University of Waterloo W. H. Cherry

ERROR: Comparison Error in Experimental Plans and Observational Plans – The Confounding Effect

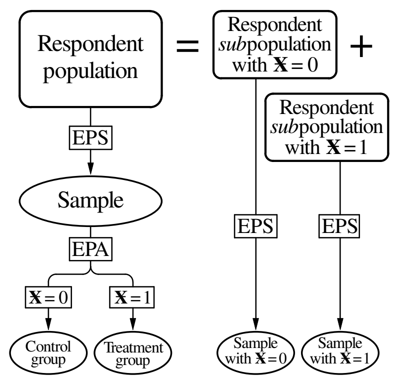
The main ideas in this Highlight #9 are in Section 6 on pages HL9.4 to HL9.7; background information (and matters of related interest) for Section 6 is given in Sections 1 to 5 on the first four sides (pages HL9.1 to HL9.4). Sections 1 to 5 deal with matters typically presented at widely-spaced points in introductory courses and each Section uses ideas from one or more of the other four Sections; this reminds us that a strictly *sequential* presentation of the concepts of introductory statistics is impossible – some terminology must be used before it is defined. The Glossary in Statistical Highlight #91 can be helpful in this regard; further background is provided in five Appendices on pages HL9.7 to HL9.12.

1. Background I – Experimental and Observational Plans [optional reading]

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).

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 (ideally) 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).

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 – see also Section 3 on pages HL63.1 to HL63.3 in Statistical Highlight #63 for further discussion.

2. Background II – Comparative Plans: The Protocol for Setting Levels and Interaction [optional reading]

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 6 on page HL68.2 in Statistical Highlight #68.

  Our concern in this Section 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) that can be applied to a unit [in the sample (or the blocks)].
- ∗ A **run** is part of the Execution stage of an experimental Plan in which all the data are collected for *one* treatment.
- ∗ A **factorial** treatment structure involves *all* combinations of the levels of the (two or more) factors.
- ∗ The **(treatment) effect** of **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**;
  - – 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 **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 **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.

There is further discussion of these matters in the middle of page HL68.1 in Statistical Highlight #68.

(*continued overleaf*)

3. Background III – Comparison Error: Lurking Variates and Confounding [optional reading]

∗ **Lurking variate:** a non-focal explanatory variate (**Z**) whose differing distributions of values over groups of elements or units with different values of the focal variate, if taken into account, would meaningfully change an Answer about an **X**-**Y** relationship.

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

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

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

∗ **Comparison error:** for an Answer about an **X**-**Y** relationship that is based on comparing attributes of groups of elements or units with different values of the focal variate(s), comparison error is the difference from the *intended* (or *true*) state of affairs arising from:

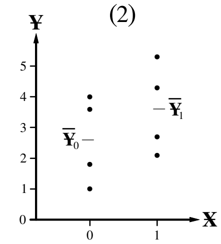
– differing distributions of lurking variate values between (or among) the groups of elements or units OR – confounding.

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

- 'lurking variates' can more readily accommodate phenomena like Simpson's Paradox – see Statistical Highlight #51.
- 'confounding' is more common in the context of comparative Plans (*e.g.*, see Statistical Highlight #68), but the variety of usage of 'confounding' can be a source of difficulty – see Statistical Highlight #3.

4. Background IV – Observational Plans: The Confounding Effect [optional reading]

(2)



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 $\bar{Y}_0$ and $\bar{Y}_1$. The difference between $\bar{Y}_1$ and $\bar{Y}_0$ 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 (HL9.1) below;

$\bar{Y}_1 - \bar{Y}_0$ = effect of change in **X** + effect of change in Z_1, ..., Z_k = treatment effect + confounding effect. -----(HL9.1)

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* **Z**s by matching and/or subdividing.

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

– it is a source of comparison error and the resulting limitation imposed on the Answer(s),

– the treatment effect and the confounding effect cannot be quantified *separately* – we can only know their sum;

thus, our efforts to manage an *inherent* limitation on Answers from observational Plans meet, at best, with only *partial* success.

The confounding effect plays a key role in schema O at the bottom right of page HL9.5 in Section 6 in this Highlight #9.

5. Background V – Equiprobable Selecting and Equiprobable Assigning [optional reading]

∗ **Equiprobable selecting (EPS):** all samples of size n units from a (respondent) population of size **N** units have probability $1/\binom{N}{n}$ of being selected. [This definition implies that all population units have the *same* sample inclusion probability, n/**N** – see Statistical Highlight #85 for further discussion of sample selecting and unit inclusion probabilities.]

∗ **Equiprobable assigning (EPA):** 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.

[There is more detail about these definitions in Statistical Highlights #21 and #63 – *e.g.*, on pages HL21.3 and HL63.3.]

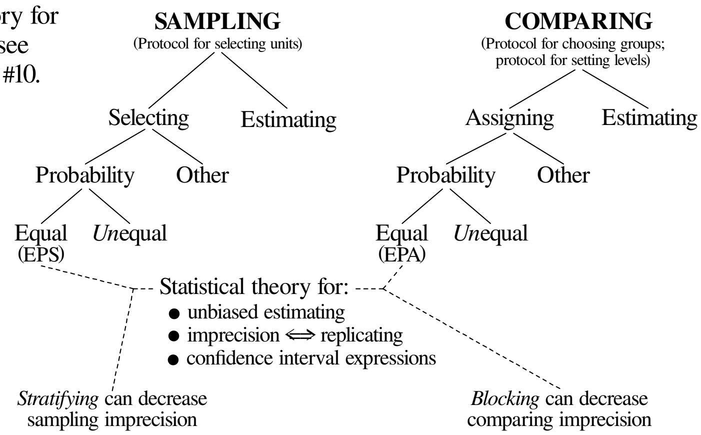
Our *equiprobable selecting* is usually **simple random selecting** (or sampling) or **random selecting** (or sampling) elsewhere; our *equiprobable assigning* is **random assigning** or **randomization** elsewhere.

Similarities of equiprobable *selecting* and equiprobable *assigning*, as components of the processes of sampling and (experimental) comparing, are portrayed by two tree diagrams in the schema at the upper right of the facing page HL9.3.

- 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, and we *hope* for a fortunate outcome of (roughly) equal numbers of units in the groups (*e.g.*, control and treatment) being compared.

(*continued*)

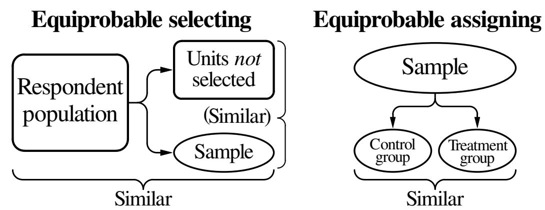
ERROR: Comparison Error in Experimental Plans and Observational Plans – The ... (continued 1)



- + Analogous to EPS, EPA is the basis of statistical theory for the behaviour of *comparison* error under repetition – see the discussion on page HL10.4 in Statistical Highlight #10.
- + Surprisingly, 'probability assigning' is not currently used elsewhere, perhaps reflecting separate development of the two large statistical areas of survey sampling and design of experiments.

– Statistical theory is *used* in the estimating branches of the two tree diagrams in the schema at the right; these branches are part of the Analysis stage of the FDEAC cycle.
  - + Selecting/assigning probabilities as the basis of the theory used for estimating is noteworthy.

– The last two lines of the schema at the right remind us of the analogous roles of stratifying and blocking in sampling and comparing; this analogy is sometimes interpreted as showing that stratifying in survey sampling is merely an instance of blocking, but this (unhelpfully) downplays the different contexts and intents of blocking and stratifying.



● As shown pictorially at the right, a common theme of EPS and EPA is dividing a group of elements into *sub*groups that are likely to be *similar* enough *under adequate replicating* for the respective limitations imposed on Answer(s) by sample error and by comparison error to be acceptable in the investigation context.
  - – When *selecting* the sample, the group of elements is the respondent population, the subgroups are the units *not* selected and the sample.

NOTES: 1. Statisticians have argued (*un*profitably) 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 8 on page HL10.4 in Statistical Highlight #10, 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 HL10.4 on page HL10.3 in Highlight #10, the first and last assignments have comparison error of substantial magnitude in the context of the hypothetical data in Table HL10.3 on page HL10.3;
- ● 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* is true but 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 difficult 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*.

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.

2. Two illustrations of the matters in Note 1 above in the context of *non*-probability assigning are:

Table HL9.1: Ribavirin Trial Data

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

○ 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 HL9.1 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

(*continued overleaf*)

NOTES: (cont.) 2. ○ 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.

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

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

Lind had direct experience of scurvy because he first went to sea with the British Navy in the late 1730s; he spent many years investigating its cause. *Our* interest in Lind's work is because, in 1747, he used an *experimental* Plan to investigate possible treatments; during a voyage which included a ten-week absence from shore and in which 80 of a crew of 350 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 FDEAC cycle, because Lind obtained what is now known to be a *correct* Answer, it is easy to overlook the severe *limitations* on his Answer imposed by:

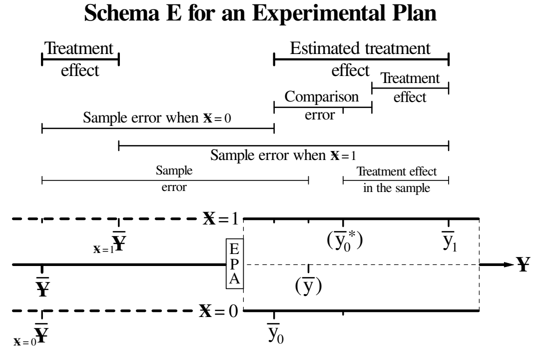
- the small sample size of 12 sailors;
- the non-probability selecting: likely *convenience* selecting of sailors who were on the ship and had scurvy;
- the non-probability assigning – not surprisingly, there is no mention by Lind of the 'modern' idea of probability assigning (*e.g.*, EPA) but some implication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

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

6. Comparison Error in Experimental Plans and Observational Plans [The title matter of this Highlight #9.]

Despite the *probabilistic* equivalence of EPS followed by EPA on the left and EPS of two samples on the right in the schema on page HL9.1, comparison error is involved in *different* ways in the two Plan types, as illustrated in schemas E and O at the right below and at the lower right of the facing page HL9.5.

Schema E for an Experimental Plan



● In schema E representing an *experimental* Plan, the respondent population has (unknown) average $\overline{Y}$ and the sample selected from it by EPS has (unobserved) average $\overline{y}$.

- The *difference* in the values of $\overline{Y}$ and $\overline{y}$ is *sample error*; its value remains *un*known in a particular investigation.
- We also denote the respondent population average, when all elements have $X = 0$, by $_{X=0}\overline{Y}$ and, when all elements have $X = 1$, by $_{X=1}\overline{Y}$; the difference of these (unknown) averages is the (unknown) *treatment effect* – the change in the *average* of Y for *unit* change in X when all the Zs remain fixed.
  + A treatment effect is an attribute describing a relationship.

(*continued*)

University of Waterloo W. H. Cherry

ERROR: Comparison Error in Experimental Plans and Observational Plans – The ... (continued 2)

- The sample is divided (roughly) in half by EPA.
  - One half yields an average $\overline{y}_0$ for the response variate when the focal variate takes assigned value $\mathbf{X} = 0$;
  - The other half yields an average $\overline{y}_1$ for the response variate when the focal variate takes assigned value $\mathbf{X} = 1$.
  - The (observed) *difference* $\overline{y}_1 - \overline{y}_0$ 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 $\overline{y}_0$ and $\overline{y}_0^*$ 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 $\overline{y}_0^*$ 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 3 overleaf on page HL9.6.
  - Comparison error (from both sources) is *eliminated* in the (unattainable) ideal of our three criteria (at the top of page HL9.12 in this Highlight #9) 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 $\mathbf{X} = 0$. -----(HL9.2)

  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 Statistical Highlight #2; see also Section 6 on page HL74.4 in Statistical Highlight #74);
  - *unbiased* estimating (*i.e.*, *zero* comparing inaccuracy) of a treatment effect (a respondent population attribute commonly of interest in a comparative Plan) – see the discussion of Table HL10.4 on page HL10.3 in Statistical Highlight #10.

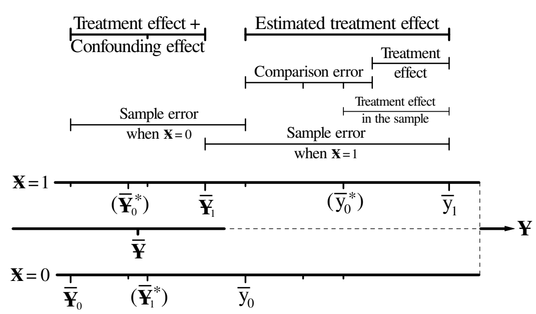
  Hence, EPS and EPA in combination provide for quantifying comparing imprecision and so, *in conjunction with adequate replicating* (or *adequate group sizes*), allow an Answer to be obtained with acceptable limitation imposed by comparison error in the context of a particular investigation with a comparative 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.

    Statistical issues arising from non-probability selecting are discussed in Statistical Highlight #83.
- ○ In schema O below representing an *observational* Plan, the two respondent subpopulations with focal variate values $\mathbf{X} = 0$ and $\mathbf{X} = 1$ have respective (unknown) averages $\overline{\mathbf{Y}}_0$ and $\overline{\mathbf{Y}}_1$.
  - The (unknown) respondent population average $\overline{\mathbf{Y}}$ is the weighted average of $\overline{\mathbf{Y}}_0$ and $\overline{\mathbf{Y}}_1$, the weights being determined by the sizes of the two subpopulations – schema O is drawn with *equal* weights.
  - $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ is the *treatment effect* [due to the different values of $\mathbf{X}$ in the two subpopulations] plus a *confounding effect* [due to differences in the (average) values for one or more lurking variate(s) $\mathbf{Z}_i$ in these subpopulations].
    + From Section 4 on page HL9.2, 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. ----(HL9.3)
    + The two components of $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ in the respondent population can*not* be separated in an observational Plan but, solely to illustrate the present discussion, one *possible* separation is indicated in schema O by the short vertical line on the lower side of the effect bar (at the upper left of schema O) a little to the right of its centre.

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

Schema O for an Observational Plan



2006-06-20 (*continued overleaf*)

- 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:

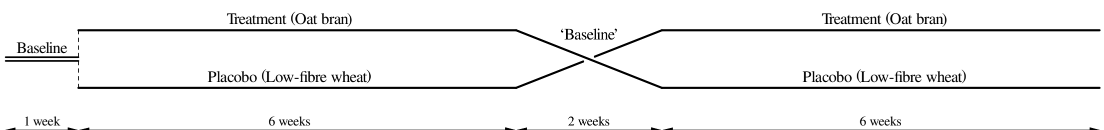
  sample error when $\mathbf{X} = 0$ + comparison error + treatment effect = treatment effect + confounding effect + sample error when $\mathbf{X} = 1$;

  $\therefore$ comparison error = confounding effect + sample error when $\mathbf{X} = 1$ − sample error when $\mathbf{X} = 0$. -----(HL9.4)

  Comparing equations (HL9.2) [overleaf on page HL9.5] and (HL9.4), 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 (HL9.4) 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 also discussed in Appendix 2 on pages HL9.9 and HL9.10.

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: 3. The (half) sample average $\overline{y}_0^*$ in schemas E and O (on pages HL9.4 and HL9.5) 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.
- 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 HL63.2 in Note 7 on page HL63.4 in Statistical Highlight #63.

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]

4. The discussion of this Section 6 starting on page HL9.4 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;

University of Waterloo W. H. Cherry

ERROR: Comparison Error in Experimental Plans and Observational Plans – The ... (continued 3)

NOTES: (cont.) 4. 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 indicate ('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).

7. Appendix 1: Introductory Background – The Formulation Stage of the FDEAC Cycle

The task of the Formulation stage is **question formulation** – turning a ***Question*** into a ***clear Question***. This can be accomplished by addressing matters involved in the components of the Formulation stage. We start with:

∗ A **population**: a well-defined group of units *other than* a sample; a **process** is considered overleaf in Note 5 on page HL9.8.

∗ An **element**: the population entity of interest to the Question(s) to be answered by an investigation and for which variate values could be obtained; *target* elements make up the *target* population. [***Informally***, an 'element' is an 'individual'.]
An *element* is to be distinguished from a **unit**, which is determined by the sampling **frame**.

∗ The **target population** is the group of elements to which the investigator(s) want Answer(s) to the Question(s) to apply.

Associated with each *element* are characteristics called *variates*.

∗ **Response variate:** a variate defined in the Formulation stage of the FDEAC cycle; an Answer describes some attribute(s) [defined below] of the response variate over the target population; our *notation* for a (population) response variate is $\boldsymbol{Y}$.

∗ **Explanatory variate:** a variate, defined in the Formulation stage of the FDEAC cycle, whose change accounts, at least in part, for change in the value of a response variate; our *notation* for a (population) explanatory variate is $\boldsymbol{X}$ or $\boldsymbol{Z}$ (or $\boldsymbol{Z}_i$).

In practice, a useful response variate may need to be a composite of several information dimensions or there may need to be several such variates; however, for simplicity in introductory discussions, we often speak as though there is just *one* (simple) response variate in an application of the FDEAC cycle: *e.g.*, an element's after-tax income in a specified year.

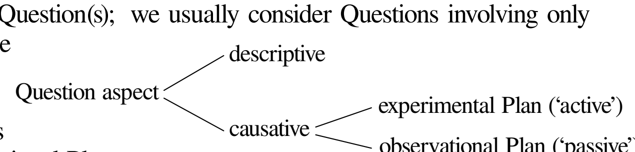
Associated with each *group* of elements (or units) are characteristics called *attributes*.

∗ **Attribute:** a quantity defined as a function of the response (and, perhaps, explanatory) variate(s) over a *group* of elements (or units), typically, the: – target population, – study population – respondent population, – non-respondent population, – sample.

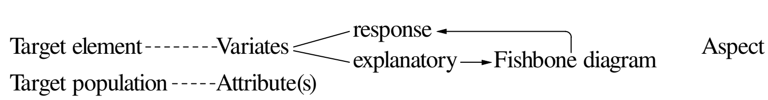
Simple numerical attributes are measures of location like an average, median or proportion and measures of variation like standard deviation or the five-number summary; simple graphical attributes are stemplots, histograms, boxplots, scatter diagrams.

∗ **Fishbone diagram:** a schematic display (reminiscent of a fish skeleton) for *organizing* the names of *explanatory* variates which may affect a particular *response* variate; there can be up to *six* main branches on the diagram, with labels (at least in an industrial context) like **measurement**, **person**, **environment**, **method**, **material** and **machine**.
 – Background information about fishbone diagrams is given in Appendix 3 on pages HL9.10 and HL9.11.

∗ **Aspect:** a binary categorization of the primary concern of the Question to be answered in an investigation, identified in the Formulation stage of the FDEAC cycle; possible aspects are **Descriptive** and **Causative**.
 – The Answer from the investigation of a **descriptive** Question addresses primarily the value(s) for target population *attribute(s)*; these values may be *current* or *extrapolations*, and *comparisons* among attribute values may be involved.
 – The Answer from the investigation of a **causative** Question addresses primarily some characteristic(s) of a *causal relationship* between a response variate and one (or more) explanatory variates, (usually) with the intent that *changing* the value(s) of the explanatory variate(s) would (or will) change the value of the response variate.
  ○ For a Question with a causative aspect, the **focal (explanatory) variate** is the *explanatory* variate whose relationship to the *response* variate is involved in the Answer(s) to the Question(s); we usually consider Questions involving only *one* focal variate in introductory courses. The schema at the right reminds us that, for an investigation to answer a Question with a *causative* aspect, an additional component of the Formulation stage of the FDEAC cycle is the decision whether to use an experimental and an observational Plan.



The components of the Formulation stage are summarized at the right, arranged to remind us that variates are associated with an *element*, attributes with a *population* (or sample) and that a fishbone diagram organizes explanatory variate information in relation to the response variate. Also, aspect refers to the *Question* but its concern is with the nature of the *Answer*.



(*continued overleaf*)

A final matter in the Formulation stage of the FDEAC cycle, after an input Question has become a *clear* Question, is to decide what would constitute a 'correct' Answer – what likely degree of (overall) error would impose *acceptable* limitation on the Answer. Addressing this matter is often a demanding task because of our *six* component categories (as listed below) and because strategies for managing them:

– differ for different error categories, AND: – often involves assumptions that are *im*perfectly met in practice.

An overview of error management strategies by error category is:

⊙ **Study error:** management is based mainly on *extra*-statistical knowledge and is usually difficult to quantify.
⊙ **Non-response error:** management is based on statistical theory seldom covered in introductory courses.
⊙ **Sample error:** in introductory courses, management is based on statistical theory which yields a *confidence interval*.
⊙ **Measurement error:** management is similar to the method for sample error.
⊙ **Comparison error:** management is similar to the method for sample error.
⊙ **Model error:** specific (often graphical) methods are used to assess (somewhat subjectively) modelling *assumptions*.

More details about error management strategies are given in Table HL6.1 on pages HL6.4 and HL6.5 in Statistical Highlight #6.

The concept of (overall) error and its six categories, as *we* use the term, is to be distinguished from the familiar **mistake**, which is *a departure from the Plan*; usually, mistakes arise from carelessness and are avoidable, in contrast to error which is the *un*avoidable consequence of incomplete information and, hence, of limitations imposed on Answers.

∗ **Uncertainty:** ignorance (incomplete knowledge) of error [*e.g.*, ignorance of its magnitude, sign or actual category].
  – *Limitations* on Answers remind us of ever-present uncertainty – an Answer we anticpate (from disciplined use of appropriate statistical methods) will be 'correct' may actually be 'wrong' (too far from the actual state of affairs to be useful) and an Answer likely to be 'wrong' may (by happenstance) be 'correct' [close (or equal) to the true state of affairs].

NOTES: 5. The foregoing discussion of the Formulation stage in this Appendix 1 starts on page HL9.7 by considering a *population*; to answer some types of Question, we may instead start with a *process*, for which 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 HL9.2 at the right above).

Table HL9.2
Populations and Processes

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

6. Our variate qualifiers 'response' and 'explanatory' are ordinary English words that best evoke relevant statistical issues for variates; elsewhere, qualifier pairs commonly used include:

● *dependent* and *independent*; ● *outcome* and *predictor*; ● *outcome* and *exposure*.

No such pair is as statistically evocative as 'response' and 'explanatory'; in addition, the first pair have multiple ordinary and technical meanings that are easily misunderstood and/or involve difficult ideas from outside statistics.

Words which evoke the essential difference between experimental and observational Plans are the **active** assignment by the investigator(s) of focal variate values compared with the **passive** acceptance of its 'natural' values.

7. Elsewhere, our population attribute may be called a population **parameter**; *we* reserve 'parameter' for one of the components of a *mathematical model* – see page HL18.6 in Statistical Highlight #18.

● Also, *we* use 'sample *attribute*' and avoid the unevocative 'sample *statistic*.'

8. Equations (HL9.5) to (HL9.7), at the right below and the top of the facing page HL9.9, give symbolic expressions for three simple numerical attributes: an average, a (data) standard deviation and a proportion, where:

● the (optional) subscript ¥ on an attribute indicates to which response variate the attribute refers – *generic* attribute notation is $a_{\text{¥}}(\mathbf{P}_{\text{Target}})$, where $\mathbf{P}_{\text{Target}}$ denotes the target population;

$$\text{average}_{\text{¥}}(\mathbf{P}_{\text{Target}}) = \frac{\sum_{\text{all } u} \text{¥}(u)}{\mathbf{N}_{\text{T}}} \quad \text{-----(HL9.5)}$$

● $\text{¥}(u)$ is the value of the response variate for element u;

$$\text{standard deviation}_{\text{¥}}(\mathbf{P}_{\text{Target}}) = \sqrt{\frac{\sum_{\text{all } u}[\text{¥}(u) - \text{average}_{\text{¥}}]^2}{\mathbf{N}_{\text{T}} - 1}} \quad \text{-----(HL9.6)}$$

● $\mathbf{N}_{\text{T}}$ is the number of elements in the target population;

(*continued*)

ERROR: Comparison Error in Experimental Plans and Observational Plans – The ... (continued 4)

NOTES: (cont.) 8. ● the summations run over all elements of the target population;

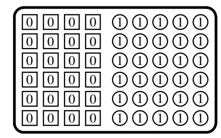
● the proportion refers to *two* categories of elements, C [for which $\mathbf{Y}(u) = 1$] and $\overline{\mathrm{C}}$ [for which $\mathbf{Y}(u) = 0$. Such use of a (binary) **indicator variate** makes the expression (HL9.7) for a proportion in two categories a special case of the expression (HL9.6) for an average.]

$$\text{proportion}_{\mathrm{C}}(\mathbf{P}_{\text{Target}}) = \frac{\sum_{\text{all } u} \mathbf{Y}(u)}{\mathbf{N}_{\mathrm{T}}} \qquad \text{-----(HL9.7)}$$

8. Appendix 2: Limitations on Answers from Observational Plans

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; this is the phenomenon of Simpson's Paradox – see Statistical Highlight #51.
- ○ 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 (HL9.8) and (HL9.9) at the right below, where:



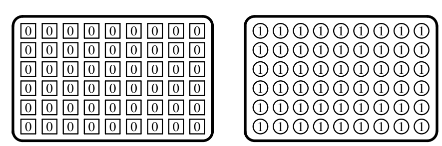
$$\text{average}_{\mathbf{Y}|\mathbf{X}=0}(\mathbf{P}_{\text{Respondent}}) = \frac{\sum_{\text{all } u} \mathbf{Y}(u)\cdot[1 - \mathbf{X}(u)]}{\mathbf{N} - \sum_{\text{all } u} \mathbf{X}(u)} \qquad \text{-----(HL9.8)}$$

$$\text{average}_{\mathbf{Y}|\mathbf{X}=1}(\mathbf{P}_{\text{Respondent}}) = \frac{\sum_{\text{all } u} \mathbf{Y}(u)\cdot\mathbf{X}(u)}{\sum_{\text{all } u} \mathbf{X}(u)} \qquad \text{-----(HL9.9)}$$

$$\text{average}_{\mathbf{Y}|\mathbf{X}=\mathrm{x}}(\mathbf{P}_{\text{Respondent}}) = \frac{\sum_{\text{all } u} \mathbf{Y}(u) \big| \mathbf{X}(u) = \mathrm{x}}{\mathbf{N}} \qquad \text{-----(HL9.10)}$$

  - – the (binary) focal variate $\mathbf{X}$ 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*;
  - – $\mathbf{Y}(u)$ is the response variate value of unit u when $\mathbf{X}(u) = \mathrm{x}$;
  - – $\mathbf{N}$ is the number of elements in the respondent population $\mathbf{P}$;
  - – the sums run over all units in the respondent population but:
    - + in equation (HL9.10) for an *experimental* Plan, *all* $\mathbf{N}$ elements contribute to *both* numerator and denominator and so, when *estimating* the two respopndent population averages for $\mathbf{X} = 0$ and $\mathbf{X} = 1$, each sample provides information about the *entire* respondent population for the relevant value of the focal variate;
    - + in equations (HL9.8) and (HL9.9) for an *observational* Plan, only the respondent *sub*population elements with a given value of $\mathbf{X}$ contributes to the relevant expression and so *neither* average (or its estimate) applies to the entire respondent 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) $\mathbf{Y}$ *would* change if units with one value of $\mathbf{X}$ were to take *another* value (the *confounding effect* discussed and illustrated in Sections 4 and 6 on pages HL9.2 and HL9.4 to HL9.7); 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 HL9.3 below), comparative *experimental* Plans with equiprobable assigning (and, preferably, blocking) usually yield (estimates of) respondent population attributes [equation (HL9.10)] 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 HL9.3: SUMMARY COMPARISON OF OBSERVATIONAL AND EXPERIMENTAL PLANS

Criterion		Observational Plan	Experimental Plan
Change the value of the *focal* explanatory variate $\mathbf{X}$		Done *passively* as values of $\mathbf{X}$ occur naturally	Done *actively* by the investigators
Lurking variate values when $\mathbf{X}$ 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* $\mathbf{X}$-$\mathbf{Y}$ 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 is able to 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:

(*continued overleaf*)

∗ 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 Statistical Highlight #87 and discussion in this Highlight #9 (*e.g.*, on pages HL9.5 and HL9.6) 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 below states:

> We must distinguish a postulate of impotence from an experimental fact and from the statements of pure mathematics, 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.

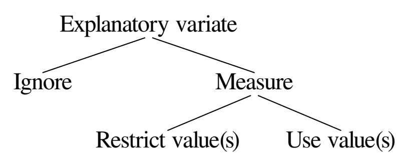
It is interesting to speculate into which of these two categories the two statistics postulates fall.

SOURCE: Whittaker, E.T.: *From Euclid to Eddington. The Tarner Lectures, 1947.* Cambridge University Press, Cambridge, U.K., 1949, pp. 58-60.

9. Appendix 3: Fishbone Diagrams for Comparative Plans

Explanatory variates and their management is the central issue in investigating statistical relationships, as indicated by the lengthy discussion in Statistical Highlights #3, #9, #10 and #57 to #68. 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* 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.

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

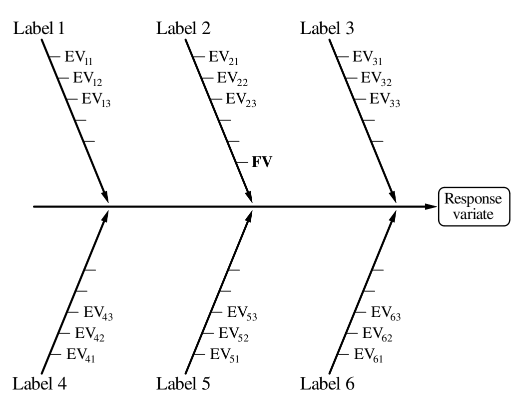
- a clinical trial of a new drug may decide to use participants of one sex and/or a restricted age range;
- when investigating a manufacturing process, the study population of parts might be specified as those parts still at the site – these would usually be parts produced consecutively over a relatively 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 in relation to blocking and matching in Statistical Highlight #63 and to stratifying in Statistical Highlight #85.

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:

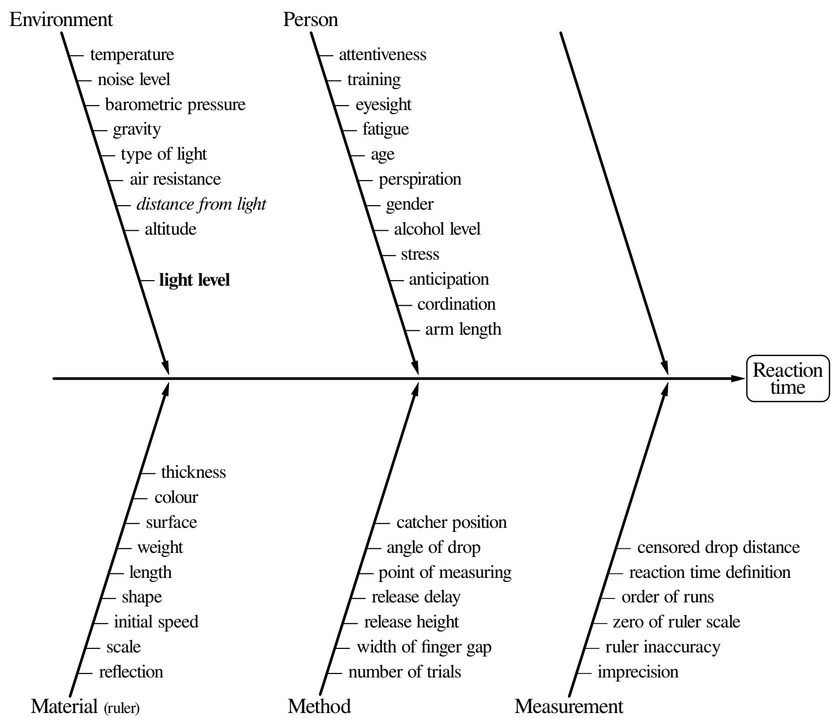
- a box on its right containing the name of the response variate,
- a central horizontal arrow pointing at this box,
- subarrows slanted from left to right and pointing either down or up at the central arrow;
  - each subarrow has a label evocative (in the investigation context) of a category of explanatory 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.



(*continued*)

ERROR: Comparison Error in Experimental Plans and Observational Plans – The ... (continued 5)

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.

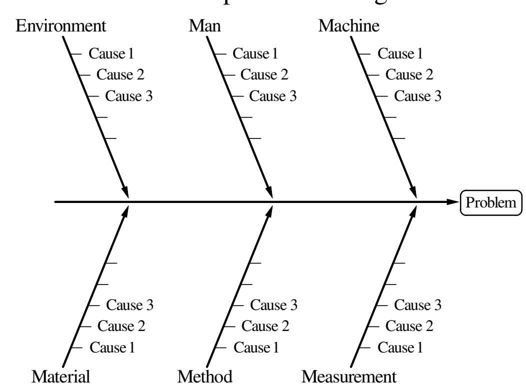


NOTE: 9. Our fishbone diagrams are an adaptation of cause-and-effect diagrams that are one of Ishikawa's seven industrial problem-solving tools:

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 bottom right of the facing page HL9.10, 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 summarized in Statistical Highlight #97 and are discussed in more detail in Figures 11.18 to 11.27 of the STAT 221 Course Materials.

(*continued overleaf*)

10. Appendix 4 – Causation in Statistics: a Definition

To define *formally* in statistics what it means to say $\mathbf{X}$ *causes* $\mathbf{Y}$ in a (target) population, *we* state three criteria:

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

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

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

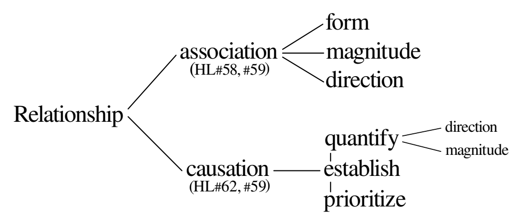
For example, the z values must be the *same* when using least squares estimates [as given in equation (HL9.11) at the right] to *compare* simple linear regression *slopes* when $\mathbf{X} = 0$ and $\mathbf{X} = 1$.

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} y_j (z_j - \bar{z})}{\sum_{j=1}^{n} (z_j - \bar{z})^2} \qquad \text{-----(HL9.11)}$$

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 Statistical Highlight #68.

11. Appendix 5 – 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 Statistical Highlights #58 and #59). This Appendix 5 is concerned with comparative Plans for investigating relationships where causation is to be established or *is* involved; the focus on the $\mathbf{X}$-$\mathbf{Y}$ relationship being *causal* means that a *change* can (potentially) be induced in $\mathbf{Y}$ by *changing* $\mathbf{X}$. These matters are summarized in the schema at the right, which reminds us that:



* association is usually characterized by its *form*, *magnitude* or *direction*;
  - correlation (see Statistical Highlight #66) is one measure of magnitude ('strength') for a straight-line association; form can also be *non*-linear;
* it is useful to distinguish three types of Questions with a causative aspect:
  - **Establishing** whether $\mathbf{X}$ *is* a cause of $\mathbf{Y}$, usually with a view to manipulating $\mathbf{X}$ to produce a (desired) change in $\mathbf{Y}$ – the quintessential example is whether cigarette smoking is a cause of lung cancer (and other life-threatening diseases), the topic of tens of thousands of data-based investigations over several decades starting in the 1940s. Establishing that an observed association of $\mathbf{X}$ and $\mathbf{Y}$ is causation of $\mathbf{Y}$ by $\mathbf{X}$ is answering the Question whether the relevant causal structure (shown again at the right from page HL59.1 in Statistical Highlight #59) is case (1) or case (9) [= case (12)].

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

    (9) $\mathbf{X} \leftarrow \mathbf{Z} \rightarrow \mathbf{Y}$

  - **Quantifying** the relationship between $\mathbf{X}$ (or, more commonly, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_q$) and $\mathbf{Y}$; this arises in the statistical area of *Design of Experiments* (DOE) – for example, the effect of temperature, humidity, light, fertilizer and insecticide levels on the growth of seedlings in a greenhouse. Quantifying a causal relationship is, in essence, investigating the case (1) causal structure – the subscripts on the case number now remind us that the Plan needs to reflect the number of focal variates involved.

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

    $(1)_2$ $\mathbf{X}_1 \rightarrow \mathbf{Y} \leftarrow \mathbf{X}_2$

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

  - **Prioritizing** causes by the size of their effect is the domain of (data-based) process improvement – trying to identify the *most important* cause (usually of excessive variation in the process output, $\mathbf{Y}$) from among many causes $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_q$.

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 STAT 231 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: 10. In STAT 231, *establishing* causation is discussed in Chapter 11 of the Course Notes but the emphasis is on *quan- tifying* 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.