University of Waterloo W. H. Cherry

ERROR: Comparison Error – Quantifying a Treatment Effect under EPA in Experimental Plans

1. Background I – Causation in Statistics: a Definition [optional reading]

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:** $\text{Attribute}(\mathbf{Y}, \text{ perhaps some of } \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k \mid \mathbf{X} = 0) \neq \text{Attribute}(\mathbf{Y}, \text{ perhaps some of } \mathbf{Z}_1, \mathbf{Z}_2, \ldots, \mathbf{Z}_k \mid \mathbf{X} = 1)$;
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 (HL10.1) at the right] to *compare* simple linear regression *slopes* when $\mathbf{X} = 0$ and $\mathbf{X} = 1$.

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} y_j (z_j - \bar{z})}{\sum_{j=1}^{n} (z_j - \bar{z})^2} \qquad \text{-----(HL10.1)}$$

2. Background II – Terminology for Comparative Plans: The Protocol for Choosing Groups [optional reading]

The three criteria defining what *we* mean by causation, reviewed in Section 1 above, involve observing a *population* under two values of the focal variate: with *all* the elements having $\mathbf{X} = 0$ and with *all* the elements having $\mathbf{X} = 1$. We try to approach this ideal in a *sampling* context by having *two* samples, one with its units having $\mathbf{X} = 0$ and the other with its units having $\mathbf{X} = 1$; each sample 'represents' the population under one of the two conditions, in the usual statistical sense of sample attributes being *estimates* of respondent population attributes. When the two samples are *compared* to quantify the change in (the average of) $\mathbf{Y}$ corresponding to a change in $\mathbf{X}$, each *non*-focal explanatory variate must have the *same* value in both samples; otherwise, there is (likely to be) comparison error. For comparative Plans for investigating relationships, we distinguish:

- ∗ an **experimental** Plan – a comparative Plan in which the *investigator(s)* (*actively*) assign the value of the focal variate to each unit in the sample (or in each block);
- ∗ an **observational** Plan – a comparative Plan in which, for each unit selected for the sample, the focal explanatory variate (*passively*) takes on its 'natural' value ***un**influenced* by the investigator(s).

This distinction reflects two types of populations encountered in data-based investigating of relationships.

- ● A population in which all (or most) elements have *one* value of a focal variate of interest, whose value it *is* feasible to change.
  - – An example is a new drug to treat a serious disease – no one would already be taking the drug but it could be given to some participants ($\mathbf{X} = 1$) and withheld from others ($\mathbf{X} = 0$) in a clinical trial (an *experimental* Plan – see Note 5 on page HL10.2).
- ● A population in which each element has one of *two (or more)* values ($\mathbf{X} = 0, 1, \ldots$) of a focal variate of interest, whose value it is *not* feasible to change for any element – see Note 6 overleaf on page HL10.2 and on page HL10.3.
  - – 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.

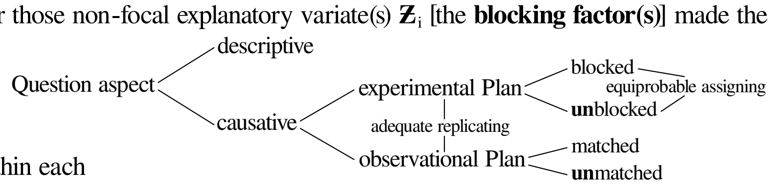
It is investigators' *in*ability to assign elements' (or units') focal variate values that restricts choice of Plan type and so weakens ability to manage comparison error, as discussed in this Highlight #10 and (in more detail) in Statistical Highlight #9.

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 right below.

- ∗ **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) $\mathbf{Z}_i$ [the **blocking factor(s)**] made the same within each block. SO THAT:

  Whether the Question involves establishing causation or quantifying a treatment effect, blocking *prevents confounding* of the focal variate with the $\mathbf{Z}_i$ made the same within each block, reducing the limitation imposed on Answer(s) by *comparison* error.



  - ● By holding one or more $\mathbf{Z}$s fixed within blocks in an experimental Plan, blocking reduces variation in $\mathbf{Y}$ 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 HL10.6 in Section 4. [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:

 (*continued overleaf*)

+ 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) [overleaf at the top of page HL10.1] for those non-focal explanatory variate(s) $\mathbf{Z}_i$ 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 $\mathbf{Z}_i$ 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 sample units 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 HL63.3 and its discussion on pages HL63.4 and HL63.5 in Statistical Highlight #63.

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: 1. Where the definitions given overleaf on page HL10.1 and above of blocking and matching refer to values of non-focal explanatory variates being the *same*, in practice the values may only be *similar.*

2. The groups of elements (or units) are called *blocks* in an experimental Plan but there is no such general term in an observational Plan; however, when the groups contain *two* elements (or units), they may be referred to as *matched pairs* – see Table HL10.1 at the right – but a *block* of two elements (or units) may also be referred to as a 'pair.'

Table HL10.1
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.

3. In design of experiments (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).

4. Just as equiprobable *selecting*, in conjunction with *adequate replicating*, provides a theoretical basis for quantifying the likely size of sample error when estimating a (respondent) population average, so equiprobable *assigning*, in conjunction with *both* EPS *and* adequate replicating, provides the *same* benefit when estimating an average difference in (two) populations in an experimental Plan. This and other parallels between EPS and EPA are discussed in Section 4 on page HL10.6.

5. A special class of comparative experimental investigation is a **clinical trial**, used in medical research to assess the efficacy of new forms of treatment (*e.g.*, drugs, surgery); because the elements are *humans*, a technique called **blinding** is used (where feasible) because of its statistical benefits.
[To be *blind* means not to know, for any element, whether it is in the *treatment* group or the *control* group (which usually receives a dummy treatment known as a **placebo**)]. As shown in Table HL10.2 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 HL10.2

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*

6. For 'focal variate' criterion (2), there are focal variates (like age and sex) whose values can*not* be *assigned* to elements by the investigator(s) in an experimental Plan. For such variates, we avoid the stronger language of saying

ERROR: Comparison Error – Quantifying a Treatment Effect under EPA in Ex... Plans (continued 1)

NOTES: (cont.) 6. *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 given in Section 1 on page HL10.1) 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 Statistical Highlight #51.

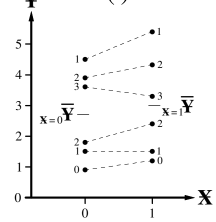
- The ideal of criterion (2) ignores any *time* difference between the realization of the two conditions **X** = 0 and **X** = 1. In actual investigations, the two groups (usually samples) with units having **X** = 0 and **X** = 1 are observed concurrently but, in a cross-over Plan (like the oat bran investigation described in Note 3 on page HL9.6 in Statistical Highlight #9), there *is* a time difference between **X** = 0 and **X** = 1 for both half samples; any changes in units' *other* explanatory variates values over time may then be a source of comparison error.

3. Experimental Plans – Quantifying a Treatment Effect Under EPA [The title matter of this Highlight #10.]

Table HL10.3: Respondent Population Responses (N = 6)

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

(1)



To illustrate properties of *experimental* Plans (and then contrast them with those of observational Plans), hypothetical data for a response variate **Y** are given in Table HL10.3 at the right for a respondent population of six elements under two values [assigned by the investigator(s)] of a focal variate **X** – the treatment effect (the change in the average of **Y** for unit change in **X** when all the **Z**s remain fixed) is 0.3 units, the average of (widely-varying) effects of changing **X** for the individual elements. The data in Table HL10.3 are also shown in diagram (1) at the right; the value of a lurking variate **Z** given beside each dot reminds us that, for our initial discussion, changing **X** does *not* affect the value of **Z** [but see the comment in the second bullet (⊙) beginning at the bottom of this page HL10.3]. The population averages when **X** = 0 and **X** = 1 are shown as short horizontal lines; the differing notation used for these averages between diagrams (1) [above] and (2) [at the centre right of page HL9.2 in Statistical Highlight #9] is to emphasize the distinction between experimental Plans [where the *investigator(s)* assign each element's **X** value (under EPA)] and observational Plans [where each element has its 'natural' **X** value *un*influenced by the investigator(s)].

Table HL10.4 at the right below (which is *un*blocked – see Note 7 overleaf on page HL10.4) shows the twenty possible assignments of the six population elements whose data are given in Table HL10.3 above, together with their response variate averages, treatment effect and comparison error; for example, the first line of Table HL10.4 shows:

- elements 1, 2 and 3 assigned **X** = 0 (often called the 'control group') with average response 1.4,
- elements 4, 5 and 6 assigned **X** = 1 (the 'treatment group') with average response 4.3,
- for this assignment, an estimated treatment effect $\overline{y}_1 - \overline{y}_0$ is 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 HL10.4 have meaning only if all 20 assignments are *equi*probable (as they are under EPA).

[The last column of ten values in *italics* at the right of Table HL10.4 is discussed in the second bullet (⊙) below and overleaf on page HL10.4.]

Table HL10.4: Data for the Set of All 20 Equiprobable Assignments of the 6 Element in Table HL10.3 above

Element nubers **X** = 0	Element nubers **X** = 1	Averages **X** = 0	Averages **X** = 1	Treatment effect	Comparison 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 HL10.4 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 (so-called)

(*continued overleaf*)

– common response to $\mathbf{X}$ [case (10) (shown again at the right) of the causal structures on page HL59.1 in Statistical Highlight #59 and see Note 2 on page HL64.3 in Statistical Highlight #64], this unbiasedness is lost, as the following illustration shows.

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

- ○ Suppose the change in $\mathbf{Z}$ (resulting from the change in $\mathbf{X}$) causes element 4 (with $\mathbf{Z} = 3$) to have a response of 3.9 and an apparent effect of 0.3 instead of its 'true' value of −0.3 – for simplicity, we assume the other five elements (with $\mathbf{Z}$ values *other than* 3) still have the effects given for $\mathbf{X} = 1$ in Table HL10.3. The 10 assignments involving element 4 with $\mathbf{X} = 1$ then have their averages increased by 0.2, as do the corresponding comparison error values (given in *italics* in the last column of Table HL10.4); the *average* of the *twenty* comparison error values is then 0.1 instead of zero, indicating biased estimating of the treatment effect.
- ○ If elements other than 4 were to *also* have their responses when $\mathbf{X} = 1$ changed by the change in $\mathbf{Z}$, some of these changes might be in *opposite* directions, resulting in some *cancellation* and a *smaller* (conceivably zero) magnitude for the comparison error of the *particular* assignment; however, the estimating bias (the *average* comparison error over the set of *all* possible assignments) is unlikely to be meaningfully changed by such (fortuitous) cancellation.

● The average of the set of 20 samples of three elements (or units) with a given $\mathbf{X}$ value is the relevant *population* average – see the right-hand column of Table HL10.3 overleaf on page HL10.3 – this is unbiased estimating of a respondent population average under EPS (see also Statistical Highlight #21 and more detail on pages HL19.1 and HL19.2 in Statistical Highlight #19).

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 $\mathbf{Y}$ are a common response to the focal variate $\mathbf{X}$; *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 $\mathbf{X}$ is changed – recall Note 3 on page HL10.2.]

NOTES: 7. If the responses in Table HL10.3 overleaf on page HL10.3 were *real* data, the Answer about the value of the treatment effect could be made more useful by managing the substantial variation among the elements' effects – *e.g.*, by blocking to decrease comparing imprecision [recall the comment (●) near the bottom of page HL10.1].

8. To avoid confounding, 'lurking variates' criterion (1) in Section 1 at the top of page HL10.1 requires the ideal of *all* non-focal explanatory variates $\mathbf{Z}_i$ holding their values for *every* population element when $\mathbf{X} = 0$ and $\mathbf{X} = 1$.

● 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 $\mathbf{Y}$ to $\mathbf{X}$, EPA addresses criterion (1) for the *other* unblocked, unmeasured and unknown lurking $\mathbf{Z}$s but it does so only *under repetition* – making their distributions (not their values *individually*) the same *on average* across elements when $\mathbf{X} = 0$ and $\mathbf{X} = 1$.

– 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); 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 FDEAC 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 HL10.1 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.

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

● EPA operates *probabilistically* and in *conjunction* with adequate replicating – as discussed in Note 8 above,

 (*continued*)

ERROR: Comparison Error – Quantifying a Treatment Effect under EPA in Ex... Plans (continued 2)

NOTES: (cont.) 9. ● 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, 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* may be true but is unhelpful because it can obscure the following two matters:

- proper use of statistical methods does not *guarantee* a 'correct' Answer – it merely makes an Answer *likely* to be close enough to the actual state of affairs to be useful (*i.e.*, proper use of statistical methods yields an Answer with *acceptable* limitations);
- *im*proper use of statistical methods does not *guarantee* a 'wrong' answer – it may (occasionally) yield a 'correct' Answer; for instance, a response variate measured *in*accurately or *in*correctly on a sample of *one* unit may happen to be close (conceivably *equal*) to the value of the respondent population average.

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

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

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

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

10. Two illustrations of the matters in Note 9 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 HL10.5 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 HL10.5: 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 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 overleaf on page HL10.6, 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. .....

(*continued overleaf*)

NOTES: 10. ⊙ **(cont.)**

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]

4. Background III – 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**.]

∗ **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:

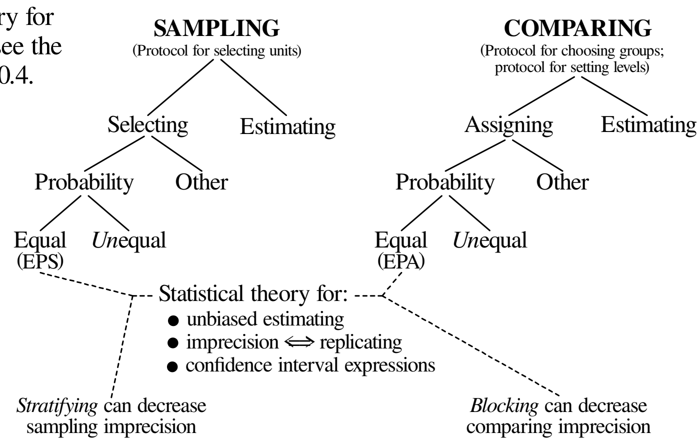
+ across the units of each block in a blocked Plan; + to each unit in the sample in an *un*blocked Plan.

[There is more detail about these definitions on page HL21.3 in Statistical Highlight #21 and on pages HL10.1 and HL10.2.]

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 right below.

- 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.
    + Analogous to EPS, EPA is the basis of statistical theory for the behaviour of *comparison* error under repetition – see the discussion on pages HL10.3 and HL10.4 of Table HL10.4.
    + 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.