}{l}\text{measuring} + \text{sampling}\\ + \text{non-responding}\\ + \text{studying bias}^2.\end{array} + \text{measuring and sampling standard deviation}^2}. \quad \text{-----(5.7.15)}$$

NOTES: 70. In printed materials other than these Course Materials (*e.g.*, see Cochran, p. 15), equation (5.7.8) [or (5.7.9)] is usually discussed only with respect to *estimating* bias. Although we have relatively little to say about estimating bias in STAT 231, it is useful to recognize the following [recall Appendix 4 on page 5.59]:

- *Estimating bias* (a *model* quantity) is the difference between the mean of an estimator and the value of the corresponding population attribute (or model parameter – recall the schema on page 5.28); for example, under EPS:
  - the random variable $\overline{Y}$ representing the sample *average* $\overline{y}$ is an *un*biased estimator of the respondent population average $\overline{\mathbf{Y}}$ because $E(\overline{Y}) = \overline{\mathbf{Y}}$ or $E(\overline{Y}) - \overline{\mathbf{Y}} = 0$; BUT
  - the sample ratio $r = \overline{y}/\overline{x}$ is a *biased* estimator of the respondent population ratio $\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}}$ because $E(R) \neq \mathbf{R}$ or $E(R) - \mathbf{R} \neq 0$, and likewise for S as an estimator of the respondent population (data) standard deviation $\mathbf{S}$ (although S^2 is an *un*biased estimater of $\mathbf{S}^2$) [*e.g.*, see the top and bottom of page 8.53 of Section 5 and Appendix 3 on page 8.57 of Figure 8.11 of the STAT 220 Course Materials].
- The rms error of an estimator is of interest because, while we prefer an *un*biased estimator of a population attribute, there are times when a *biased* estimator has only *small* bias and appreciably *smaller* standard deviation than an available *un*biased estimator; we *may* then prefer the biased estimator with *smaller* rms error.
- *Un*like (real-world) inaccuracy, estimating bias *de*creases in magnitude with increasing sample size – see, for example, Note 21 at the bottom of page 8.57 of Figure 8.11 of the STAT 220 Course Materials.

71. Equation (5.7.15) can be thought of as an extension to *variation* of the behaviour of averages (when answering a Question with a *descriptive* aspect) in equation (5.7.1) and its pictorial version at the lower right of page 5.25.

(*continued overleaf*)

31. Appendix 7: Illustrative Newspaper Article (cited on page 5.27)

EM0442: The Globe and Mail, November 23, 2004, page B9

Hooked on your cell? You must be Canadian

Study finds we talk 49 minutes a day on cellphones, double the global average

BY RICHARD BLOOM

As a representative of 16,000 students, Jennifer Green's cellphone is always on and almost always attached to her ear.

"On a given day, I could get between 15 and 50 phone calls from students, ranging anywhere from one minute all the way to 30 minutes," says Ms. Green, 24, a marketing student and the president of Humber College's students' federation in Toronto. She also frequently uses her phone for personal reasons.

"I don't know anyone in this day and age that doesn't own a cellphone. Constantly, cellphones are going off."

Ms. Green isn't alone in her chattiness. From salespeople to senior citizens to students, more and more Canadians are both buying mobile phones and ratcheting up talk time.

In fact, when you ask them, Canadians appear to be among the most talkative in the world, according to a study to be released today by the Canadian arm of mobile phone giant Telefon AB LM Ericsson.

Respondents said they talked an average of 49 minutes a day on cellphones, nearly double the global average of 27.

Only the United States is higher as respondents there said they talked an average of 63 minutes a day, the study reveals.

The study was conducted in conjunction with Starch Research and consisted of 2,000 in-home one-hour interviews with Canadians aged 15-69 across six provinces.

It didn't measure exact talk minutes provided by cellphone carriers but instead asked people's "perception" of how much they talk.

"It's very important for our customers to understand how they're perceived in the market ..... Perception is reality for most cases from a consumer perspective," said Vishnu Singh, Ericsson Canada Inc's manager of traffic and revenue growth.

He added that Canadians and Americans are used to unlimited talking on their wireline phones and that habit is being transferred to mobile phone usage.

What's more, most carriers offer free evenings and weekend packages, which means non-stop conversations have little impact on consumers' wallets, he said.

"In North America, we have big buckets of minutes and the cost of usage is quite low compared to many of the European countries where the tariffs are much higher," Mr. Singh said in an interview.

Users in Britain talked 32 minutes while those in Italy and China round out the top five at 30 and 27 minutes, respectively. The global figures were compiled from more than 14,000 interviews in 10 countries and are accurate to within 2.2 percentage points. (*sic*.)

The study also shows that 63 per cent of Canadians own a mobile phone, up from 56 per cent in 2003. That number, it says, should grow to 69 per cent in 2005.

"[Cellphones] are really becoming a lifeline for Canadians from a communications perspective," Mr. Singh said.

Mark Quigley of consultancy Yankee Group Canada said that, while he agrees cellphone usage has "grown dramatically" in recent years, he questions the talk-time figures. He said that according to data from the phone companies, Canadian talk an average of 347 minutes a month, or about 11.5 minutes a day – also below that of the United States – and cellphone penetration is about 45 percent.

"The minutes of use sounds very much out of whack," Mr. Quigley said.

Mr. Singh responded to that concern by repeating that the study was based on consumers' perception not actual billed minutes.

Still, Messrs. Singh and Quigley do agree that the cellphone won't be replacing the traditional home telephone any time soon.

"We're still a long way from outright replacing our landline," said Mr. Quigley, adding that Canadians use cellphones mainly as supplementary lines and are "cost sensitive" enough to make longer calls on a flat-fee home phone versus a pay-per-minute service.

Earlier this year, Humber's Ms. Green experimented with canceling her landline phone but ended her trial after three months, citing poor quality.

"It was horrible. The area that I was in, the signal wasn't very good ... I was wasting my minutes hanging up and trying calls again," she said.

Other highlights of the study:

- Use of short message service (SMS), also known as text messaging, has doubled since 2003, with 23 percent saying they send or receive an SMS message on a monthly basis. Nearly half of young Canadians (aged 15-24) say they use text message on a weekly basis.
- One in 10 young Canadian cellphone owners uses multimedia messaging services at least once a month, even though they were only introduced to this country last year.
- Of those with cellphones, 69 per cent say they never leave home without it.
- Fifty-seven per cent did not know it is possible to access the Internet on cellphones.

Matters of statistical interest raised by this article are:

- ○ A *clear* Question is needed – is the attribute of interest the average talk time over the population of *adults*, *telephone* users or *cellphone* users (or owners)?
  - – For a variate with a lower limit of zero and a few very high values, the *median* may be a better attribute than the average.
- ○ What was the *method* of selecting, and was the *frame* from which the sample was selected a list of:
  - – Canadians (who may or may not own a phone or a cellphone);
  - – times of the day (people who talk *longer* are then more likely to be selected)
  - – Canadian cellphone users (or owners)?
- ○ People who are more likely to *answer* the telephone are more likely to be selected for the sample.
- ○ Limitations on Answers due to measurement error arise from using respondents' *self-reported* 'perception' of their talk times – note the comments of Mr. Singh (near the bottom of the left-hand and top of the middle columns) and of Mr. Quigley (at the bottom of the middle and top of the right-hand columns).

(*continued*)

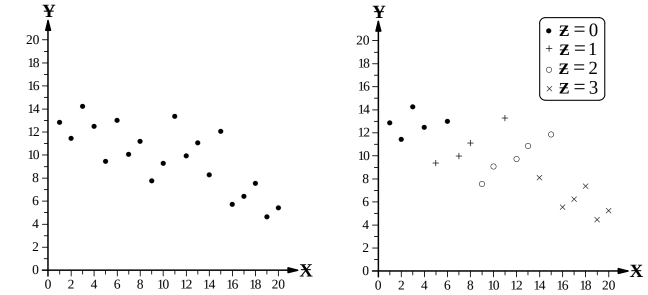
Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 23)

32. Appendix 8: Lurking Variates – Scatter Diagrams (cited on pages 5.29, 5.31, 5.73 in Appendix 11, 5.76 in Appendix 12 and 5.79 in Appendix 13)

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

∗ in the left-hand version in which Z values are *ignored*, we see an X-Y relationship that could reasonably be modelled by a straight line with a *negative* slope.

∗ in the right-hand version, where different symbols for the points denote four different values of some (non-focal) explanatory variate Z, the straight-line X-Y relationship can have a slope which is (close to) zero (when Z is 0), positive (when Z is 1 or 2) or negative (when Z is 3).



NOTES: 72. When looking at a scatter diagram of bivariate data to assess an X-Y relationship, we again recognize that experience *out*side statistics with diagrams involving Cartesian axes provides poor preparation for statistics – it is difficult for later statistical training to overcome a mindset (*un*concerned with lurking variates) that arises from more formative earlier experience with such diagrams, starting in elementary school, with on-going exposure in the media, and continuing up to post-secondary level courses, including calculus and algebra.

73. 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*.

74. The foregoing discussion and scatter diagrams in this Appendix 8 draw attention to the distinction between *conditioning* on Z and *ignoring* Z when investigating relationships.

∗ Conditioning is *subdividing*, as discussed at the upper left of page 5.67 in Section 1 of Appendix 9.

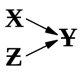
∗ A **marginal** (probability) distribution, referred to in Section 5 on page 5.68 of Appendix 9 and illustrated in Tables 5.7.30 to 5.7.33, is an example of 'ignoring' the variate which is absent from the marginal distribution – for instance, in Table 5.7.31, X_2 is absent, in Table 5.7.32, X_1 is absent, and in Table 5.7.33, Y is absent.

The scatter diagram at the right above shows the marginal distribution of X and Y if we think of the Z direction as coming vertically up from the page. With the Z values as given at the upper right of the right-hand version of the diagram and thinking of the page as the plane Z = 0, the first five points of the cloud would lie *on* the page; the remaining 14 points would then lie progressively further *above* the page in groups as one moves to the right across the diagram. This discussion reminds us that a marginal distribution is a *projection* – we *see* the marginal distribution of X and Y if we look vertically *down* on the diagram (*i.e.*, we look along the Z axis) to project the three-dimensional point cloud on to the two-dimensional plane of the page.

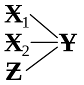
- It is interesting to speculate on the extent to which the ideas of conditioning and marginalizing (or projecting) provide a basis for understanding the ways in which mathematical models *approximate* reality (recall the maxim quoted in Note 19 on page 5.28).

33. Appendix 9: Lurking Variates – a Broader Perspective (cited on pages 5.30, 5.33, 5.40, 5.65, 5.70, 5.71, 5.72, 5.73, 5.75, 5.77, 5.78, 5.79 and 5.83)

(8)



$(A)_2$



In Section 10 on pages 5.29 and 5.30, the context of our introductory discussion of comparison error due to lurking variate(s)/confounding is comparative investigating of a *treatment* effect; the relevant *causal* structure from near the middle of page 5.34 is case (8), shown again at the upper right, with *focal* variate X, *respone* variate Y and lurking variate/confounder Z. In this Appendix, as summarized in the structure $(A)_2$ at the lower right, we broaden the discussion in two ways:

- we have two (or three) 'focal' variates [not necessarily all of equal interest in the Question context];
- we are *un*concerned with *causation* as the reason for the X_i-Y and Z-Y associations, because the nature of the focal variates is such that we can*not* set their levels and this precludes using such focal variate(s) to manipulate the value of Y – recall the discussion in Note 26 on page 5.33;
  – this is why the lower structure at the right has *lines rather than arrows* between the variate symbols.

(*continued overleaf*)

The phenomenon known as Simpson's Paradox can arise in a comparative investigation where the attributes are proportions – that is, the response variate ¥ is *qualitative* [discrete (categorical)] in nature; the dramatic name ('Paradox') is a reflection of how the effect of lurking variate(s) can *reverse* the sign of a relationship. The data in seventeen of the eighteen Tables 5.7.21 to 5.7.38 used in discussion of Simpson's Paradox in this Appendix 9 are hypothetical. The discussion is in six sections:

1. Illustrations of Simpson's Paradox
2. 'Simpson's Paradox' with a quantitative response variate.
3. Reasons for Simpson's Paradox – properties of proportions.
4. Reasons for Simpson's Paradox – population subgroups and weighted averages.
5. Reasons for Simpson's Paradox – probability distributions.
6. A Plan for an investigation to answer the Question of sex discrimination.

The discussion is framed in terms of *populations*, because there are no *inherent* sampling issues in Simpson's Paradox; when the groups being compared are *samples*, there is the additional statistical issue of managing sample error.

1. Illustrations of Simpson's Paradox.

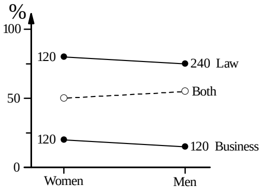
The data in Table 5.7.21 below come from the discussion of Simpson's Paradox in Program 11 of *Against All Odds: Inside Statistics*; the context is possible sex discrimination in graduate admissions. Overall, the admission *rate* [or *proportion* (an *attribute*)] is *lower* for women (50% *vs.* 55% for men – see the bottom line of the Table) but, when the data are subdivided by school (Law and Business), the female admission rate is *higher* (by 5 percentage points) for *each* school. The (binary) response variate is school admission (Yes, No) and the lurking variate is women-to-men ratio among applicants; its effect is because:

∗ the two schools had appreciably *different* admission rates: 80 and 75% for Law, 20 and 15% for Business;

∗ *half* as many women as men (120 *vs.* 240) applied to Law but *equal* numbers of women and men (120) applied to Business.

Table 5.7.21:

	...........WOMEN............			................MEN..............		
SCHOOL	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%
Law	120	96	**80**	240	180	**75**
Business	120	24	**20**	120	18	**15**
Both	240	120	**50**	360	198	**55**



The diagram to the right of Table 5.7.21 shows its data in graphical form; Simpson's Paradox is the *positive* slope of the middle dashed line for the data for *both* schools changing to a *negative* slope in the upper and lower lines for the schools *individually.*

In this illustration, the variates in the lower structure (A)$_2$ at the lower right overleaf on page 5.65 are:

Ӿ$_1$ is an applicant's sex (female, male), Ӿ$_2$ is the school applied to (Law, Business),

[In Tables 5.7.25 and 5.7.26 on the facing page 5.67, Ӿ$_3$ is the level of study (Masters, Doctoral)],

Ƶ is the (lurking variate) women-to-men ratio among applicants (discussed further in Sections 2 and 4 on pages 5.67 and 5.68),

¥ is the response to an applicant (admitted, not admitted). [Overleaf, ¥ is time for degree completion (minimum, longer).]

*Un*like investigating a treatment effect when there is more than one focal variate (*e.g.*, using a factorial treatment structure), the focal variate of primary interest in *this* Question context is Ӿ$_1$, an applicant's sex.

The limitation imposed by lurking variates on an Answer to a Question about an Ӿ-¥ relationship is illustrated further by the data in Tables 5.7.22 to 5.7.24; as the diagrams to the right of the tables emphasize, it is also possible to have:

∗ the *same* overall admission rate for women and men but a *higher* rate for women in the two schools individually (Table 5.7.22);

∗ a *lower* overall admission rate for women but the *same* rate for women and men in the two schools individually (Table 5.7.23);

∗ a *higher* rate overall *and* in the two schools individually for women (see Table 5.7.24).

Table 5.7.22:

	...........WOMEN............			................MEN..............		
SCHOOL	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%
Law	120	96	**80**	168	126	**75**
Business	120	24	**20**	120	18	**15**
Both	240	120	**50**	288	144	**50**

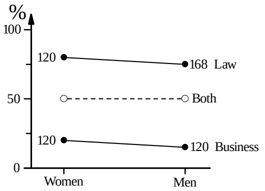


Table 5.7.23:

	...........WOMEN............			................MEN..............		
SCHOOL	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%
Law	120	96	**80**	240	192	**80**
Business	120	24	**20**	120	24	**20**
Both	240	120	**50**	360	216	**60**

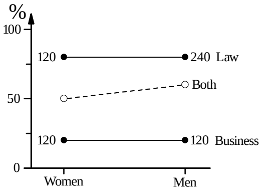
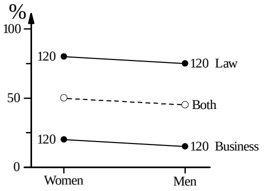


Table 5.7.24:

	...........WOMEN............			................MEN..............		
SCHOOL	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%
Law	120	96	**80**	120	90	**75**
Business	120	24	**20**	120	18	**15**
Both	240	120	**50**	240	108	**45**



The effect of lurking variates on an Ӿ-¥ relationship at a *second* level of subdivision is illustrated in Tables 5.7.25 and 5.7.26 at the upper right of the facing page 5.67; a context for these data is the proportion of graduate students who complete their degree in the minimum time. In Table 5.7.25, the proportion for women is *lower* overall, *higher* when subdivided by subject area (Law or Business) but again *lower* when subect area is subdivided by level (Masters or Doctoral). Similar effects are seen in Table 5.7.26, except the proportions for women become *equal* when subdivided by subject area and *higher* when further subdivided by level.

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 24)

Probabilistically, subdividing is *conditioning* so that Tables 5.7.21 to 5.7.26, in illustrating Simpson's Paradox, show the limitation on an Answer which involves comparing *conditional* probabilities for a response variate with *different* conditionings; that is, comparing probabilities for ¥ given $\cancel{\mathbf{X}}_1$ and $\cancel{\mathbf{X}}_2$ with ¥ given only $\cancel{\mathbf{X}}_1$ (in Tables 5.7.21 to 5.7.24) or for ¥ given $\cancel{\mathbf{X}}_1$, $\cancel{\mathbf{X}}_2$ and $\cancel{\mathbf{X}}_3$ with ¥ given $\cancel{\mathbf{X}}_1$ and $\cancel{\mathbf{X}}_2$ or ¥ given only $\cancel{\mathbf{X}}_1$ (in Tables 5.7.25 and 5.7.26) – see Section 5 overleaf on page 5.68. Four other illustrations of Simpson's Paradox are given in Note 76 on pages 5.69 and 5.70 and three more illustrative tables (like Table 5.7.29 overleaf on page 5.68) are discussed on pages 5.77 and 5.78 in Appendix 13.

Table 5.7.25:	............**WOMEN**............			................**MEN**..............		
	Number of	COMPLETIONS		Number of	COMPLETIONS	
SCHOOL	Students	Number	%	Students	Number	%
Law: Masters	60	51	**85**	60	54	**90**
Doctoral	60	33	**55**	300	180	**60**
Bus.: Masters	60	27	**45**	20	10	**50**
Doctoral	60	9	**15**	100	20	**20**
Law	120	84	**70**	360	234	**65**
Business	120	36	**30**	120	30	**25**
Both	240	120	**50**	480	264	**55**

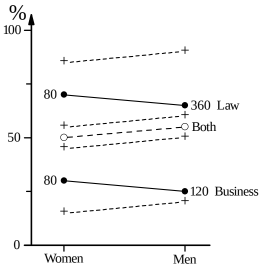
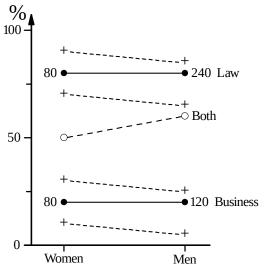


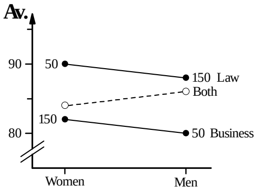
Table 5.7.26:	............**WOMEN**............			................**MEN**..............		
	Number of	COMPLETIONS		Number of	COMPLETIONS	
SCHOOL	Students	Number	%	Students	Number	%
Law: Masters	60	54	**90**	240	204	**85**
Doctoral	60	42	**70**	80	52	**65**
Bus.: Masters	60	18	**30**	120	30	**25**
Doctoral	60	6	**10**	40	2	**5**
Law	120	96	**80**	320	256	**80**
Business	120	24	**20**	160	32	**20**
Both	240	120	**50**	480	288	**60**



2. 'Simpson's Paradox' with a quantitative response variate

Simpson's Paradox is usually presented in the context of comparing *proportions* but the same phenomenon can occur with a *continuous* response variate. As illustrated by the data in Table 5.7.27 and the diagram to its right, whose context is graduate studies admission averages, the average is *lower* overall for women than men (84% *vs.* 86%) but, when the data are subdivided by school, both averages are *higher* for women. The response variate here is an applicant's average, the attribute is the *average* of these averages (*e.g.*, 90 and 88 for Law, 82 and 80 for Business) and the lurking variate is women-to-men ratio among applicants (1:3 for Law, 3:1 for Business). With 1:1 ratios, there is *no* 'paradox'.

Table 5.7.27:	........**WOMEN**........		...........**MEN**...........	
	Number of	Applicants'	Number of	Applicants'
SCHOOL	Applicants	Average (%)	Applicants	Average (%)
Law	50	**90**	150	**88**
Business	150	**82**	50	**80**
Both	200	**84**	200	**86**

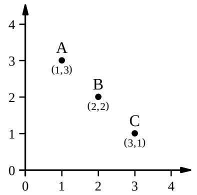


The illustration in Table 5.7.27 shows that Simpson's Paradox is *not* solely a phenomenon which may arise when comparing *proportions*. Its origin lies in the *relative* 'natural' group sizes arising from the process of subdividing (or its inverse of combining) used to manage comparison error in observational Plans. Such a lurking variate (called the women-to-men ratio in the discussion of Table 5.7.21 on page 5.66 and Table 5.7.27 above) is *different* in nature to Ƶ in the upper causal structure of case (8) at the lower right of page 5.65, which we think of as being able to *cause* a unit to change the value of its response variate. Thus, we now recognize *two* ways a change in a lurking variate can affect attribute value(s):

- by *causing* units' response variate (and, hence, their attribute) values to change, AND:
- by distorting attribute *calculation* when subdividing is used to manage comparison error in an observational Plan.

3. Reasons for Simpson's Paradox – properties of proportions

Quantities (like variate and attribute values) which are *single* numbers are relatively straightforward to compare: 4 is greater than 2 is greater than −6, although the latter has a larger *magnitude* than the first two. However, when quantities (like proportions or fractions and the coordinates of points on a scatter diagram) involve *two* numbers, comparisons may raise complications. For example, in the diagram at the right, points A and C with *different* coordinates are the *same* distance from the origin and point B is *closer* to the origin than A and C despite its coordinates being *larger* than one of those of A and C. The (surprising) result for fractions (or proportions), exhibited as Simpson's Paradox, is that for eight (positive) integers *a, b, ..., h,* it is possible to have (as in Table 5.7.21 on page 5.66):



$$\frac{a}{b} > \frac{c}{d} \text{ and } \frac{e}{f} > \frac{g}{h} \text{ but at the } \textit{same time} \text{ to have: } \frac{a+e}{b+f} < \frac{c+g}{d+h}; \qquad \textit{e.g.}, \frac{32}{40} > \frac{90}{120} \text{ and } \frac{12}{60} > \frac{9}{60} \text{ but } \frac{44}{100} < \frac{99}{180}.$$

This property also applies to *more* than two pairs of fractions (as in Table 5.7.35 on page 5.69) and for other combinations of inequality and equality (as in Tables 5.7.22, 5.7.23, 5.7.25 and 5.7.26 on the facing page 5.66 and above).

(*continued overleaf*)

When the fraction $\frac{90}{120}$ is instead $\frac{30}{40}$, we see at the right that there is *no* 'paradox' (as in Table 5.7.24), reminding us that it is the group sizes (in the denominators) under subdividing that may engage the property of proportions which generates the 'paradox' – recall Section 2 on page 5.67.

$$\frac{32}{40} > \frac{30}{40} \quad \text{and} \quad \frac{12}{60} > \frac{9}{60} \quad \textit{and} \quad \frac{44}{100} > \frac{39}{100}$$

4. Reasons for Simpson's Paradox – population subgroups and weighted averages

The distorted calculation of the values of (population) attributes (like proportions and averages), which generates the 'paradox' illustrated in Sections 1 and 2, is an instance of *weighted* combinations of the corresponding attributes of population *subgroups*. As shown in Table 5.7.28 at the right, the attribute values in the last line of each of Tables 5.7.21 to 5.7.24 are weighted combinations of the attributes in the two table lines above them; what produces the changes in attribute values *relative* to each other is a change in *weights*. Each weight is determined by the (natural) *size* of a population subgroup; this size is the *lurking* variate whose change is responsible for the change in the (sign of the) Ӿ-¥ relationship. The same idea applies to *each* of the *two* levels of subdivision in Tables 5.7.25 and 5.7.26 and to the averages in Table 5.7.27. When the weights are *equal* (as in Table 5.7.24), there is *no* 'paradox'.

Table 5.7.28:

	Weighted percentage	**Weights**	
Table 5.7.21:	$\frac{120}{240}\times\mathbf{80} + \frac{120}{240}\times\mathbf{20} = \mathbf{50}$	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{240}{360}\times\mathbf{75} + \frac{120}{360}\times\mathbf{15} = \mathbf{55}$	$\frac{2}{3}$	$\frac{1}{3}$
Table 5.7.22:	$\frac{120}{240}\times\mathbf{80} + \frac{120}{240}\times\mathbf{20} = \mathbf{50}$	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{168}{288}\times\mathbf{75} + \frac{120}{288}\times\mathbf{15} = \mathbf{50}$	$\frac{7}{12}$	$\frac{5}{12}$
Table 5.7.23:	$\frac{120}{240}\times\mathbf{80} + \frac{120}{240}\times\mathbf{20} = \mathbf{50}$	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{240}{360}\times\mathbf{80} + \frac{120}{360}\times\mathbf{20} = \mathbf{60}$	$\frac{2}{3}$	$\frac{1}{3}$
Table 5.7.24:	$\frac{120}{240}\times\mathbf{80} + \frac{120}{240}\times\mathbf{20} = \mathbf{50}$	$\frac{1}{2}$	$\frac{1}{2}$
	$\frac{120}{240}\times\mathbf{75} + \frac{120}{240}\times\mathbf{15} = \mathbf{45}$	$\frac{1}{2}$	$\frac{1}{2}$

5. Reasons for Simpson's Paradox – probability distributions

Table 5.7.21 on page 5.66 provides data from which the probability function of a discrete trivariate distribution can be estimated. To obtain this model, we first extend Table 5.7.21 as in Table 5.7.29 below to include three extra columns for 'Both sexes'. We then define five events and use estimates for ten probabilities – the vertical line means 'given that' in the eight *conditional* probabilities and ∩ denotes an *intersection* of events.

Table 5.7.29:

	...........WOMEN............			...............MEN..............			.......BOTH SEXES........		
	Number of	ADMISSIONS		Number of	ADMISSIONS		Number of	ADMISSIONS	
SCHOOL	Applicants	Number	%	Applicants	Number	%	Applicants	Number	%
Law	120	96	**80**	240	180	**75**	360	276	**76.6̇**
Business	120	24	**20**	120	18	**15**	240	42	**17.5**
Both schools	240	120	**50**	360	198	**55**	600	318	**53**

Event A: Applicant is admitted (¥ = yes; the complement Ā is ¥ = no)
Event F: Applicant is female (Ӿ$_1$ = female) Pr(F) = 0.4 Pr(A|F) = 0.5 Pr(A|F∩L) = 0.8
Event M: Applicant is male (Ӿ$_1$ = male) Pr(M) = 0.6 Pr(A|M) = 0.55 Pr(A|F∩B) = 0.2
Event L: Applicant applies to Law (Ӿ$_2$ = Law) Pr(A|L) = 0.76̇ Pr(A|M∩L) = 0.75
Event B: Applicant applies to Business (Ӿ$_2$ = Business) Pr(A|B) = 0.175 Pr(A|M∩B) = 0.15

The (joint) trivariate model is shown in Table 5.7.30 at the right below; summing its probabilities for one variate, we obtain the three (marginal) *bi*variate models in Tables 5.7.31 to 5.7.33. The smaller **bold** annotations in Tables 5.7.30 to 5.7.32 show how eight of the nine percentages in Table 5.7.29 arise; for example, the 80% of women admitted to Law is $\frac{0.16}{0.2}$.

Table 5.7.30: Trivariate model for ¥, Ӿ$_1$ and Ӿ$_2$

	.....F.....		....M.....		
	L	B	L	B	
A	0.16	0.04	0.3	0.03	0.53
Ā	**0.8** 0.04	**0.2** 0.16	**0.75** 0.1	**0.15** 0.17	0.47
	0.2	0.2	0.4	0.2	

Table 5.7.31: Bivariate model for ¥ and Ӿ$_1$

	F	M	
A	0.2	0.33	0.53
Ā	**0.5** 0.2	**0.55** 0.27	0.47
	0.4	0.6	

We see that Table 5.7.21 on page 5.66 involves *parts* of the two multivariate distributions in Tables 5.7.30 and 5.7.31; it is therefore *un*surprising if comparisons among these parts, taken in isolation, yield seeming 'paradoxes'. It can be confusing that Table 5.7.21 and those like it do not show *explicitly* percentages involving *complements* [like applicants '*not* admitted' (event Ā)].

Table 5.7.32: Bivariate model for ¥ and Ӿ$_2$

	L	B	
A	0.46	0.07	0.53
Ā	**0.76̇** 0.14	**0.175** 0.33	0.47
	0.6	0.4	

Table 5.7.33: Bivariate model for Ӿ$_1$ and Ӿ$_2$

	L	B	
F	0.2	0.2	0.4
M	0.4	0.2	0.6
	0.6	0.4	

6. A Plan for an investigation to answer the Question of sex discrimination

Comparing proportions of women and men admitted among applicants to graduate studies (as in the context of Table 5.7.21 on page 5.66) is *not* an adequate Plan to answer the Question of possible sex discrimination, for two reasons:

- there is the possibility of Simpson's Paradox and no clear way to define the level of subdivision at which to make comparisons;
- applicants' *qualifications* are not taken into account.

Both matters are addressed by a Plan which involves taking *pairs* of applicants, one female and one male, with the *same* qualifications for admission and then comparing the proportions of women and men who are admitted across a number of such pairs that is adequate, in the investigation context, to manage all relevant categories of error.

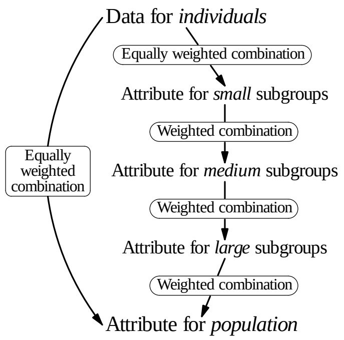
∗ For comparison error, pairing manages the *group sizes* (and hence, the weights in the attribute calculations) in a way that precludes Simpson's Paradox; matching manages equality of qualifications for the groups of women and men being compared.

However, as with *any* observational Plan (that gathers data from a population in its *natural* state), there is still the limitation on Answer(s) imposed by comparison error due to other (unrecognized) lurking variates.

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 25)

∗ When investigating the much-discussed issue of comparable worth (whether women are paid the same as men for the same work), relevant explanatory variates to manage include qualifications, experience and hours worked per month or per year.


Data for *individuals*
Equally weighted combination
Attribute for *small* subgroups
Weighted combination
Attribute for *medium* subgroups
Weighted combination
Attribute for *large* subgroups
Weighted combination
Attribute for *population*
Equally weighted combination

The schema at the right is a pictorial reminder of the lurking variate of group (population or sample) sizes when developing an observational Plan to answer a Question with a *causative* aspect, which (usually) involves comparing attribute values (calculated or obtained from a scatter diagram in the *Analysis* stage of the PPDAC cycle) for broad subpopulations (like women and men). By contrast, when answering a Question with a *descriptive* aspect (*e.g.*, a Question about *both* sexes), differing attribute values at different levels of subdivision are more obvious and so lurking variate(s) are usually less troublesome. These matters are illustrated, using information from Table 5.7.25 (at the top of page 5.67), in Table 5.7.34 at the right below.

Table 5.7.34: (Based on Table 5.7.25 data) **Group**		**Women** **Group size**	**%**	**Men** **Group size**	**%**	**Both sexes** **Group size**	**%**
Individuals		1	--	1	--	1	--
Smaller subgroups	Law: Masters	60	85	60	90	120	88
	Doctoral	60	55	300	60	360	59
	Bus.: Masters	60	45	20	50	80	46
	Docroral	60	15	100	20	160	18
Larger subgroups	Law	120	70	360	65	480	66
	Business	120	30	120	25	240	27
Population		240	50	480	55	720	53

NOTES: 75. Simpson's Paradox is so surprising, particularly when first encountered, that it is easy to lose sight of key *statistical* issues.

- The proportions *are* correctly calculated – Simpson's Paradox is *not* the result of mistakes in arithmetic.
- Simpson's Paradox is *not* confined to attributes that are proportions (as discussed in Section 2 on page 5.67).
- Simpson's Paradox occurs when subdividing (or combining) data for categories and only in *some* circumstances.

Lessons for data-based investigating are:

∗ recognize and manage the (surprising) property of proportions discussed in Section 3 on pages 5.67 and 5.68;

∗ manage relevant non-focal explanatory variates – this includes the possibility of sometimes being able to identify an appropriate level of subdivision at which to make comparisons (recall Table 5.7.12 on page 5.39).

There is then no 'paradox' for a *clear* Question investigated with an adequate Plan, suggesting that the name Simpson's *Paradox* can be misleading;

76. Four more illustrations of Simpson's Paradox are:

Table 5.7.35: The context is the same as that of Table 5.7.21 on page 5.66 but there are now *six* programs (A, ..., F) instead of two schools (Law, Business).

Like Table 5.7.21, there is a *lower* percentage of women admitted overall but a *higher* percentage for *each* of the six programs.

Table 5.7.35:	............**WOMEN**............			................**MEN**...............		
PROGRAM	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%
Archeology	108	89	**82**	825	512	**62**
Biology	25	17	**68**	560	353	**63**
Chemistry	593	219	**37**	325	114	**35**
Drama	375	131	**35**	417	138	**33**
English	393	106	**27**	191	48	**25**
French	341	27	**8**	373	22	**6**
All	1,825	589	**32**	2,691	1,187	**44**

Table 5.7.36: Baseball batting averages – the batter with the *lower* average for the whole season has a *higher* average in both *half* seasons. Recalling Section 6 and Note 75 above, it is of interest to develop a Plan to answer the Question of which batter to take if only one can be chosen.

Table 5.7.36:	......**BATTER #1**.....			.....**BATTER #2**......		
Time Period	Hits	At bats	**Average**	Hits	At bats	**Average**
First half	15	70	**.214**	25	130	**.192**
Second half	15	50	**.300**	80	280	**.286**
Whole season	30	120	**.250**	105	410	**.256**

Table 5.7.37: Death rates (per 1,000 lives) in two regions of the U.S. for smokers and non-smokers.

Table 5.7.37:	......**SMOKERS**......			**NON-SMOKERS**		
LOCATION	Deaths	Policies	**Rate**	Deaths	Policies	**Rate**
Nashville	6	900	**6.67**	7	1,100	**6.36**
Los Angeles	5	1,100	**4.55**	3	700	**4.29**
Either	11	2,000	**5.50**	10	1,800	**5.56**

[These data were gathered by a life insurance company which was issuing whole life policies countrywide on a non-medical issue basis; in 1986, 3,800 policies were issued to males aged 40-45. The company's files were kept in two locations – Nashville for policies issued east of the Mississippi and Los Angeles for policies issued west of the Mississippi. Nashville issued 2,000 policies and processed 13 deaths, Los Angeles issued 1,800 policies and processed 8 deaths.]

REFERENCE: Dolins, J.G.: Actuaries ... be careful!
The Actuary, March, 1989, page 11.

Table 5.7.38: Effect of jury challenges on conviction rates in trials in the U.K.

[In early 1987, an article by Bernard Levin in *The Times* raised the question of whether jury challenges

Table 5.7.38:	...**NO CHALLENGE**...			.......**CHALLENGE**......		
DEFENDENT STATUS	Number of Trials	CONVICTIONS Number	%	Number of Trials	CONVICTIONS Number	%
Guilty	20	16	**80**	70	42	**60**
Innocent	10	0	**0**	0	0	**0**
Either	30	16	**53**	70	42	**60**

(*continued overleaf*)

NOTES: (cont.) 76. assist those who are guilty in avoiding conviction. Mr. Levin concluded this was *not* the case, on the basis of data showing a conviction rate of 53% in trials with no challenges, *lower* than the conviction rate of 60% in trials *with* challenges. However, this answer does *not* necessarily follow from these conviction rates; in the *hypothetical* data in Table 5.7.38 (at the lower right overleaf on page 5.69), the conviction rate for *guilty* defendents is substantially *higher* in trials with *no* challenges. Unfortunately, this counter-argument is speculative because the number of defendents *actually* guilty and innocent, and the rates of challenge and of conviction in both these groups, are not readily accessible. Nevertheless, an article in a major newspaper which uses flawed reasoning from data to answer a Question on a substantive issue is a serious matter.]

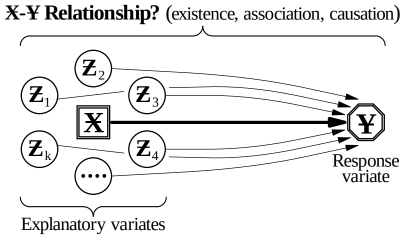
REFERENCE: Hill, I.D.: Rebutting the media. *The Royal Statistical Society NEWS & NOTES* **16** (#1), September, 1989, page 4.

There is discussion and further illustrations of Simpson's Paradox in Wagner, C.H.: Simpson's Paradox in Real Life. *American Statistician* **36** (#1, February): 46-48 (1982).

There are three other matters of statistical interest about the data in Tables 5.7.37 and 5.7.38 overleaf on page 5.69.

- What is a plausible explanation for the *lower* death rates for both smokers and non-smokers whose files were kept in Los Angeles, compared with those kept in Nashville?
- Develop a Plan which would be expected to obtain an Answer with fewer limitations about the effect(s) of jury challenges on conviction rates for the guilty in the U.K.
  - Even if the data in Table 5.7.38 were *real*, they would provide no Answer about the effect(s) of jury challenges on the conviction of *innocent* defendents because these data (rightly) show *no* such convictions.

34. Appendix 10: Confounding – Usage in Statistics (cited on pages 5.30, 5.45, 5.76 in Appendices 11 and 12 and 5.79 in Appendix 13)



As background to answering Question(s) about an X-Y relationship between a focal variate X and a response variate Y, Z_1, Z_2, ..., Z_k in the schema at the right are called **lurking variates**, a phrase that means lurking *explanatory* variates in that each Z accounts, at least in part, for changes from unit to unit in the value of the response variate. The importance of lurking variates is that if the distributions of their values *differ* between groups of units [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 Zs are taken into account.

A practical difficulty for data-based investigating of an X-Y relationship is that lurking variates are often *numerous* and so:

- important Zs or their differing distributions for different values of the focal variate can easily be overlooked, AND:
- substantial resources may be needed to measure values on the sampled units for those Zs 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 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 Z_i for differemt 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 units) differs for different values of the focal variate.

The following definitions (repeated from page 5.30) summarize the foregoing discussion:

∗ **Lurking variate:** a non-focal explanatory variate whose differing distributions of values over groups of 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 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 units with different values of the focal variate(s), comparison error is the difference from the *intended* (or *true*) state of affairs arising from:
– differing distributions of lurking variate values between (or among) the groups of units OR – confounding.

The alternate wording of the last phrase accommodates the equivalent terminologies of lurking variates and confounding; in a particular context, we use the version of the definition appropriate to that context:

- 'lurking variates' can more readily accommodate phenomena like Simpson's Paradox – see Appendix 9 on pages 5.65 to 5.70;
- 'confounding' is more common in the context of comparative Plans, as in Section 15 which starts on page 5.36, but the variety of usage of 'confounding' can be a source of difficulty (as discussed starting overleaf on page 5.71).

The dictionary meanings of 'confounding' in ordinary English include *confused*, *bewildered* and *mixed up* – the last of

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 26)

these three is closest to the statistical meaning given on page 5.30 in Section 10 (and repeated on the facing page 5.70), because the effects on $\mathbf{Y}$ of differences in $\mathbf{X}$ and in one or more of the $\mathbf{Z}$s are 'mixed up' (or 'cannot be separated' as it is also expressed) – recall the discussion of the confounding effect on page 5.49 in Section 22. The *difficulty* with the statistical usage is that different statisticians in different places may, without distinction, use 'confounding' to refer to any one of four of its facets:

- ⊙ the **definition:** inability (or failure) to separate the effects of $\mathbf{X}$ and $\mathbf{Z}_i$ [or $\mathbf{X}_i$ and $\mathbf{X}_j$] (which are *associated*) on $\mathbf{Y}$,
- ⊙ the **idea:** non-focal explanatory (or lurking) variate(s) $\mathbf{Z}_i$ *differ in value* for different $\mathbf{X}$ values,
- ⊙ the **limitation:** an Answer to a Question about an $\mathbf{X}$-$\mathbf{Y}$ relationship that may be meaningfully different from the 'truth',
- ⊙ the **consequence:** an Answer may be *altered* in a meaningful (*i.e.*, practically important) way if the values of (one or more) $\mathbf{Z}_i$ are taken into account.

This variety of usage, reflecting lack of agreement among statisticians about how broadly 'confounding' is to be interpreted, can obscure its underlying *idea*, the facet emphasized in our introductory discussion in Section 10 on pages 5.29 and 5.30; it can also be a source of confusion. [There is, of course, common ground among these facets (most obviously among the last three) because they all refer to the *same* phenomenon.]

As summarized in Table 5.7.39 below, one way to make these matters more transparent is to distinguish four contexts for 'confounding' in statistics; to do so in these Course Materials, *we* qualify 'confounding' with one of four adjectives:

● perfect, ● partial, ● general, ● selecting.

However, these adjectives and distinctions are particular to these Course Materials and are unlikely to be encountered or understood elsewhere – this is like our use of 'EPS from an unstratified population' instead of the usual 'SRS' (see Note 98 on page 5.86 in Appendix 18) and of 'EPA' instead of the usual 'randomization' (see the bottom of page 5.48 in Note 53 in Section 21).

Table 5.7.39: SUMMARY OF USAGE OF 'CONFOUNDING' IN STATISTICS
(Simpson's Paradox referred to below is discussed in Appendix 9 on pages 5.65 to 5.70)

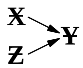
Description	Type	Impact	Facet	Illustration
Perfect confounding	1	Positive: Exploited in DOE	Definition	Fractional factorial treatment structure
Partial confounding	2	Negative: Imposes limitation on an Answer	Idea, limitation	'Confounding' in comparative Plans
General confounding	3		Limitation, consequence	'Confounding' *and* Simpson's Paradox
Selecting confounding	4		Consequence, limitation	Judgement selecting

The latter three facets of confounding are encompassed by our definition of comparison error on page 5.30 in Section 10, and repeated for convenience near the bottom of the facing page 5.70. A minor change in this definition, from that on page 5.30 in our introductory discussion, allows for the possibility of *more than one* focal variate to accommodate, for example, experimental Plans with a factorial treatment structure and phenomena like Simpson's Paradox (recall Appendix 9 on pages 5.65 to 5.70).

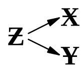
More details about the four facets of confounding and our distinctions are as follows:

- ○ Confounding ('perfect or type 1 confounding') is a term in the statistical area of *Design of Experiments* (DOE), where it indicates inability to (fully) separate the effects of two (or more) ***focal*** variates on a response variate; it can be exploited to achieve statistical benefits in Plans with a *fractional* factorial treatment structure – recall Note 47 on page 5.45.
  - – The adjective *perfect* for type 1 confounding indicates that levels of (some) *focal* variates and/or their interactions are associated with correlation of magnitude 1 – this is why (some of) the effects on a response variate *cannot* be separated (except by using a Plan with a *full* factorial treatment structure).
    - + A Plan with a *fractional* factorial treatment structure accepts the limitation on Answers imposed by ('perfect or type 1') confounding to obtain the advantage of using *fewer* resources resulting from a *smaller* number of runs – in this sense, confounding of *focal* variates [introduced by the investigator(s)] in DOE has a *positive* impact.
  - – This is likely the original usage of 'confounding' in statistics – the emphasis in this usage is on the (original) *definition*.
    - + When 'confounding' (in the sense of its definition) is introduced among two or more ***focal*** variates and/or their interactions in a fractional factorial treatment structure, this facet is *not* encompassed by our definition of comparison error. However, 'partial or type 2 confounding' among *non*-focal lurking variates *is* a potential source of *our* comparison error.
- ○ Confounding ('partial or type 2 confounding') in the context of comparative Plans imposes a limitation on an Answer to a Question with a causative aspect, due to one (or more) [*non*-focal] confounders changing (or differing) as the focal variate changes (or differs) in value. The impact of type 2 confounding is *negative* and the emphasis is on the *idea* of confounding and the resulting *limitation* imposed on Answers by comparison error.
  - – We distinguish two cases of type 2 confounding – either may give rise to comparison error that distorts reality (creates illusion) and so leads to a 'wrong' Answer about an $\mathbf{X}$-$\mathbf{Y}$ relationship:
    - + when $\mathbf{Z}$ and $\mathbf{X}$ *both* cause $\mathbf{Y}$ (type 2a) – this situation is that of our introduction to confounding on page 5.30 in Section 10, and the relevant causal structure from near the middle of page 5.34 is case (8) [equivalent to case (1) with the confounder shown explicitly] given again at the right;
    - + when $\mathbf{Z}$ is a *common cause* of $\mathbf{X}$ and $\mathbf{Y}$ (type 2b) – the relevant causal structure is case (9) at the right and see also Appendices 11 and 12 on pages 5.73 to 5.76 and 5.76 and 5.77.

(8)



(9)



(*continued overleaf*)

- The adjective *partial* for type 2 confounding indicates that the association of [the (*un*wanted) change in] the confounder **Z** and (the change in) the focal variate **X** has a correlation that is (usually) *less* than 1 in magnitude;
  - + The special case of *zero* correlation is discussed briefly in relation to diagram (5) near the bottom of page 5.31.
- In the 2004 STAT 231 Course Notes, 'confounding' means our 'partial confounding'.

○ Confounding ('general or type 3 confounding') is a broader meaning used by some statisticians to encompass both the 'partial' confounding of comparative Plans *and* the effects of lurking variates in phenomena like Simpson's Paradox. The impact of type 3 confounding is (again) *negative* and the emphasis is on the *limitation* on Answers and its *consequence*.

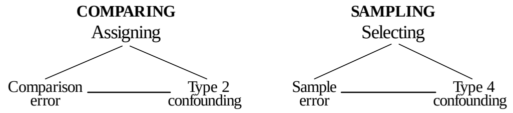
- The adjective *general* is to remind us an Answer [usually to a Question about a (*causal*) **X**-**Y** relationship] may be *altered* in a meaningful (*i.e.*, a practically important) way if the values of **Z** are taken into account.
- When phenomena like Simpson's Paradox are considered to be an instance of ('general or type 3') confounding, discussion of its management (in an observational Plan) in Section 6 on pages 5.68 and 5.69 in Appendix 9 supplements earlier discussion of managing confounding (*e.g.*, as summarized in Table 5.7.10 on page 5.38).
- Simpson's Paradox and related phenomena (discussed in the previous Appendix 9 on pages 5.65 to 5.70) would not usually be considered to involve *causation* in the sense of the discussion of Figure 10.6 of these Course Materials. As a consequence, inclusion of Simpson's Paradox in 'general or type 3 confounding' affects the wording of two definitions:
  - ∗ **Causative aspect:** the Answer from the investigation of a **causative** Question addresses some characteristic(s) of a *relationship* between a response variate and one (or more) explanatory variates; if the relationship is *causal*, the intent is usually that *changing* the value(s) of the explanatory variate(s) would (or will) change the response variate value.
  - ∗ **Focal variate:** an explanatory variate whose *relationship* to the response variate is given in the Answer to the Question.

  If Simpson's Paradox and related phenomena are *not* regarded as instances of 'confounding', a causative aspect and the focal variate would both be defined as involving a *causal* relationship and our distinction involving 'general or type 3 confounding' would not be needed.

○ Confounding ('selecting or type 4 confounding') involves the possible *creation* of an *unwanted* relationship (*e.g.*, by judgement selecting) between unit sample inclusion probabilities and response variate values – see Appendix 14 on pages 5.79 to 5.82. The *relationship* here is between $\mathbf{X}^*$ [which indicates whether a unit *is* selected for the sample ($\mathbf{X}^* = 1$) or is in the group of units *not* selected ($\mathbf{X}^* = 0$)] and **Y**, distinct from the *Question* which may have a descriptive *or* a causative aspect.

- Type 4 confounding is unique to these Course Materials and is included in this Appendix 10 primarily to provide statistical insight from recognizing common themes (from the page 5.48 schema) of probability *assigning* and probability *selecting*;
  - + probability assigning (*e.g.*, EPA) manages type 2 confounding,
  - + probability selecting (*e.g.*, EPS) manages type 4 confounding;

  'manages' here means 'provides a basis for statistical theory that quantifies the likely magnitude of (comparison or sample) error' – this theory shows that both processes are more likely to achieve their goal of acceptable limitation on an Answer with *in*creasing group or sample size(s).

  Type 2 confounding (both cases) *distorts* a (wanted) relationship; type 4 confounding *creates* an *un*wanted relationship.
- The impact of type 4 confounding is (again) *negative* and the emphasis is on the *consequence* (and *limitation*).

NOTES: 77. A further difficulty with 'confounding' is that its root may be used in any of three forms; we can say, for example:

- ● there is *confounding* of the effects of variates **X** and **Z** on variate **Y**, OR:
- ● the effects of variates **X** and **Z** on variate **Y** are *confounded*; ALSO:
- ● if **X** is the focal variate, then **Z** (which is associated with **X**) is a possible *confounder*.

78. The *association* (*e.g.*, non-zero correlation) of confounded variates is really an *incidental* feature of the phenomenon – association in the *usual* state of affairs for variates that change together.

- ● *Zero* correlation of confounded variates is usually introduced by the investigator(s) – for instance, in a factorial treatment structure [see also the brief discussion of diagram (5) immediately before Note 23 on page 5.31].

79. Key ideas to take from the lengthy discussion of confounding in this Appendix 10 and earlier in this Figure 5.7 are:

- ● the use and meaning of 'confounding' in DOE,
- ● the idea and the management of 'confounding' (or of 'lurking variates') in comparative Plans, taking into account the *two* ways a change in a lurking variate can affect attribute value(s):
  - – by *causing* units' response variate (and, hence, their attribute) values to change, AND:
  - – by distorting attribute *calculation* when subdividing is used to manage comparison error in an observational Plan.

For an introductory statistics course, whether Simpson's Paradox and related phenomena are instances of 'confounding' is of no consequence and the concept of 'selecting or type 4 confounding' is *solely* for enrichment.

Surprisingly, 'confounding' may not be mentioned elsewhere in discussion of statistical methods – for instance, it does not appear in the index (p. 500) of the widely-cited text by G.W. Snedecor and W.G. Cochran, *Statistical*

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 27)

NOTES: (cont.) 79. *Methods*, The Iowa State University Press, Ames, Iowa, Seventh Edition, 1980.

- ● No mention of ‘confounding’ may indicate its interpretation in this text as solely our ‘perfect or type 1 confounding’ (the ‘original’ definition), coupled with no *formal* discussion of the topic of DOE.

35. Appendix 11: Reality and Illusion in Statistical Association and Causation (cited on pages 5.30, 5.34, 5.35, 5.71, 5.76 and 5.79)

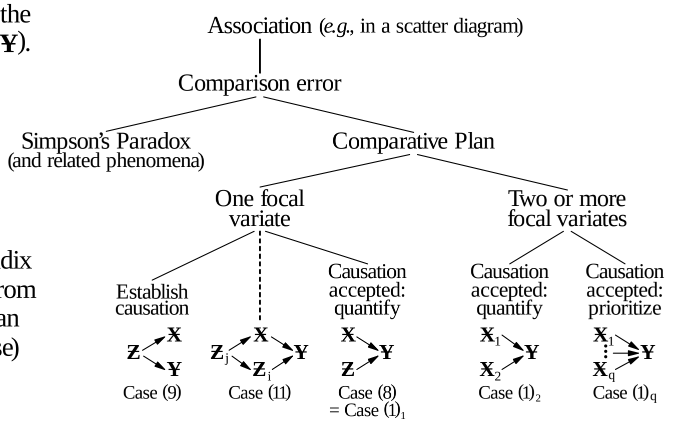
A precept of investigating relationships in statistics is *looking at the data*; that is, *visualizing* the relationship – recall the discussion of scatter diagrams on page 5.29 and note that the course STAT 442 has the title *Data Visualization*. A caveat to this precept is that deciding *how* to visualize the data and how to *interpret* the display(s) must be informed by adequate statistical understanding – seeing may be illusory. In this Appendix 11, we extend the discussion in Section 14 on pages 5.35 and 5.36, dealing in turn with the three main types of Question which arise when investigating statistical relationships:

1. Is there an *association* of **X** and **Y** and, if so, what is its *form* and/or *magnitude* and/or *direction*?
2. What is the *reason* for an association – *e.g.*, can we *establish* that **X** *causes* **Y**?
3. *Accepting* a relationship as causal: what is its *form* – *e.g.*, what is the *effect* of **X** on (the average of) **Y**?
   which explanatory variate is the *most important* cause of (variation in) **Y**?

This ‘natural’ order of such Questions may differ appreciably from the one in which they are discussed in an introductory course; also, the emphasis in introductory discussion is usually on the relationship of *one* focal variate (**X**) to *one* response variate (**Y**).

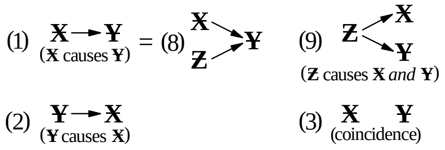
The main issue in our discussion of reality and illusion in data-based investigating of statistical relationships is the limitation imposed on Answers by comparison error arising from confounding by lurking variate(s). A framework for our present discussion is shown in the schema at the right.

- ∗ This schema reminds us of relevant matters from Appendix 8 on page 5.65, Appendix 9 on pages 5.65 to 5.70, and from Sections 9 to 23 on pages 5.28 to 5.52; it starts with an observed association (*e.g.*, of **X** and **Y** in the simplest case) [or with *lack* of (expected) association].


Association (e.g., in a scatter diagram)
Comparison error
Simpson's Paradox (and related phenomena)
Comparative Plan
One focal variate
Two or more focal variates
Establish causation
Causation accepted: quantify
Causation accepted: quantify
Causation accepted: prioritize
Z X Y
Case (9)
Z_j X Z_i Y
Case (11)
X Z Y
Case (8) = Case (1)_1
X_1 X_2 Y
Case (1)_2
X_1 X_q Y
Case (1)_q

We have encountered the first type of Question in Appendix 8 on page 5.65 and in the discussion of Simpson's Paradox and related phenomena (which involve comparing proportions) in Appendix 9 on pages 5.65 to 5.70. Statistical issues are the role of lurking variates and formulating a *clear* Question.

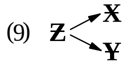
For the second type, we distinguish four reasons (‘cases’) for *association* of variates **X** and **Y** in the presence of a possible confounder (or lurking variate) **Z**; the four relevant cases from page 5.34 are shown symbolically at the right, where an arrow denotes causation:

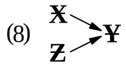

(1) X → Y (X causes Y) = (8) X Z Y
(9) Z X Y (Z causes X and Y)
(2) Y → X (Y causes X)
(3) X Y (coincidence)

- ∗ **X** causes **Y** (in the presence of confounder **Z**);
- ∗ **Y** causes **X**;
- ∗ **Z** causes **X** *and* **Y** – that is, **Z** is a **common cause** of **X** and **Y**;
- ∗ coincidence [which often means that both **X** and **Y** are associated with *time* – *i.e.*, coincidence is often case (9) in which **Z** is time (whatever ‘causation’ by time means)].

A Question about the *actual* reason for an observed **X**-**Y** association can be answered by a process of *elimination*:

- ○ coincidence [case (3)] requires extra-statistical knowledge to rule it out;
  - – in a universe with an almost uncountable number of variates changing in value over time, coincidental associations are likely *numerous* but may largely go *un*noticed;
- ○ identifying correctly which is the response and which the explanatory variate can rule out case (2);
- ○ the remaining two reasons are direct causation [case (1) = case (8)] and common cause [case (9)] – an investigation with a comparative Plan that yields acceptable limitation on the Answer imposed by comparison error (due to type 2b confounding) can then be used to try to rule out case (9) [or case (12)]; if successful, this leaves (direct) causation [case (1)] as the (likely) reason for the association.
  - – The Plan that manages confounding by a common cause for **X** and **Y** [case (9)] will also manage possible confounder(s) **Z** (or **Z**$_i$) changing as focal variate **X** changes [case (8)] in a way that makes *acceptable* the limitation imposed by comparison error (due to type 2a confounding); the common theme of managing type 2 confounding [cases (9) and (8)] is holding **Z** *fixed* as **X** changes in value.
    - + Case (8) is, of course, case (1) with confounder **Z** shown explicitly – **Y** is a **common response** to **X** and **Z**.





For Questions of the second type, comparison error results in an Answer that *mis*identifies the cause of **Y** – for instance:

(*continued overleaf*)

- an Answer which says Ӿ *is* a cause of Ɏ when (in reality) it is not [case (9)],
- an Answer which says Ӿ is *not* a cause of Ɏ when (in reality) it is [case (8)].

Diagrams (4) to (8) below and near the top of the facing page 5.75 illustrate these matters in more detail. A real-world example occurred in *The Globe and Mail* on March 7, 2015, pages M1 and M5, in the article: **$42 AN HOUR** WHY CANADA'S YOUNG ACADEMICS ARE ON THE PICKET LINES.

The excerpt (from page M5) relevant to this discussion is given at the right. The three variates are:

Ӿ: proportion of the Canadian population with a PhD,
Ɏ: Canada's level of productivity and innovation,
Ƶ: Canada's level of development.

The assumption is *that more graduate students equal increased productivity and innovation*. BUT:
It could well be that countries at high levels of development produce both innovation and PhDs.

Thus, there is a need to establish whether the applicable causal structure is our case (1) [= case (8)] or case (9).

In addition to the *causal* Question, there are also statistically challenging matters of *measuring* a country's productivity and innovation (number of patents?) and level of development.

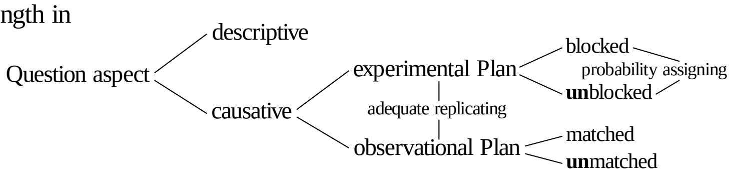
(1) $Ӿ \rightarrow Ɏ$

(8) $$\begin{tikzcd} Ӿ \arrow[dr] & \\ & Ɏ \\ Ƶ \arrow[ur] & \end{tikzcd}$$

(9) $$\begin{tikzcd} & Ӿ \\ Ƶ \arrow[ur] \arrow[dr] & \\ & Ɏ \end{tikzcd}$$

> **Do we have too many PhDs?**
>
> Yet, if graduate programs have grown, it has been because governments have waved money at university administrations, operating under the assumption that more graduate students equal increased productivity and innovation.
>
> In 2011, Ontario announced that it would fund an additional 6,000 master's and doctoral spots even as its undergraduate per-student funding, adjusted for inflation, has been declining for decades.
>
> Turns out that the current instructions for building a PhD student may not translate into the economic gains that governments assume.
>
> "It is an open question whether PhDs are necessary to produce innovation and productivity, or can you do that with MAs," said Daniel Munro, an analyst at the Conference Board of Canada who will be releasing a report he co-authored on PhDs in the Canadian labour market this spring.
>
> Dr. Munro says it is true that countries with a higher percentage of PhDs have higher level of productivity and innovation, and tend to produce more patents. However, the relationship between very advanced education and innovation is uncertain. It could well be that countries at high levels of development produce both innovation and PhDs.

For the third type of Question, provided that the Ӿ-Ɏ relationship *is* causal [case (1) above], acceptable limitation imposed by comparison error (due to type 2a confounding) can usually be achieved by a comparative experimental (but *not* an observational) Plan developed and executed as previously discussed at length in this Figure. The schema at the right, from the upper right of page 5.37, reminds us of Plan components available to investigators to manage possible confounding by known and by unknown and unmeasured *non*-focal explanatory variates.



For Questions of the second and third types, which involve an inference from association to causation, the limitation imposed by comparison error (due to type 2 confounding) on Answers can be summarized in two precepts:

∗ **Partial or complete false positive:** when *coincidence* can be ruled out as a reason, an Ӿ-Ɏ *association* indicates *causation* of Ɏ but ***not necessarily*** (only or even in part) by Ӿ.

∗ **False negative:** association of Ӿ and Ɏ may be *absent* even when they have a *causal* relationship.

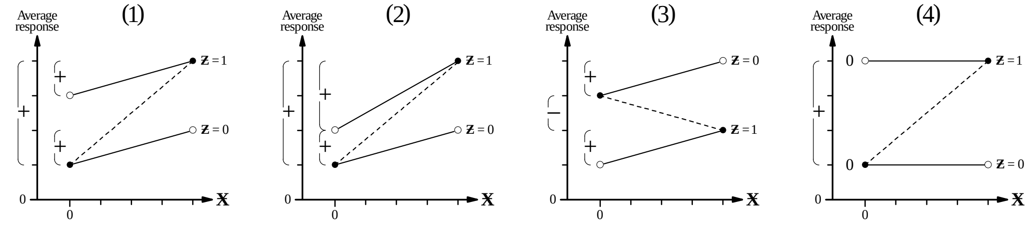
These precepts formalize the ideas in Note 30 on page 5.35. The precepts are illustrated in more detail in two sets of four diagrams below and on the facing page 5.75, which show how the *average* of Ɏ changes with Ӿ in the presence of a (binary) confounder Ƶ; except for diagram (4) [case (9) above], the causal connections are those in case (8). In these eight diagrams:

- the circles (○) represent response variate averages that do *not* occur under the Plan because these averages would require the units' Ƶ values to be *different* from their actual values (counterfactuals – recall the middle of page 5.50).
- the dots (•) represent response variate averages for units whose Ƶ values mean these averages *are* observed (data),
- the solid lines are the *un*observed 'reality,' ● the dashed lines are the observed 'illusion';
- the horizontal axis has a *quantitative* scale, rather than showing Ӿ as an indicator variate with values of 0 and 1.

○ In the left-hand diagram (1) below, the (positive) effect of Ӿ on (the average of) Ɏ is represented by the two braces to the *right* of the Ɏ-axis; however, if there is confounding (Ƶ *changing* from 0 to 1 as Ӿ changes), the effect of Ӿ on Ɏ would be *observed* as the brace to the *left* of the axis. Thus, in diagram (1):

- the dashed line involving comparison error yields a *wrong (exaggerated) magnitude* for the effect of Ӿ on Ɏ.
- the slope of the solid lines showing the relationship of Ӿ to the average of Ɏ are *un*affected by the value of Ƶ – that is, there is no *interaction* of Ӿ and Ƶ.

○ Diagram (2) is like diagram (1) except there *is* interaction of Ӿ and Ƶ – the solid lines have *different* slopes.



(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 28)

- Diagrams (1) and (2) illustrate partial false positives; association of **X** and **Y** *does* correspond to causation of **Y** by **X** but confounding by **Z** distorts reality (creates illusion).

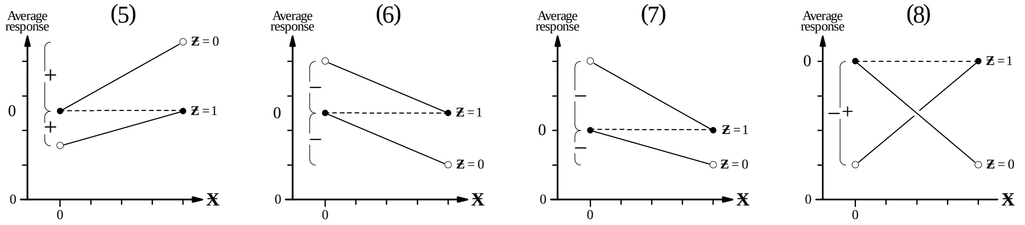
○ In diagram (3) with a *different* (negative) **Z**-**Y** relationship, comparison error is *wrong direction* for the effect of **X** on **Y**.

- An Answer which, due to comparison error, is a wrong *direction* for an **X**-**Y** relationship is also a case of wrong *value* (and, usually, wrong *magnitude*) but, as in Simpson's Paradox (see Appendix 9 on pages 5.65 to 5.70), the more dramatic (directional) manifestation of comparison error is usually emphasized.

○ The right-hand diagram (4) illustrates (with*out* interaction) a (complete) *false* positive Answer – there is **X**-**Y** *association* without **X**-**Y** *causation*; this is the situation when **Z** is a common cause of **X** and **Y** [case (9) at the lower right of page 5.73].

- The two lines in diagram (4) have *zero* slope so we say **X** and **Y** are *independent conditional on* **Z** – for (unit) change in **X**, there is *no* change in (the average of) **Y** *provided* **Z** is (held) fixed – see also Notes 87 and 88 on page 5.79 in Appendix 13.

⦿ A false negative Answer, as in diagram (5), can occur in other ways; in diagram (6), such an Answer again arises from confounding with*out* interaction, and in diagrams (7) and (8) [as in diagram (5)] from *both* confounding *and* interaction.



- In diagrams (2), (5) and (7), interaction of **X** and **Z** is *incidental* to the manifestation of comparison error when quantifying the magnitude and/or direction of the effect of **X** on **Y**; only in diagram (8) is the interaction *essential* to this manifestation.
  + (Complete) false negatives involve the special case of *exact* cancellation of the effects of **X** and **Z** on **Y**.

The foregoing discussion of diagrams (1) to (8), involving type 2 confounding, is summarized in Table 5.7.40 at the right.

Table 5.7.40

X-Average response RELATIONSHIPS IN EIGHT DIAGRAMS

Diagram	Relationship	Interaction	Comparison error
(1)	Positive	None	Wrong magnitude – exaggerated
(2)	Positive	Incidental	Wrong magnitude – exaggerated
(3)	Positive	None	Wrong direction: wrong value
(4)	None	None	Wrong value: false positive
(5)	Positive	Incidental	Wrong value: false negative
(6)	Negative	None	Wrong value: false negative
(7)	Negative	Incidental	Wrong value: false negative
(8)	Mixed	Essential	Wrong value: false negative

NOTES: 80. For Questions of the third type, concerned with quantifying the effect of **X** on **Y**, the eight **X**-Average response diagrams on the facing page 5.74 and above are only *illustrative* of comparison error because its particular manifestation depends on the interplay of several matters which affect the appearance of any such diagram, including:

- the magnitude(s) and direction(s) of the slope(s) of the **X**-**Y** relationships;
- the magnitude of the **Z**-**Y** relationship, reflected in the *vertical separation* of the solid lines;
- the absence or presence of interaction (of **X** and **Z** in their effects on **Y**);
- the *forms* of the **X**-**Y** and **Z**-**Y** relationships (*e.g.*, linear or *non*linear).

81. The *two* descriptions in the last column of Table 5.7.40 above for diagrams (4) to (8) on the facing page 5.74 and above remind us that, for quantitative variates, comparison error is [despite the second (more dramatic) descriptions] a *wrong value* (and, usually, a wrong *magnitude*) for the effect of **X** on (the average of) **Y**. As a consequence, diagrams (4) to (8) illustrate the occurrence of comparison error for Questions of *both* the second and third type.
   - Similarly, Simpson's Paradox (discussed in Appendix 9 on pages 5.65 to 5.70) is, under subdivision on an additional explanatory variate, a *change in value* for a 'treatment' effect, resulting in a (dramatic) change of direction of the 'effect' [analogous to diagram (3) at the bottom of the facing page 5.74.]

82. For the (unattainable) ideal [criterion (1) near the top of page 5.32] of *all* non-focal explanatory variates remaining fixed when **X** changes to make apparent its relationship to **Y**, the Answer has *no* limitation imposed by comparison error but, when a possible confounder **Z** does *not* remain fixed, comparison error *does* impose a limitation on the Answer about the **X**-**Y** relationship. These limitations, from the foregoing discussion in this Appendix 11, are summarized in Table 5.7.41 below; it reminds us that the nature of the limitation depends on the Question context.

Table 5.7.41

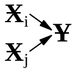
Question context	Question	Comparison error	Confounding type	Other name
Comparing proportions	Does **Y** increase or decrease with **X**?	Wrong direction for an **X**-**Y** association	3	Simpson's Paradox
Establishing causation	Is **X** the (or a) *cause* of **Y**?	Wrong cause identified for **Y**	2b	Common cause
Quantifying a treatment effect	What is the effect of **X** on **Y**?	Wrong magnitude for effect of **X** on **Y**	2a	Confounding
		Wrong direction for effect of **X** on **Y**	2a	Confounding
Improving a process	Is **X** the *most important* cause of **Y**?	Wrong main cause identified for **Y**	2a	Confounding
More than one focal variate	What is the effect of each $\mathbf{X}_i$ on **Y**?	Wrong effect of one or more $\mathbf{X}_i$ on **Y**	1, 2a	-----

(*continued overleaf*)

NOTES: (cont.) 82. ● If Simpson's Paradox (in the first line of Table 5.7.41) is observed in a *sample* but the Question involves the corresponding *population*, limitation is imposed on the Answer by *sample* error as well as by *comparison* error.

– Regardless of whether the investigation involves a (respondent) population census or sample, study, non-response and measurement error also likely need to be managed in the Plan – see Appendix 16 on page 5.84.

● The second-last line of Table 5.7.41 deals with **process improvement** – prioritizing the explanatory variates responsible for variation in **Y**; the relevant causal connections can then be thought of as case $(1)_2$ [repeated at the right from page 5.36]. Prioritizing a set of m explanatory variates $\mathbf{X}_l$ (l = 1, 2, ..., m) can be achieved by successively prioritizing *pairs* of variates $\mathbf{X}_i$ and $\mathbf{X}_j$. The effects of confounding are the *same* as for *quantifying* the effect of **X** on **Y** and the manifestation of comparison error is identifying the **wrong *main* cause** of (variation in) **Y**.

$(1)_2$ 

– Because prioritizing explanatory variates involves *two or more* focal variates, the Plan should make provision for estimating *interaction* effect(s).

+ ['Perfect or type 1 confounding' among some treatment effects arising from using a *fractional* factorial treatment structure may be acceptable in the Question context – recall Note 47 on page 5.45 and the discussion of type 1 confounding on page 5.71 in Appendix 10.]

● In the *last* line of Table 5.7.41, again involving estimating treatment effects for two or more focal variates, the Plan should provide for estimating both main *and* interaction effects.

36. Appendix 12: Connections Among Three Variates (cited on pages 5.35, 5.44, 5.71 and 5.79 in Appendix 13)

For *three* variates [two expanatory (**X** and **Z**) and one response (**Y**)] involving *two* causal relationships, there are *five* causal structures, as shown in the first two columns of Table 5.7.42 at the right below; the first structure has *two* contexts, making six lines in the Table. The structures are five of the twelve cases given near the middle of page 5.34 in Section 13 [cases (4), (6), (8), (9) and (10)] plus case $(1)_2$ from the upper right of page 5.36 [also discussed above in the second bullet (●) of Note 82]. [A reminder of two definitions from pages 5.30 and 5.44 is:

∗ **Confounding:** differing distributions of values of one or more *non*-focal explanatory variate(s) among two (or more) groups of units [like (sub)populations or samples] with different values of the focal variate.

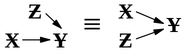
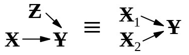
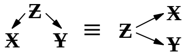
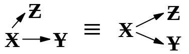
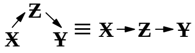
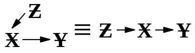
∗ **Interaction** of two factors $\mathbf{X}_1$ and $\mathbf{X}_2$ is said to occur when the effect of one factor on a response variate **Y** depends on the level of the other factor. Interaction means the combined effect of two factors is *not* the sum of their individual effects.]

Several matters are noteworthy.

○ What tends to distinguish the cases is the pattern of the *causal* relationships in the last column of Table 5.7.42 – as shown in the *third* column, each variate is *associated* with each of the other two, except in the case of interaction, when the **X**-**Z** relationship is not relevant.

○ *Confounding* and *interaction* have similarities (+) and differences (–).

+ both involve two explanatory variates which cause a response variate;

Variate *causal* connections		Table 5.7.42 Name	Association X-Y	X-Z	Z-Y	Causation X-Y	X-Z	Z-Y
(8)		Confounding [Common response **Y**]	Yes	Yes	Yes	Yes	No	Yes
$(1)_2$		Interaction	Yes	---	Yes	Yes	No	Yes
(9)		Common cause **Z**	Yes	Yes	Yes	No	Yes	Yes
(10)		Common cause **X**	Yes	Yes	Yes	Yes	Yes	No
(6)		Causal chain **XZY**	Yes	Yes	Yes	Yes	Yes	Yes
(4)		Causal chain **ZXY**	Yes	Yes	Yes	Yes	Yes	Yes

+ both have the same pattern of causal relationships and (except as noted above) associations in Table 5.7.42;

– their focus is different:

⊙ confounding is concerned with the impact on investigating the **X**-**Y** relationship of (unwanted) *changes* in (confounder) **Z** as (focal variate) **X** changes;

⊙ interaction is concerned with the impact on the $\mathbf{X}_1$-**Y** relationship *and* the $\mathbf{X}_2$-**Y** relationship of the *value* of (focal variates) $\mathbf{X}_2$ and $\mathbf{X}_1$ respectively (see also Note 88 on page 5.79 of Appendix 13 – recall also Note 48 on page 5.45).

The same components but different focus of confounding and interaction are somewhat reminiscent of the conditioning-ignoring distinction discussed in Note 74 of Appendix 8 on page 5.65 because, in each case, statistical *mis*handling provides opportunity for comparison error to impose unnecessary (and so, possibly unacceptable) limitation on Answers.

The *manifestation* of confounding as comparison error may be (adversely) affected *if* there is *interaction* of **Z** and **X** – for example, (puzzling) 'inconsistencies' may be exhibited in the **X**-**Y** relationship – recall Table 5.7.40 and the discussion on pages 5.74 and 5.75 in Appendix 11.

○ The possibility of **Z** as a **common cause** of **X** and **Y** [case (9)] is relevant when establishing the reason for an **X**-**Y** association, as discussed on pages 5.73 and 5.74 in Appendix 11.

– Common cause **Z** responsible for *mis*identifying **X** as a cause of **Y** is our type 2b confounding – recall the discussion at the bottom of page 5.71 in Appendix 10 and at the bottom of page 5.73 and the top half of page 5.74 in Appendix 11.

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 29)

+ From *this* perspective, **Z** as a common cause of **X** and **Y** could be regarded as an extreme case of our type 2a confounding where **Z** is *solely* responsible for the change in **Y** as **X** changes.

- ○ **X** as a **common cause** of **Z** and **Y** [case (10)] is really the causal structure at the right below, because **Z** is an *explanatory* variate; thus, elsewhere we consider case (10) to involve *three* causal relationships, *not* two as in this Appendix 12. We have seen earlier in this Figure 5.7 that case (10) with *three* causal relationships:

$$(10)\quad \begin{tikzcd} & \mathbf{Z} \arrow[dr] & \\ \mathbf{X} \arrow[ur] \arrow[rr] & & \mathbf{Y} \end{tikzcd}$$

  - – is *not* a viable basis for a comparative Plan, as discussed in Note 41 on pages 5.42 and 5.43;
  - – can result in *biased* estimating of a treatment effect, as illustrated on page 5.46 in the discussion of Table 5.7.16.
- ○ As discussed on page 5.32 in Note 24 (and also in the middle of the first side of Figure 10.6 of the Couse Materials), we think of causation of **Y** by **X** as proceeding via a (long) **causal chain** of explanatory variates leading to the response of interest. The Question context identifies (arbitrarily) *one* (focal) variate (**X**) in this chain as being of interest, but we recognize that this variate is *preceded* by and followed by other 'focal' variates; the context also (arbitrarily) defines the *end* of the chain in terms of a particular *response* variate (**Y**). However, this response can become part of an *explanatory* variate chain if a different Question context identifes a *different* (later) response variate. From this perspective:
  - – The causal chain of case (6) is merely the upper branch of the (real) causal structure of case (10) shown above at the right;
    - ⊙ case (6) reminds us to distinguish **X** causing **Y** *via* **Z** from **X** and **Z** as *separate* causes of **Y** [case (8)].
  - – The causal chain of case (4) is really case (1) [= case (8)] – **Z** in case (4) is merely an explanatory variate *preceding* the focal variate **X** in the causal chain and so is (generally) of no statistical interest in the Question context.

The (surprising) number of statistical issues arising with relationships among *only three* variates is further complicated if the **X**-**Z**-**Y** relationship is modelled mathematically; such a model (for use in the Analysis stage of the PPDAC cycle) needs to consider:

- ● the *form* in the model (*e.g.*, first power, second power, square root, logarithm, product) of **X** and **Z**;
- ● the *distribution* of **Z** (*e.g.*, its mean and standard deviation) [and perhaps of **X**];
- ● the *relationship* of **X** and **Z** (*e.g.*, their correlation).

Association of (focal variate) **X** and (confounder) **Z** in case (8) is one feature of confounding, a source of comparison error and limitation on Answers from comparative Plans (recall Note 78 on page 5.72). Similarly, association among variates in the structural component (on the right-hand side) of a response model [like equation (5.7.3) on page 5.28] is also a source of such limitation, manifested as *uncertainty* in the estimates of model parameters [*e.g.*, β_1 – the treatment effect for **X** – in equation (5.7.3)].

- ○ This uncertainty becomes apparent from *stepwise* model fitting, a process to assess [*e.g.*, based on the coefficient of multiple determination, a measure of the proportion of the variation in **Y** accounted for by the fitted model] *which* explanatory variates to include in the model. For instance, in the case of two (focal) variates **X**$_1$ and **X**$_2$, three models are fitted – one with *both* variates, one with **X**$_1$ only and one with **X**$_2$ only. The stronger the association (in the data) of **X**$_1$ and **X**$_2$, the greater the likely difference in the estimates of their coefficients β_1 and β_2 among the three models.
  - – In the extreme situation where two variates **X**$_i$ and **X**$_j$ have correlation of magnitude 1 (*i.e.*, **X**$_i$ and **X**$_j$ are the *same* variate statistically), the model fitting process cannot be achieved computationally – the design matrix is not of full rank and so cannot be inverted.

In introductory statistics courses, emphasis on comparative Plans with *one* focal variate, together with similarities of confounding and interaction when there are three variates, should not be allowed to obscure the continuing importance of possible confounding in comparative Plans with *two or more* focal variates. With *three* variates and possible confounders **Z**$_i$, **Z**$_j$ and **Z**$_k$, statistical issues like those in the foregoing discussion may arise for connections among:

○ **X**$_1$, **Z**$_i$ and **Y**, ○ **X**$_2$, **Z**$_j$ and **Y**, AND ○ **X**$_1$, **X**$_2$, **Z**$_k$ and **Y** [**Z**$_k$ may be a *common cause* of **X**$_1$, **X**$_2$ *and* **Y**].

37. Appendix 13: Simpson's Paradox and Interaction (cited on pages 5.44, 5.60, 5.67, 5.75 and 5.76 in Appendices 9 and 12)

For extending the discussion of Simpson's Paradox in Appendix 9 on pages 5.65 to 5.70, for convenience in this Appendix 13 (including labelling the three diagrams to the right of Tables 5.7.43 to 5.7.45 overleaf) we use the notation defined on page 5.66:

X$_1$ is an applicant's sex (female, male), **X**$_2$ is the school applied to (Law, Business), **X**$_3$ is the level of study [Masters, Doctoral],
Y is the response to an applicant (admitted, not admitted) or time for degree completion (minimum, longer),
$\overline{\mathbf{Y}}$ [the *average* of (**Y**)] is the *percentage* of applicants admitted or who complete their degree in the minimum time.

The diagrams illustrating Simpson's Paradox to the right of Tables 5.7.21 to 5.7.26 (on pages 5.66 and 5.67 in Appendix 9) are reminiscent of a diagram showing interaction (*e.g.*, in Note 48 on page 5.45); however, there are *differences*:

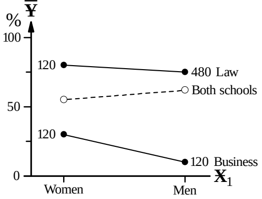
- ○ the Simpson's Paradox diagrams have an additional (dashed) line for the overall **X**$_1$-$\overline{\mathbf{Y}}$ relationship;
- ○ the instances of Simpson's Paradox in Tables 5.7.21 to 5.7.26 have only *parallel* (solid) lines for the **X**$_1$-$\overline{\mathbf{Y}}$ relationships for different values of **X**$_2$ – that is, there is *no* interaction of **X**$_1$ and **X**$_2$ in their effects on **Y**.

This restriction is *removed* in (another) reworking of Table 5.7.21 in Table 5.7.43 overleaf at the top of page 5.78, where there *is* interaction of **X**$_1$ and **X**$_2$ in their effects on **Y** because the two solid lines in the diagram to the right of the Table are *not* parallel.

(*continued overleaf*)

Table 5.7.43:

	...........WOMEN............			...............MEN..............			.......BOTH SEXES......		
SCHOOL	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%
Law	120	96	**80**	480	360	**75**	600	456	**76**
Business	120	36	**30**	120	12	**10**	240	48	**20**
Both schools	240	132	**55**	600	372	**62**			



Thus, interaction *may* be involved in Simpson's Paradox but is not *required* for it to occur.

Discussion in Appendix 9 at the upper left of page 5.67 and on page 5.68 in Section 5, and in this Appendix 13, reminds us that Simpson's Paradox and interaction *both* involve (estimated) values of *conditional* probabilities for $\mathbf{Y}$, BUT:

- ○ Simpson's Paradox involves comparing these probabilities conditioned on two (or three) of the $\mathbf{X}$s with probabilities conditioned on one *fewer* (one or two) $\mathbf{X}$s; WHEREAS:
- ○ interaction is absent or present depending on the values of probabilities with the *same* conditioning on the $\mathbf{X}$s – these values determine whether the corresponding lines are or are not parallel.

NOTES: 83. Illustration of Simpson's Paradox from comparing *across* Tables 5.7.21 to 5.7.26 can overshadow comparisons *down* such tables. For example, in Table 5.7.21 (reworked as Table 5.7.29 on page 5.68), the six **bold** percentages for $\mathbf{X}_2$ (80 and 20, 75 and 15, $76.\dot{6}$ and 17.5) address a Question *different* from possible sex discrimination:

- ● *How do the admission standards of the Law and Business schools compare?*

The (hypothetical) data in Table 5.7.29 (and Tables 5.7.43 to 5.7.45 on this page) indicate an appreciably *higher* admission standard for Business than for Law, unless the abilities of the two applicant pools are remarkably different.

84. A second reworking of Table 5.7.21 and its diagram is given below in Table 5.7.44; as in Table 5.7.43, there *is* interteraction of $\mathbf{X}_1$ and $\mathbf{X}_2$ in their effects on $\mathbf{Y}$ but, for both schools combined, there is *no* sex difference in proportions, due to cancellation of effects in *opposite* directions for the schools individually.

Table 5.7.44:

	...........WOMEN............			...............MEN..............			.......BOTH SEXES......		
SCHOOL	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%
Law	120	96	**80**	180	153	**85**	300	249	**83**
Business	120	24	**20**	180	27	**15**	300	51	**17**
Both schools	240	120	**50**	360	180	**50**			

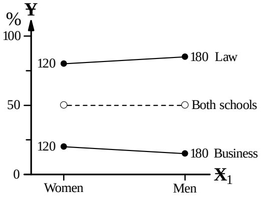
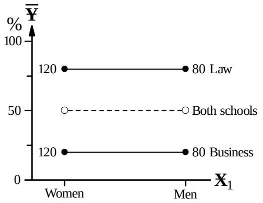


Table 5.7.45 below and its diagram show, like Table 5.7.44, *no* sex difference for both schools combined but this is now a consequence of the *individual schools* also showing this same behaviour – there is *no* interaction.

Table 5.7.45:

	...........WOMEN............			...............MEN..............			.......BOTH SEXES......		
SCHOOL	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%	Number of Applicants	ADMISSIONS Number	%
Law	120	96	**80**	80	64	**80**	200	160	**80**
Business	120	24	**20**	80	16	**20**	200	40	**20**
Both schools	240	120	**50**	160	80	**50**			



85. Across Tables 5.7.21 to 5.7.26 on pages 5.66 and 5.67 in Appendix 9 and Tables 5.7.43 to 5.7.45 in this Appendix 13, different weights in the proportion calculations (like those in Table 5.7.28 on page 5.68) yield a noteworthy *variety* in the percentages for women compared to those for men. This is summarized in Table 5.7.46 at the right below; three categories are distinguished.

- ○ In four tables, there *is* an $\mathbf{X}_1$-$\overline{\mathbf{Y}}$ relationship, there is no interaction of $\mathbf{X}_1$ and $\mathbf{X}_2$ in their effects on $\mathbf{Y}$ and, in two of the tables, the $\mathbf{X}_1$-$\overline{\mathbf{Y}}$ relationship is *un*exceptional in light of the effect of subdivision by $\mathbf{X}_2$; by contrast, in Table 5.7.23 and Table 5.7.26 between the first and second levels of subdivision by $\mathbf{X}_2$, the exceptional behaviour is the $\mathbf{X}_1$-$\overline{\mathbf{Y}}$ relationship *disappearing* when the data are subdivided by $\mathbf{X}_2$.
  - – Notation like (1,3) [or (1,2)] on Table 5.7.25 (or Table 5.7.26) in Table 5.7.46 refers to the first and third (or first and second) levels of subdivision by $\mathbf{X}_2$.
- ○ In four tables, there is *no* $\mathbf{X}_1$-$\overline{\mathbf{Y}}$ relationship but three of these are '*false* negative' Answers – when the data are subdivided by $\mathbf{X}_2$, there *is* an $\mathbf{X}_1$-$\overline{\mathbf{Y}}$ relationship and so they are designated 'exceptional' in the fourth column.
  - – In Table 5.7.44, interaction is the *reason* for the exceptional behaviour but interaction is absent in the other three tables.
- ○ In five tables, there *is* (again) an $\mathbf{X}_1$-$\overline{\mathbf{Y}}$ relationship, inter-

Table 5.7.46: $\mathbf{X}_1$-$\overline{\mathbf{Y}}$ **RELATIONSHIPS IN NINE TABLES**
(SP in the fourth column denotes 'Simpson's Paradox')

Table	Relationship	Interaction	Exceptional behaviour
5.7.23	Yes	No	Yes
5.7.24	Yes	No	No
5.7.25 (1,3)	Yes	No	No
5.7.26 (1,2)	Yes	No	Yes
5.7.22	No	No	Yes
5.7.26 (2,3)	No	No	Yes
5.7.44	No	Essential	Yes
5.7.45	No	No	No
5.7.21	Yes	No	Yes: SP
5.7.25 (1,2)	Yes	No	Yes: SP
5.7.25 (2,3)	Yes	No	Yes: SP
5.7.26 (1,3)	Yes	No	Yes: SP
5.7.43	Yes	Incidental	Yes: SP

(*continued*)

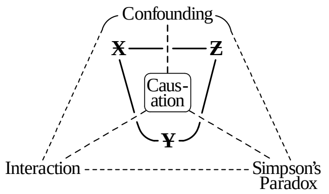
Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 30)

NOTES: (cont.) 85. action is absent or incidental, and each is a case of the exceptional behaviour known as Simpson's Paradox.

Apart from understanding the properties of proportions and weighted averages and using an adequate Plan (discussed in Sections 3, 4 and 6 and Note 75 on pages 5.67 to 5.69 in Appendix 9), the summary in Table 5.7.46 at the bottom right of the facing page 5.78 reminds us that:

- Simpson's Paradox is merely the *most* exceptional case (*change of direction* of an X_1-Y relationship) in a context (involving *discrete* variates) that can give rise to less exceptional or even *un*exceptional behaviour;
- interaction is rarely the reason for the exceptional behaviour (only in Table 5.7.44 on the facing page 5.78).

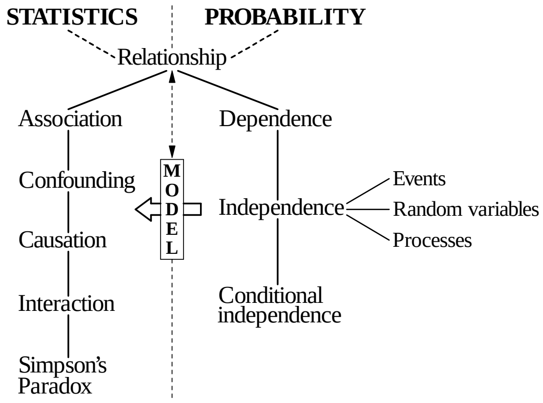
86. In meeting the obligation to deal with *relationships* in an introductory statistics course, the lengthy discussion (*e.g.*, in this Figure, Sections 10 to 23 on pages 5.29 to 5.52 and Appendices 8 to 13 on pages 5.65 to 5.79) shows the (unexpected) complexities, for only *three* variates, arising from issues of causation, confounding, interaction and Simpson's Paradox. The schema at the right reminds us there are common themes *and* differences among these four matters – recall also Appendix 12 on pages 5.76 and 5.77.



87. As summarized on the left of the schema at the right, a relationship in *statistics* is often considered in terms of one or more of association, confounding, causation, interaction and Simpson's Paradox (recall also the schema on page 5.29).

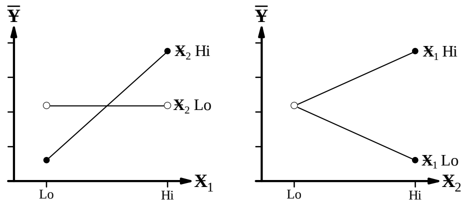
In *probability* (on the right of the schema), a relationship is considered in terms of *dependence*, which comes in great variety and is often difficult to mathematize; as a consequence, introductory courses emphasize ***in***dependence, as it applies to events, random variables and processes. Even the first two of these three involve an appreciable set of ideas and may be all a course has time to discuss.

Connection between statistical and probabilistic considerations of a relationship arises in the probability *models* statistics uses in the Analysis stage of the PPDAC cycle.



- Emphasis on *in*dependence in introductory courses can obscure the fact that independence is a mathematical *idealization*. In the real world, *de*pendence is the norm – it may be that the behaviour of *every* particle in the universe depends on (*i.e.*, is affected by) *every other* particle, no matter how minute the degree of dependence.
  - This may be why lurking variates are usually so *numerous* when answering Questions with a causative aspect.

88. The (equivalent) diagrams at the right show the effects of two (binary) focal variates X_1 and X_2 on (the average of) Y; one focal variate is on the horizontal axis of a diagram, the other distinguishes the two lines by its level.



- The *non*parallel lines show there is an X_1-X_2 *interaction*.
- The left-hand diagram shows that X_1 and Y are *conditionally independent* when X_2 is Lo (the relevant line has *zero* slope) but *not* when X_2 is Hi.
- The right-hand diagram shows there is *no* conditional independence of X_2 and Y – *neither* line has zero slope [recall also diagram (4) at the bottom of page 5.74 and its discussion near the top of page 5.75].

Thus, equivalences between statistical and probabilistic views of relationships are not always straight forward.

38. Appendix 14: Question Aspect and Method of Sample Selecting (cited on pages 5.23, 5.40, 5.44, 5.56 and 5.72)

The Question aspect (identified as early as the Problem stage of the PPDAC cycle) has implications for the method of sample *selecting*. To discuss this matter, we first adapt ideas from the fourth bullet (●) of Note 53 on page 5.49 (immediately before Section 22). We take (binary) focal variate X^* to indicate whether a unit *is* selected for the sample ($X^* = 1$) or is in the group of units *not* selected ($X^* = 0$). The value of a (possibly confounding) explanatory variate Z determines which X^* value each respondent population unit receives. [The asterisk (*) on X is to remind us that the nature of this focal variate differs from X elsewhere in this Figure 5.7; for instance, its values are *imposed* on the units of the respondent population but, *un*like a 'treatment,' it (usually) does not actively change a unit's response variate value (but see Note 93 on page 5.82)]

* **Question with a descriptive aspect:** under *probability* selecting (*e.g.*, EPS), a suitable probabilistic process (*e.g.*, equiprobable digits) determines the values of Z (and, hence, of X^*), so these values are *un*influenced by the units' *other* variate values; with *adequate replicating*, we can therefore usually come acceptably close to the ideal of there being *no* Z-Y (and,

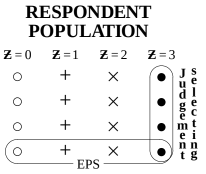
(*continued overleaf*)

hence, no $\mathbf{X}^*$-$\mathbf{Y}$) relationship over the units of the respondent population. This means in practice that the value of the attribute of interest in the sample will usually be acceptably close to that for the units *not* selected, which means in turn an acceptable limitation on the Answer due to sample error. Schema (A) at the right shows the relationships (or their *absence*) among $\mathbf{Z}$, $\mathbf{X}^*$ and $\mathbf{Y}$.

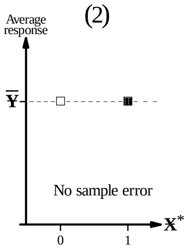
(A) $$\begin{tikzcd} & \mathbf{X}^* \\ \mathbf{Z} \arrow[ur] & \\ & \mathbf{Y} \end{tikzcd}$$

- ● While probability selecting *may* obtain a sample with an attribute value (*e.g.*, an average) meaningfully different from that of the respondent population (*e.g.*, $\overline{\mathbf{Y}}$), statistical theory quantifies, under repetition, the probability of obtaining such a sample – that is, the theory makes explicit the dependence (and its form) of sampling imprecision on degree of replicating (*i.e.*, sample size), as well as providing, when estimating an *average*:
  - + a *confidence interval* expression, + *unbiased* estimating.
  - – Diagram (1) at the right is a representation of a respondent population of $\mathbf{N}=16$ units with four different $\mathbf{Z}$ values; a sample consisting of the bottom row of four units would yield (the ideal of) diagram (2), in which the sample average (■) and that of the units *not* selected (□) are (exactly) *equal* – that is, there is *zero* sample error.

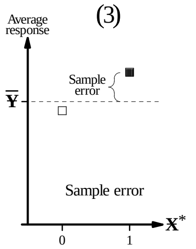
(1)



(2)



(3)



∗ **Question with a descriptive aspect:** under *judgement* selecting, $\mathbf{Z}$ may be an explanatory variate of the respondent population units, in which case a unit's $\mathbf{Z}$ value influences *both* its $\mathbf{X}^*$ and $\mathbf{Y}$ values so that, as shown in schema (B) at the right, $\mathbf{X}^*$ is *associated* with $\mathbf{Y}$ (the *dashed* line), due to their *common cause* (or 'confounder') $\mathbf{Z}$; an example, in the respondent population in diagram (1) above, would be if judgement selecting obtained the four units with $\mathbf{Z} = 3$.

(B) $$\begin{tikzcd} & \mathbf{X}^* \arrow[d, dash, dashed] \\ \mathbf{Z} \arrow[ur] \arrow[dr] & \\ & \mathbf{Y} \end{tikzcd}$$

- ● A possible outcome of judgement selecting is illustrated in diagram (3) at the right above – this diagram [and diagrams (6), (7) and (9) on the facing page 5.81] assume the sample size is one-quarter of the respondent population size and sample error is *positive*.
  - – As well as illustrating the (*un*acceptable) limitation imposed by sample error under judgement selecting when answering a Question with a descriptive aspect, diagram (3) *also* reminds us of the usual (*non*-ideal because there *is* sample error) situation under *probability* selecting; the *critical* differences are:
    - + judgement selecting does *not* have the three benefits from sampling theory under EPS, reiterated above (see also Note 10 on page 5.23 and the schema in Note 53 at the lower right of page 5.48), which allow investigators to manage the inherent uncertainty (arising from incomplete information) of sampling and so to try to make acceptable in the Question context the limitation on an Answer imposed by sample error;
      - ○ when estimating an average under EPS, a consequence of the Central Limit Theorem is a *higher* probability of selecting a sample with sample error of *smaller* magnitude, a *lower* probability of selecting one with *larger* magnitude;
        - ⊙ this may imply that *judgement* selecting, to which the Central Limit Theorem does *not* apply, is prone to sample error of *larger* magnitude than is EPS for a given sample size.

  Of course, it is *possible* that, in diagram (1) above, EPS might select the four units with $\mathbf{Z} = 3$ and judgement selecting might select the bottom row of four units – recall also Note 51 on pages 5.47.

When answering a Question with a descriptive aspect and when the value of the respondent population attribute being estimated subsequently becomes *known*, statistical experience shows that 'confounding' by $\mathbf{Z}$, resulting in a sample error of unacceptably large magnitude, is *common* under judgement selecting compared with probability selecting.

- ● For a sample obtained by judgement selecting, the limitation imposed by sample error on an Answer to a Question with a descriptive aspect is so severe it raises doubt as to whether the investigation should have been undertaken.
  - – Judgement (rather than probability) selecting, usually done to conserve resources, is thus statistical *false* economy when answering a Question with a descriptive aspect.

∗ **Question with a causative aspect answered using an experimental Plan:** for a Question with a causative aspect, EPS, despite its benefits, is seldom feasible and judgement selecting is the practical alternative (recall the oat-bran investigation described in Note 55 on pages 5.51 and 5.52). We deal with this Question aspect and Plan type only for the special case of estimating a treatment effect of focal variate $\mathbf{X}$ which is a *difference* of two *averages* (${}_{\mathbf{X}=1}\overline{\mathbf{Y}} - {}_{\mathbf{X}=0}\overline{\mathbf{Y}}$) and so is itself an average; also, we assume that the sample is divided into the treatment ($\mathbf{X}=1$) and control ($\mathbf{X}=0$) groups by EPA. From discussion in Section 23 near the middle of page 5.50, we recall from schema E that *comparison* error has *two* sources:

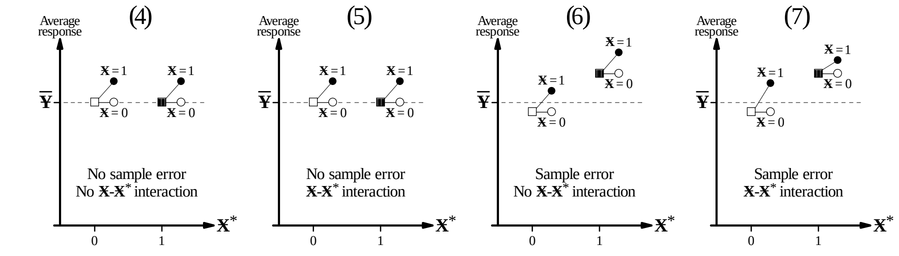
- ○ the two half samples obtained under EPA would likely have *different* averages $\overline{y}_0$ and $\overline{y}_0^*$ when $\mathbf{X} = 0$; AND:
- ○ the treatment effect in the (half) *sample* with $\mathbf{X} = 1$ is likely to differ from the *true* treatment effect.

In the present discussion, we need consider only the *second* source, because the first, *under EPA*, has no *preferential* effect on comparison error in relation to method (probability or judgement) of sample selecting. Diagrams (2) and (3) above now each become *two* diagrams, depending on the absence or presence of an $\mathbf{X}$-$\mathbf{X}^*$ (*i.e.*, an $\mathbf{X}$-$\mathbf{Z}$) *interaction*. The respective pairs of diagrams, shown at the upper right of the facing page 5.81, are (4) and (5), (6) and (7) – these diagrams assume a *positive* treatment effect; for clarity, they omit comparison error from the first source given above (they all show $\overline{y}_0 = \overline{y}$).

- ● Under EPS, in the 'ideal' case of diagrams (4) and (5) with *no* sample error, relationships (or their absence) are as shown

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 31)



in schema (C) at the right below diagram (7). With *no* Z-Y (and, hence, no X^*-Y) relationship over the units of the respondent population, it is *im*material whether there is an X-X^* (*i.e.*, an X-Z) interaction; this is why diagrams (4) and (5) are the *same*, reflecting *zero* comparison error from the second source (restated on page 5.80).

- When there *is* sample error, as in diagrams (6) and (7) above and relationships are as shown in schema (D) at the lower right (where the *dashed* line denotes *association*), there is comparison error from the *second* source *only* when there is an X-X^* (*i.e.*, an X-Z) interaction [diagram (7)].

(C)
$$\begin{tikzcd} Z \arrow[r] & X^* \\ X \arrow[r] & Y \end{tikzcd}$$

(D)
$$\begin{tikzcd} Z \arrow[r] \arrow[dr] & X^* \arrow[d, dash, dashed] \\ X \arrow[r] & Y \end{tikzcd}$$

We see from this discussion that, when estimating a treatment effect by a *difference* of sample averages (${}_{X=1}\overline{y} - {}_{X=0}\overline{y}$) in an experimental Plan, the *intuitive* idea that there may be *cancellation* between the two sample errors is an *over*-simplification – rather, the absence of interaction makes the difference in average response the *same* for both values of X^*.

∗ **Question with a causative aspect answered using an observational Plan:** for this Question aspect and Plan type, we first distinguish:

○ Z^*: the variate that determines which X^* value each respondent population unit receives, FROM:

○ Z: the 'confounder' whose distribution differs between the respondent *sub*populations with X = 0 and X = 1.

Relevant patterns of variate relationships (or their absence) are shown at the right in schemas (F) and (G) – in schema (G), Z^* and Z *may* be the *same* variate.

(F)
$$\begin{tikzcd} Z^* \arrow[r] & X^* \\ Z \arrow[r] \arrow[d, dash] & Y \\ X \arrow[ur, dash] & \end{tikzcd}$$

(G)
$$\begin{tikzcd} Z^* \arrow[r] & X^* \arrow[d, dash, dashed] \\ Z \arrow[u] \arrow[r] \arrow[d, dash] & Y \\ X \arrow[ur, dash] & \end{tikzcd}$$

⊙ From schema O and equations (5.7.5) and (5.7.7) on page 5.51, we recall that the difference $\overline{Y}_1 - \overline{Y}_0$ being estimated in an *observational* Plan is *not* simply the *treatment* effect; the *inherent* limitation of an observational Plan arising from the *confounding* effect of equation (5.7.5) is represented in the lower part of schemas (F) and (G), where the two (solid) lines represent three possible situations which show an X-Y association:

– (focal variate) X is a cause of (response variate) Y [so there *is* a treatment effect of X on Y];

– (possible 'confounder') Z is a common cause of both X and Y [so there is *no* treatment effect of X on Y];

– (possible 'confounder') Z is associated with X which is *not* a cause of Y [so there is (again) *no* treatment effect of X on Y].

The lower part of schemas (F) and (G) is redrawn at the right below with these three situations shown explicitly in schemas (H), (I) and (J). For observational Plans where these schemas represent the *actual* (but *un*known) state of affairs, the confounding effect of equation (5.7.5) on page 5.51 is:

+ under schema (H), the (main) effect of Z on Y plus, *if* there is an X-Z interaction, the X-Z interaction effect;

+ under schemas (I) and (J), the effect of Z on Y.

(H)
$$\begin{tikzcd} Z \arrow[r] & Y \\ X \arrow[ur] & \end{tikzcd}$$

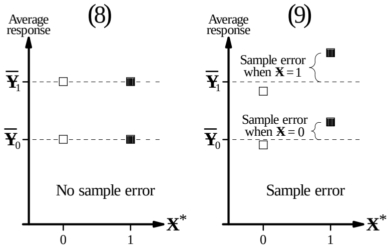
(I)
$$\begin{tikzcd} Z \arrow[r] \arrow[d] & Y \\ X \arrow[ur, dash, dashed] & \end{tikzcd}$$

(J)
$$\begin{tikzcd} Z \arrow[r] \arrow[d, dash, dashed] & Y \\ X \arrow[ur, dash, dashed] & \end{tikzcd}$$

[Limitations *inherent* in observational Plans are pursued in Appendix 15 on pages 5.82 to 5.84].

For a Question with a causative aspect investigated with an observational Plan, the relevant diagrams (8) and (9) below have more in common with diagrams (2) and (3) on the facing page 5.80 than with diagrams (4) to (7) above, except there are now *two* samples selected from the two respondent *sub*populations.

- Under EPS, whether in the 'ideal' case of diagram (8), or diagram (9) where there *are* sample errors but likely of different magnitudes in the two samples, the three benefits from statistical theory are offset by the *inherent* limitation of an observational Plan – there is thus (even under EPS) a *severe* limitation on Answers due to comparison error.
- Under judgement selecting, *lack* of theory and its benefits compounds the already severe limitation from comparison error under EPS to (usually) make *un*acceptable the limitation on Answers imposed by comparison error.



In summary, it may be that, under judgement selecting:

∗ an X^*-Y *relationship* created by Z is (relatively) *common*, thus imposing a (usually) *un*acceptable limitation:

- due to sample error on Answer(s) to Question(s) with a *descriptive*: aspect, AND:
- due to comparison error (as the manifestation of sample error *and* the confounding effect) on Answer(s) to Question(s) with a *causative* aspect in an *observational* Plan (but taking account of the comments in Note 39 about Case-Control Plans on page 5.40 and in Note 55 on pages 5.51 and 5.52), BUT:

∗ an X-X^* (*i.e.*, an X-Z) *interaction* is (relatively) *un*common, thus imposing a (usually) *acceptable* limitation due to compari-

(*continued overleaf*)

son error (as the manifestation of sample error) on Answer(s) to Question(s) with a *causative* aspect in an *experimental* Plan.

NOTES: 89. The foregoing discussion in this Appendix 14 is illuminated by a sampling exercise used over more than a decade in teaching introductory statistics in the 4-year Bachelor of Mathematics program at the University of Waterloo.

A population of 100 'blocks' (irregular polygons cut from 6-mm grey plastic sheet, numbered from 1 to 100) is laid out on a table in the classroom and each of the (50 to 80) students selects a sample of 10 blocks by EPS (using a table of equiprobable digits) and by judgement selecting. From a list of the 100 block weights, each student calculates their two sample averages, which are then used by the instructor to construct, on an overhead projector at the front of the classroom, a bar-graph (in 2-gram intervals) of the averages from each selecting method.

- Under EPS, the bar-graph is usually centred close to the population average block weight (32.4 grams), is roughly Gaussian (or at least symmetrical), and has most of its values within about 10 grams of its centre.
- Under judgement selecting, the centre of the bar-graph is typically at least 40 grams (more than 20% too *high*), the shape is more 'ragged' and the width is appreciably greater than for EPS.

Although this is a *restricted* sampling context, the persistence of sampling *inaccuracy* under judgement selecting is noteworthy – no substantial exception to the characteristics noted above for the judgement-selecting bar-graph was observed in one to two hundred classroom uses of the exercise.

90. An illustration of this Appendix 14 discussion overleaf on page 5.81 is provided by the U.S. Physicians' Health Study (summarized in Figure 10.2), which investigated the effect of aspirin on the risk of heart attack in males.

The sample was 22,071 male doctors, who were assigned to aspirin or placebo under EPA – the treatment and control groups were thus each of size about 11,000. In the notation of Appendix 14, the binary variates were:

- focal variate **Ӿ**: taking placebo (**Ӿ** = 0, 11,034 doctors) or taking aspirin (**Ӿ** = 1, 11,037 doctors),
- 'confounder' **Ƶ**: not being a doctor (**Ƶ** = 0) or being a doctor (**Ƶ** = 1),
- response variate **Ɏ**: not having a heart attack (**Ɏ** = 0) or having a heart attack (**Ɏ** = 1) during the investigation.

Here, the pattern of relationships among the variates is probably more like schema (C) than schema (D) overleaf on page 5.81 – it seems reasonable to assume that the effect of aspirin on heart attack risk in males is not (or is only *weakly*) related to whether a person is a doctor. Also, under EPA, the large sample size reduces the limitation imposed by comparison error from its *first* source. Thus, despite the use of judgement selecting to obtain the sample, there should be acceptable limitation due to comparison error on the Answer from this investigation.

91. The discussion of judgement selecting in this Appendix 14 reminds us how statistics deals with uncertainty [and the resulting limitation on Answer(s)] due to sample (and comparison) error – that is, how statistics deals with *inductive* reasoning from the sample (or the treatment and control groups) to the respondent population.

- In contrast to *predictable* benefits and acceptable limitation due to sample (or comparison) error under *probability* selecting, *judgement* selecting (usually) imposes an *un*acceptable limitation on Answer(s) primarily because of *lack of predictability* of its behaviour under repetition.
  - Judgement selecting *might*, in a particular investigation, yield sample (or comparison) error of *smaller* magnitude than EPS but there is no *theory* to identify *when* this is likely to be the case.
  - This matter is a *statistical* version of the precept that *knowledge is more useful than ignorance*.

  An Answer (*e.g.*, from judgement selecting), no matter *how* severe its limitation, *may* be 'correct' – for instance, early (before 1940) investigations with a Case-Control Plan *correctly* identified cigarette smoking as an explanatory variate associated with the difference between surgery patients admitted to hospital because they had lung cancer and those admitted for *other* diseases – recall also Notes 51 and 52 on pages 5.47 and 5.48.
  - Likewise, an Answer with *acceptable* limitation will sometimes be 'wrong' – too far from the 'truth' to be useful.

92. The discussion in this Appendix 14 also reminds us of the *common* ground in dealing with 'confounder(s)' by EPS in sampling and by EPA in assigning, as portrayed by the schema in Note 53 at the lower right of page 5.48.

93. The parenthetical comment at the end of the first paragraph near the bottom of page 5.79 of this Appendix 14 – that the imposed value of **Ӿ*** does not (usually) change a unit's **Ɏ** value – may not apply in some samples selected from human populations: being in the sample may *change* a unit's response(s). An illustration is a person or family changing TV viewing habits when they keep a diary of programs they watch to provide data for (say) Nielson ratings – recall also the discussion of the 'unit measured' on the lower half of page 5.61 in Appendix 5.

39. Appendix 15: Limitations on Answers from Observational Plans (cited on pages 5.51 and 5.81)

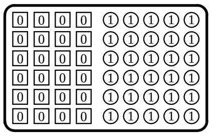
Observational Plans – 'passively' investigating a population in its 'natural' state – to answer a Question with a causative aspect (*i.e.*, a Question about a relationship) impose an *inherent* limitation on Answers.

○ Proportions may behave in surprising ways when comparing population or sample subgroups at different levels of subdivision;
- the reason is changes with level of subdivision of the ('natural') weights involved in the calculations of the proportions;

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 32)

this is the phenomenon of Simpson's Paradox – recall Appendix 9 on pages 5.65 to 5.70.



- ○ A comparative observational Plan is based on the ('natural') *sub*populations of units of the population defined by different values of the focal variate, as represented pictorially for *two* values at the right; for visual convenience, the two subpopulations are separated left and right in the display. The attribute (an average, say) whose values are compared (usually as their *estimates* from samples) are defined for these (two) subpopulations as in equations (5.7.16) and (5.7.17) at the right below, where:
  - – the (binary) focal variate **Ӿ** takes values x of 0 or 1, representing the two 'treatments';
  - – the vertical line in the three subscripts is conditional probability notation and means *given that*;
  - – **¥**(*u*) is the response variate value of unit *u* when **Ӿ**(*u*) = x;
  - – **Ꞥ** is the number of units in the respondent population **P**;
  - – the sums run over all units in the respondent population but:

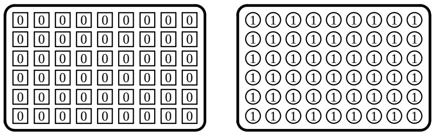
$$\text{average}_{\text{¥}|\text{Ӿ}=0}(\mathbf{P}_{\text{Respondent}}) = \frac{\sum_{\text{all }u}\text{¥}(u)\cdot[1-\text{Ӿ}(u)]}{\text{Ꞥ}-\sum_{\text{all }u}\text{Ӿ}(u)} \qquad \text{-----(5.7.16)}$$

$$\text{average}_{\text{¥}|\text{Ӿ}=1}(\mathbf{P}_{\text{Respondent}}) = \frac{\sum_{\text{all }u}\text{¥}(u)\cdot\text{Ӿ}(u)}{\sum_{\text{all }u}\text{Ӿ}(u)} \qquad \text{-----(5.7.17)}$$

$$\text{average}_{\text{¥}|\text{Ӿ}=\text{x}}(\mathbf{P}_{\text{Respondent}}) = \frac{\sum_{\text{all }u}\text{¥}(u)\,|\,\text{Ӿ}(u)=\text{x}}{\text{Ꞥ}} \qquad \text{-----(5.7.18)}$$

- + in equation (5.7.18) for an *experimental* Plan, *all* **Ꞥ** units contribute to *both* numerator and denominator and so, when *estimating* the two respopndent population averages for **Ӿ** = 0 and **Ӿ** = 1, each sample provides information about the *entire* respondent population for the relevant value of the focal variate;
- + in equations (5.7.16) and (5.7.17) for an *observational* Plan, only the respondent *sub*population units with a given value of **Ӿ** contributes to the relevant expression and so *neither* average (or its estimate) applies to the entire respon- dent population (or, at two stages removed, to the entire target population).

Even when the Plan uses matching (or subdividing) of the units selected for the sample, there is no information on how (the average of) **¥** *would* change if units with one value of **Ӿ** were to take *another* value (the *confounding effect* discussed and illustrated in Sections 22 and 23 on pages 5.49 to 5.52); thus, comparative *observational* Plans only provide Answers about causation with *severe* limitation imposed by comparison error.



- ○ In contrast to an obervational Plan (and as portrayed pictorially at the right and summarized in Table 5.7.47 below), comparative *experimental* Plans with equiprobable assigning (and, preferably, blocking) usually yield (estimates of) respondent population attributes [equation (5.7.18)] with *acceptable* (in context) limitation imposed by comparison error.
  - – *Any* Plan must also deal adequately with limitations imposed by *other* categories of error (study, non-response, sample, sample attribute measurement and model error).

Table 5.7.47: SUMMARY COMPARISON OF OBSERVATIONAL AND EXPERIMENTAL PLANS

Criterion		**Observational Plan**	**Experimental Plan**
Change the value of the *focal* explanatory variate **Ӿ**		Done *passively* as values of **Ӿ** occur naturally	Done *actively* by the investigators
Lurking variate values when **Ӿ** changes	● Hold one or a few fixed	Matching and/or Subdividing (where feasible)	Blocking (where feasible)
	● Quantify the effect of their likely differences	-----	Probability assigning (*e.g.*, EPA)
Effect of *comparison* error on Answers about a *causal* **Ӿ**-**¥** relationship		*Severe* (often *un*acceptable) limitation imposed	Usually *acceptable* limitation imposed

Another way of describing the difference between the two Plan types is:

- ⊙ an *observational* Plan yields Answer(s) about the respondent population *as it **is***;
- ⊙ an *experimental* Plan yields Answer(s) about the respondent population *as it might be enabled to* ***become.***

The *inherent* limitation on the Answer(s) observational Plans can provide may be analogous to the *Postulates of Impotence* in physics – assertions of a conviction that all attempts to do a certain thing, however made, are bound to fail. For example:

- ∗ The Postulate of Relativity: it is impossible to detect uniform translatory motion, possessed by a system as a whole, by observations of phenomena taking place wholly within the system;
- ∗ The Postulate of Thermodynamics: it is impossible to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the surrounding objects.
- ∗ The Postulate of Imperfect Definition: it is impossible to measure precisely the momentum of a particle at the same time as a precise measurement of its position is made.

From Figure 1.1 of these Course Materials and previous discussion in this Figure 5.7, in statistics we assert:

- ∗ The Postulate of Uncertainty: it is impossible to obtain a certain Answer from incomplete information.
- ∗ The Postulate of a Causative Question: it is impossible to establish causation using an observational Plan.

The source given overleaf near the top of page 5.84 states:

(*continued overleaf*)

> We must distinguish a postulate of impotence from an experimental fact and from the statements of pure mathemetics, which do not depend in any way on experience but are necessitated by the structure of the human mind; such a statement as, for instance, 'It is impossible to find any power of two which is divisible by three.' We cannot conceive any universe in which this statement would be untrue, whereas we can quite readily imagine a universe in which any physical postulate of impotence would be untrue.

It is interesting to speculate into which of these two categories the two statistics postulates fall.

SOURCE: Whittaker, E.T.: *From Euclid to Eddington*. *The Tarner Lectures, 1947.* Cambridge University Press, Cambridge, U.K., 1949, pp. 58-60.

40. Appendix 16: Error Categories, Samples and Populations (cited on pages 5.55 and 5.76 in Notes 57 and 58 and Appendix 11)

From the start of Figure 1.1, we have recognized that statistics is concerned with data-based investigating of *populations* (or *processes*), but that resource constraints usually impose *sampling* with its components of *selecting* and *estimating*; thus, sampling is to be set against investigating *all* the (respondent) population units (a census). At the right, the two lists of investigative processes remind us that selecting and estimating are what distinguish investigating based on a sample and on a whole population – the processes of specifying the study population, obtaining responses, measuring variate values and comparing (for a *causative* Question) are *common* to both types of Plan.

A Plan involving a .....

sample:	**population:**
Specifying	Specifying
Selecting	Responding
Responding	Measuring
Measuring	Comparing
Estimating	
Comparing	

We can classify *error categories* on the basis of whether they arise in the context of a Plan involving a sample and/or one involving a whole population as shown at the right in Table 5.7.48; we see that:

- ○ sample error arises *only* when the Plan involves a sample;
- ○ study error, non-response error and attribute measurement error arise *regardless* of whether some of or all the (respondent) population units are being investigated;
- ○ because a (response) model is usually constructed to describe a *sampling* process, model error commonly arises in the context of a Plan involving a sample but models for *populations* are constructed in some investigating.
- ○ comparison error arises only when the investigating involves a Question with a *causative* aspect and, like model error, is usually encountered in the context of a *sample* of units because most comparative Plans involve sampling, but comparison error in phenomena like Simpson's Paradox can arise when the Plan involves *either* a population *or* a sample (recall Appendix 9).

Table 5.7.48: ERROR CATEGORIES

	Arises with a	
Error category	**Sample**	**Population**
Sample	Yes	No
Study	Yes	Yes
Non-response	Yes	Yes
Attribute measurement	Yes	Yes
Model	Yes	(Yes)
Comparison	Yes	Yes

NOTES: 94. Specifying the units which comprise the study population is usually thought of in terms of a **frame** – a *list* of units that may be real or conceptual [*e.g.*, a rule that would, if implemented, generate the list (recall page 5.56)].

95. Investigators may have the opportunity to trade study error and sample error; an illustration is an investigation with a target population of all Canadian adults and resources to select equiprobably 1,000 people for the sample.

- ● A study population of all Canadians residing in Canada would have *smaller* study error but (likely) *larger* sample error because of greater variation among the units of the study population.
- ● A study population of all Canadian university students would have *larger* study error but (likely) *smaller* sample error because of smaller variation among the units of the study population.

A Plan involving *smaller* study error and (likely) *larger* sample error is usually preferred because:

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

+ Under EPS, sampling theory describes the behaviour of sample (and, perhaps, measurement) error under repetition of the selecting and estimating processes [recall the second bullet (●) of Note 10 on page 5.23].

96. When answering a Question with a *descriptive* aspect, there is terminology – **sample survey** and **census** – to distinguish whether the investigating involves some or all of the units of the respondent population.

- ● No such distinguishing terminology exists when answering a Question with a *causative* aspect, perhaps because essentially *all* such investigating involves *sampling*.

41. Appendix 17: Error Categorization in the Social Sciences (cited in Note 58 on page 5.55)

The *idea* of error and its categorization in data-based investigating is widely recognized but terminology differs among disciplines; Table 5.7.49 at the right compares the terminology in this Figure with that in the social sciences. Advantages of the former are:

- ○ several words with the same (or similar) meanings are avoided;
- ○ sources of different categories of error are distinct and well defined;
- ○ the individual case and repetition, and the real world and the model, are clearly distinguished.

Table 5.7.49: ERROR TERMINOLOGY

This Figure 5.7	**Social Sciences**
Study error	External validity, sample bias, generalizability
Non-response error	Sample bias
Sample error	Sampling error, generalization
Measurement error	(Construct validity)
Comparison error	Internal validity
Model error	-----
Measuring inaccuracy	Measurement (in)validity, systematic measurement error
Measuring imprecision	Measurement (un)reliability, random measurement error

(*continued*)

Figure 5.7. DATA-BASED INVESTIGATING: Error – Its Categories and Sources (continued 33)

SOURCE: (for social sciences terminology): Singleton, R.A., Straits, B.C. and M.M. Straits: *Approaches to Social Research*. Second Edition, Oxford University Press, New York, 1993, pages 114-118, 171-172, 185-187, 216, 393, 402, 433 and 452-456.

NOTE: 97. Because they unnecessarily duplicate terminology in STAT 231, words (or phrases) to **avoid** include:

- ∗ **Applicability** (of an Answer) refers to study error and/or sample error.
- ∗ **Generality** (of an Answer) usually refers to sample error; in DOE, **generality** (or a **wider inductive basis**) may refer to whether the Plan involves a factorial treatment structure so that interaction effect(s) can be estimated.
  - – **Generalizability** refers to *study* error and **generalization** to *sample* error (pp. 185-187 in the Source above).
- ∗ **Reliability** [usually] refers to adequate precision (attained by managing *im*precision) [sometimes to adequate accuracy].
- ∗ **Sensitivity** (ability to detect an effect) refers to adequate precision (attained by managing *im*precision).
- ∗ **Strength** (of an Answer) means precision so **weakness** means *im*precision.
- ∗ **Trustworthiness** (of an Answer) means accuracy so **untrustworthiness** means *in*accuracy.
- ∗ **Validity** (of an Answer) means accuracy so **invalidity** means *in*accuracy.

42. Appendix 18: Selecting Protocols and Unit Inclusion Probabilities (cited on pages 5.55, 5.56 and 5.71 in Appendices 1, 2 and 10)

This Appendix 18 illustrates the idea (introduced on page 5.56 in Appendix 2) of the *distinction* between sample selecting and unit inclusion probabilities. The illustration involves a (respondent) population of $\mathbf{N} = 10{,}000$ units and a sample of $n = 100$ units, obtained using six protocols for selecting units (listed in roughly their order of discussionin a course like STAT 332):

- ○ equiprobable selecting of 100 units from the *un*stratified population;
- ○ systematic selecting: selecting equiprobably 1 unit from the *first* 100 population units and then every 100*th* unit;
- ○ equiprobable selecting of 10 *clusters* of 10 units from the population of 1,000 such clusters;
- ○ equiprobable selecting of 10 units from each of the 10 population *strata* each of $\mathbf{N}_h = 1{,}000$ units ($h = 1, 2, \ldots, 10$);
- ○ two-stage selecting: selecting 100 clusters equiprobably and then selecting 1 unit equiprobably from each cluster;
- ○ two-stage selecting: selecting 2 strata equiprobably and then selecting 50 units equiprobably from each stratum.

Table 5.7.50 below uses the symbol $\binom{\mathbf{N}}{n}$, the number of ways n items can be selected from $\mathbf{N}$ items if order of selecting is *un*important; this symbol and its use are discussed in a course like STAT 230 (or in Figure 7.5 of the STAT 220 Course Materials).

Relevant calculations for this Appendix 18 are summarized in Table 5.7.50 below, where the six protocols are now listed in order of *de*creasing number of possible samples. The *short* names for the protocols in the second column of the Table should generally be avoided because their brevity can (temporarily) obscure the nature of, and differences among, the protocols.

Table 5.7.50

			Ratio	..........Selecting probability....................	
Protocol for selecting units	**Short name**	**Number of samples**	**to EPS**	**Sample**	.... **Unit** ....
EPS from an unstratified population	EPS	$\binom{10{,}000}{100} \simeq 6.5\times10^{241}$	1	1.5×10^{-242}	$\binom{1}{1}\binom{9{,}999}{99}/\binom{10{,}000}{100} = \frac{1}{100}$
2-stage EPS from a population in equal-sized clusters	2-stage cluster selecting	$\binom{1{,}000}{100}\binom{10}{1}^{100} \simeq 6.4\times10^{239}$	$\sim10^{-2}$	1.6×10^{-240}	$\binom{1}{1}\binom{999}{99}/\binom{1{,}000}{100}\bullet\binom{1}{1}\binom{9}{0}/\binom{10}{1} = \frac{1}{10}\bullet\frac{1}{10} = \frac{1}{100}$
EPS from a stratified population	Stratified selecting	$\binom{1{,}000}{10}^{10} \simeq 1.6\times10^{234}$	$\sim10^{-7}$	6.2×10^{-235}	$\binom{1}{1}\binom{999}{9}/\binom{1{,}000}{10} = \frac{1}{100}$
2-stage EPS from a stratified population	2-stage stratified slecting	$\binom{10}{2}\binom{1{,}000}{50}^{2} \simeq 4.0\times10^{171}$	$\sim10^{-70}$	2.5×10^{-172}	$\binom{1}{1}\binom{9}{1}/\binom{10}{2}\bullet\binom{1}{1}\binom{999}{49}/\binom{1{,}000}{50} = \frac{1}{5}\bullet\frac{1}{20} = \frac{1}{100}$
1-stage EPS from a population in equal-sized clusters	Cluster selecting	$\binom{1{,}000}{10} \simeq 2.6\times10^{23}$	$\sim10^{-218}$	3.8×10^{-24}	$\binom{1}{1}\binom{999}{9}/\binom{1{,}000}{10} = \frac{1}{100}$
1-in-100 systematic selecting from an unstratified population	Systematic selecting	$100 = 10^{2}$	$\sim10^{-240}$	$\frac{1}{100} = 10^{-2}$	$\frac{1}{100}$

The last four columns of (sometimes approximate) *numerical* table entries are, for each of the six protocols:

- ⊙ the number of samples that can be selected; *i.e.*, the size of the set of all possible samples;
- ⊙ the ratio of the number of samples a protocol can select to the number for EPS from an unstratified population;
- ⊙ the probability any *sample* is selected; here, the *reciprocal* of the number of samples (but see Note 100 overleaf on page 5.86);
- ⊙ the probability any *unit* is included for the sample.
  - – In contrast to the *extreme* variation (over nearly 240 orders of magnitude) of the *sample* selecting probabilities among the protocols, the six *unit* inclusion probabilities are *all* 1 in 100 in this illustration (the *equality* of these six probabilities is a characteristic of this illustration, *not* a general result).
    - + Although *calculations* for *unit* inclusion probabilites for the one-stage and the two-stage clustered and stratified protocols have the *same* structure, they yield vastly different numbers of samples; there are also other important *statistical* distinctions between clustered and stratified protocols – see Appendix 5 on the last side (page 5.96) of Figure 5.8.

(*continued overleaf*)

NOTES: 98. EPS from an unstratified population yields the (exhaustive) set of all *possible* samples of a given size from a population of a given size; this set contains about 6.5×10^{241} samples when **N** = 10,000 units and n = 100 units.

- ● Each of the other five sampling protocols can select only a *sub*set of this (exhaustive) set of samples.
  - – These five protocols are useful because EPS from an unstratified population can rarely meet Plan requirements.
  - = When these protocols are properly implemented, they *preferentially* exclude samples with an extreme value for an attribute like an average, thus *de*creasing sampling imprecision (*e.g.*, recall Note 10 on page 5.23).
- ● As well as yielding all possible samples, EPS is emphasized in introductory discussions because it is:
  - – involved in more practically useful protocols like the last five in Table 5.7.50 overleaf on page 5.85;
  - – the basis of sampling theory for quantifying the behaviour of sample error under repetition, *i.e.*, for quantifying sampling imprecision – recall Note 10 on page 5.23.

Thus, we need to distinguish:

∗ **EPS from an unstratified population**: a protocol for selecting units which is seldom used in practice but which is the basis of sampling theory; FROM:

∗ **EPS** (unqualified): *part* of a protocol for selecting units which involves *other* statistical ideas like stratifying and/or clustering and/or systematic selecting – this is the more *common* usage of ‘EPS’.

99. In practice, selecting uses a **frame** – a real or conceptual *list* of the (respondent) population units; ‘population’ in the protocol descriptions in the first column of Table 5.7.50 overleaf on page 5.85 could therefore be replaced by ‘frame’.

- ● An advantage of two-stage selecting protocols is that, at the second stage, a frame is required *only* for those clusters or strata selected at the first stage – recall Note 62 page 5.57 in Appendix 2.
  - – A protocol for selecting units with three or more stages (see Note 62 on page 5.57v’-.06’) enhances this advantage.

100. Use of EPS at the one or both stages of each protocol means that, in *this* illustration, all *samples* the protocol can select are *equally* likely; as a consequence, the *sample* selecting probability for each protocol in the fifth (‘Sample’) column of Table 5.7.50 overleaf on page 5.85 is the *reciprocal* of its number of samples. [but see the first comment (◎) in Note 62 on the upper half of page 5.57]. Thus, from the perspective of *samples*, the protocols are:

● *alike* in having their possible samples *equi*probable; ● *different* in their numbers of possible samples.

101. A multistage *stratified* selecting protocol for a sample survey of a large geographic area (like a Canadian province) may need to place each large *urban* area in its own stratum. For example, a sample survey in Ontario would rarely want to omit Toronto; its inclusion is *assured* by making Toronto a stratum with an inclusion probability of 1.

102. The ideas of clustering and multistage selecting require us to distinguish the **units** of the selecting process from the **elements** determined by the Question(s) – recall Note 59 at the end of Appendix 1 at the bottom of page 5.55.

- ● For example, an investigation to answer Question(s) about people (*elements*) may use a frame of households (*units*).
  - – If the Question(s) are about households, a unit and an element would be the *same* (a household).

Like most introductory statistics courses, this Figure 5.7 has, for simplicity, largely ignored the element-unit distinction.

Elsewhere, elements may be called **elementary units** or **observation units**; units may be called **sampling units**.

Multistage sampling Plans have **primary sampling units**, **secondary sampling units**, etc., at their successive stages.

SOURCE: MacKay, R.J. *Experimental Design and Sampling*. Course Notes for Statistics 332/362, University of Waterloo, Fall, 2005, page VII – 1.

43. Appendix 19: Themes, Symbols and Acronyms for this Figure 5.7

We use appropriate terminology and notation (see also page 5.25 and its Table 5.7.3) to help maintain distinctions between:

● the population(s) and the sample; ● the individual case and behaviour under repetition; ● the real world and the model.

– Upper case **bold** letters are **population** variates: *e.g.*, **X** (the focal variate), **Y** (the response variate) and **Z** (a confounder); for the respondent population (page 5.34): **N** is its number of units, $\overline{\mathbf{Y}}$ is its average and **S** is its (data) standard deviation.

– Upper case *italic* letters represent *random variables* like Y_j in the response model (5.7.3) on page 5.28.

– Lower case *italic* letters are *values* of random variables; lower case Roman letters are data values (except n is the sample size). The Plan for an investigation must try to make it reasonable statistically to treat data values as values of random variables.

Overall error is the difference between the Answer provided by data-based investigating and the (unknown) answer that reflects the *actual* (or ‘*true*’) state of affairs in the population or process; *we* distinguish six components (pages 5.19, 5.25, 5.52, 5.54). Error is inherent in answers obtained from incomplete information (*e.g.*, from sampling and measuring).

Statistical methods for managing error in data-based investigating have two goals:

∗ to reduce the likely size of error as much as is feasible (but only as much as is needed) in the Question context;

∗ to quantify remaining error in terms of behaviour under repetition.

Acronyms are: – EPS: equiprobable selecting (pages 5.56, 5.86); – EPA: equiprobable assigning (pages 5.37, 5.48); – DOE: Design of Experiments (page 5.31);

– PPDAC: a five-stage structured process for data-based investigating: **P**roblem-**P**lan-**D**ata-**A**nalysis-**C**onclusion, more evocatively renamed FDEAC: **F**ormulation-**D**esign-**E**xecution-**A**nalysis-**C**onclusion.