Figure 10.8. DATA-BASED INVESTIGATING: Question Aspect and Method of Sample Selecting

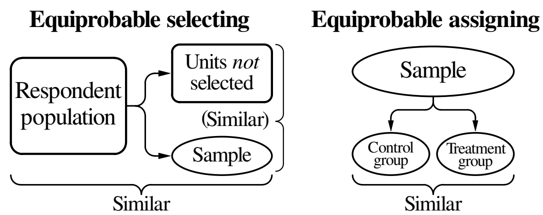
1. Background – Question Aspect and Dividing a Group of Units into Two Subgroups [optional reading]

The Problem stage of the PPDAC cycle introduces terminology which enables a (statistical) Question (the 'input' to this stage) to be turned into a *clear* (statistical) Question (the 'output'). One component of this terminology is the Question

Aspect: a binary categorization of the primary concern of a Question.

- **Descriptive:** a Question whose Answer will involve primarily values for *population/process attributes* (past, present, future).
- **Causative:** a Question whose Answer will involve primarily whether and/or how the focal explanatory variate is *causally* related to the response variate in a population/process.

As discussed in this Figure 10.8, the Question aspect has implications for the method of sample *selecting*.

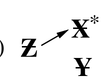


To pursue this discussion, we give at the right a schema from page 5.49 in Figure 5.7, which shows pictorially that a common theme of equiprobable selecting (EPS) and equiprobable assigning (EPA) is dividing a group of units into (two) *sub*groups that are likely to be *similar* enough *under adequate replicating* for the respective limitations imposed on Answer(s) by sample error and by comparison error to be acceptable in the investigation context. The idea from this schema *we* use here is that when *selecting* the sample, the group of units is the respondent population, the subgroups are the units *not* selected and the sample.

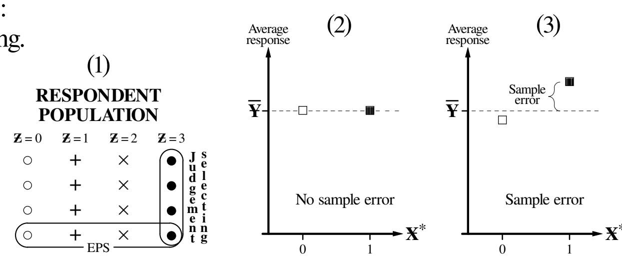
Specifically, we take (binary) focal variate $\mathbf{X}^*$ to indicate whether a unit *is* selected for the sample ($\mathbf{X}^* = 1$) or is in the group of units *not* selected ($\mathbf{X}^* = 0$). The value of a (possibly confounding) explanatory variate $\mathbf{Z}$ determines which $\mathbf{X}^*$ value each respondent population unit receives. [The asterisk (∗) on $\mathbf{X}$ is to remind us that the nature of this focal variate differs from $\mathbf{X}$ in most of our discussion elsewhere; for instance, its values are *imposed* on the units of the respondent population but, *un*like a 'treatment', it (usually) does not actively change a unit's response variate value (but see Note 5 on page 10.20 of this Figure 10.8).] The emphasis in our discussion is to contrast *probability* selecting with *judgement* selecting for Questions with a descriptive aspect and with a causative aspect, under an experimental Plan and under an observational Plan in the latter case.

2. Question with a Descriptive Aspect: probability selecting

Under *probability* selecting (*e.g.*, EPS), a suitable probabilistic process (*e.g.*, equiprobable digits) determines the values of $\mathbf{Z}$ (and, hence, of $\mathbf{X}^*$), so these values are *un*influenced by the units' *other* variate values; with *adequate replicating*, we can therefore usually come acceptably close to the ideal of there being *no* $\mathbf{Z}$-$\mathbf{Y}$ (and, hence, no $\mathbf{X}^*$-$\mathbf{Y}$) relationship over the units of the respondent population. This means in practice that the value of the attribute of interest in the sample will usually be acceptably close to that for the units *not* selected, which means in turn an acceptable limitation on the Answer due to sample error. Schema (A) at the right shows the relationships (or their *absence*) among $\mathbf{Z}$, $\mathbf{X}^*$ and $\mathbf{Y}$.

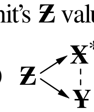
(A) 

- While probability selecting *may* obtain a sample with an attribute value (*e.g.*, an average) meaningfully different from that of the respondent population (*e.g.*, $\overline{\mathbf{Y}}$), statistical theory quantifies, under repetition, the probability of obtaining such a sample – that is, the theory makes explicit the dependence (and its form) of sampling imprecision on degree of replicating (*i.e.*, sample size), as well as providing, when estimating an *average*:
  - **+** a *confidence interval* expression, **+** *unbiased* estimating.
  - **–** Diagram (1) at the right is a representation of a respondent population of $\mathbf{N} = 16$ units with four different $\mathbf{Z}$ values; a sample consisting of the bottom row of four units would yield (the ideal of) diagram (2), in which the sample average (▪) and that of the units *not* selected (▫) are (exactly) *equal* – that is, there is *zero* sample error.



3. Question with a Descriptive Aspect: judgement selecting

Under *judgement* selecting, $\mathbf{Z}$ may be an explanatory variate of the respondent population units, in which case a unit's $\mathbf{Z}$ value influences *both* its $\mathbf{X}^*$ and $\mathbf{Y}$ values so that, as shown in schema (B) at the right, $\mathbf{X}^*$ is *associated* with $\mathbf{Y}$ (the *dashed* line), due to their *common cause* (or 'confounder') $\mathbf{Z}$; an example, in the respondent population in diagram (1) above, would be if judgement selecting obtained the four units with $\mathbf{Z} = 3$.

(B) 

- A possible outcome of judgement selecting is illustrated in diagram (3) at the right above – this diagram [and diagrams (6), (7) and (9) overleaf on page 10.18 and on page 10.19] assume the sample size is one-quarter of the respondent population size and sample error is *positive*.
  - **–** As well as illustrating the (*un*acceptable) limitation imposed by sample error under judgement selecting when answering

(*continued overleaf*)

a Question with a descriptive aspect, diagram (3) *also* reminds us of the usual (*non*-ideal because there *is* sample error) situation under *probability* selecting; the *critical* differences are:

+ judgement selecting does *not* have the three benefits from sampling theory under EPS, reiterated overleaf on page 10.17 (see also the schema in Note 4 at the centre right of page 10.20 in this Figure 10.8 and Note 10 on page 5.23 in Figure 5.7), which allow investigators to manage the inherent uncertainty (arising from incomplete information) of sampling and so to try to make acceptable in the Question context the limitation on an Answer imposed by sample error;
  - ○ when estimating an average under EPS, a consequence of the Central Limit Theorem is a *higher* probability of selecting a sample with sample error of *smaller* magnitude, a *lower* probability of selecting one with *larger* magnitude;
    - ⊙ this may imply that *judgement* selecting, to which the Central Limit Theorem does *not* apply, is prone to sample error of *larger* magnitude than is EPS for a given sample size – see Note 1 on the lower half of page 10.19.

Of course, it is *possible* that, in diagram (1) overleaf on page 10.17, EPS might select the four units with $\mathbf{Z} = 3$ and judgement selecting might select the bottom row of four units – recall also Note 51 on page 5.47 in Figure 5.7.

When answering a Question with a descriptive aspect and when the value of the respondent population attribute being estimated subsequently becomes *known*, statistical experience shows that 'confounding' by $\mathbf{Z}$, resulting in a sample error of unacceptably large magnitude, is *common* under judgement selecting compared with probability selecting.

- For a sample obtained by judgement selecting, the limitation imposed by sample error on an Answer to a Question with a descriptive aspect is so severe that it raises doubt as to whether the investigation should have been undertaken.
  – Judgement (rather than probability) selecting, usually done to conserve resources, is thus statistical *false* economy when answering a Question with a descriptive aspect.

4. Question with a Causative Aspect Answered using an Experimental Plan

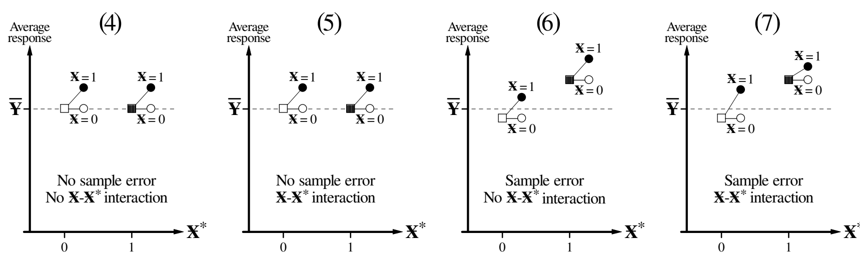
For a Question with a causative aspect, EPS, despite its benefits, is often not feasible and judgement selecting is the *feasible* alternative (recall the oat-bran investigation described in Note 55 on pages 5.51 and 5.52 in Figure 5.7). We deal with this Question aspect and Plan type only for the special case of estimating a treatment effect of focal variate $\mathbf{X}$ which is a *difference* of two *averages* ($_{\mathbf{X}=1}\overline{\mathbf{Y}} - {}_{\mathbf{X}=0}\overline{\mathbf{Y}}$) and so is itself an average; also, we assume that the sample is divided into the treatment ($\mathbf{X}=1$) and control ($\mathbf{X}=0$) groups by EPA. From discussion in Section 23 near the middle of page 5.50 of Figure 5.7, we recall from schema E that *comparison* error has *two* sources:

○ the two half samples obtained under EPA would likely have *different* averages $\overline{y}_0$ and $\overline{y}_0^*$ when $\mathbf{X} = 0$; AND:

○ the treatment effect in the (half) *sample* with $\mathbf{X} = 1$ is likely to differ from the *true* treatment effect.

In the present discussion, we need consider only the *second* source, because the first, *under EPA*, has no *preferential* effect on comparison error in relation to method (probability or judgement) of sample selecting. Diagrams (2) and (3) at the lower right overleaf on page 10.17 now each become *two* diagrams, depending on the absence or presence of an $\mathbf{X}$-$\mathbf{X}^*$ (*i.e.*, an $\mathbf{X}$-$\mathbf{Z}$) *interaction.* The respective pairs of diagrams, shown at the right below, are (4) and (5), (6) and (7) – these diagrams assume a *positive* treatment effect; for clarity, they omit comparison error from the first source given above (they all show $\overline{y}_0 = \overline{y}$).

- Under EPS, in the 'ideal' case of diagrams (4) and (5) with *no* sample error, relationships (or their absence) are as shown in schema (C) at the right below diagram (7). With *no* $\mathbf{Z}$-$\mathbf{Y}$ (and, hence, no $\mathbf{X}^*$-$\mathbf{Y}$) relationship over the units of the respondent population, it is *im*material whether there is an $\mathbf{X}$-$\mathbf{X}^*$ (*i.e.*, an $\mathbf{X}$-$\mathbf{Z}$) interaction; this is why diagrams (4) and (5) are the *same*, reflecting *zero* comparison error from the second source given above in the middle of the page.



- When there *is* sample error, as in diagrams (6) and (7) above and relationships are as shown in schema (D) at the lower right (where the *dashed* line denotes *association*), there is comparison error from the *second* source *only* when there is an $\mathbf{X}$-$\mathbf{X}^*$ (*i.e.*, an $\mathbf{X}$-$\mathbf{Z}$) interaction [diagram (7)].

(C) $\mathbf{Z} \rightarrow \mathbf{X}^*$
$\mathbf{X} \rightarrow \mathbf{Y}$

We see from this discussion that, when estimating a treatment effect by a *difference* of sample averages ($_{\mathbf{X}=1}\overline{y} - {}_{\mathbf{X}=0}\overline{y}$) in an experimental Plan, the *intuitive* idea that there may be *cancellation* between the two sample errors is an *over*-simplification – rather, the absence of interaction makes the difference in average response the *same* for both values of $\mathbf{X}^*$.

(D) $\mathbf{Z} \rightarrow \mathbf{X}^*$, $\mathbf{Z} \rightarrow \mathbf{Y}$
$\mathbf{X} \rightarrow \mathbf{Y}$

5. Question with a Causative Aspect Answered using an Observational Plan

For this Question aspect and Plan type, we first distinguish:

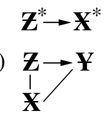
○ $\mathbf{Z}^*$: the variate that determines which $\mathbf{X}^*$ value each respondent population unit receives, FROM:

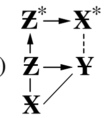
○ $\mathbf{Z}$: the 'confounder' whose distribution differs between the respondent *sub*populations with $\mathbf{X} = 0$ and $\mathbf{X} = 1$.

(*continued*)

Figure 10.8. DATA-BASED INVESTIGATING: Question Aspect and ..... Selecting (continued 1)

Relevant patterns of variate relationships (or their absence) are shown in schemas (F) and (G) at the right – in schema (G), $\mathbf{Z}^*$ and $\mathbf{Z}$ *may* be the *same* variate.

(F) 

(G) 

$\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ = treatment effect + confounding effect ----(10.8.1)

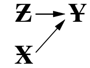
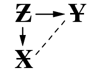
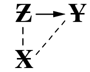
comparison error = confounding effect + sample error when $\mathbf{X}=1$ – sample error when $\mathbf{X}=0$ ----(10.8.2)

⊙ From schema O and equations (5.7.5) and (5.7.7) on page 5.51 in Figure 5.7 – given again for convenience as equations (10.8.1) and (10.8.2) above – we recall that the difference $\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_0$ being estimated in an *observational* Plan is *not* simply the *treatment* effect; the *inherent* limitation of an observational Plan arising from the *confounding* effect of equation (5.7.5) ≡ (10.8.1) is represented in the lower part of schemas (F) and (G), where the two (solid) lines represent three possible situations which show an **X**-**Y** association:

- (focal variate) **X** is a cause of (response variate) **Y** [so there *is* a treatment effect of **X** on **Y**];
- (possible 'confounder') **Z** is a common cause of both **X** and **Y** [so there is *no* treatment effect of **X** on **Y**];
- (possible 'confounder') **Z** is associated with **X** which is *not* a cause of **Y** [so there is (again) *no* treatment effect of **X** on **Y**].

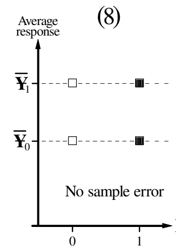
The lower part of schemas (F) and (G) is redrawn at the right below with these three situations shown explicitly in schemas (H), (I) and (J). For observational Plans where these schemas represent the *actual* (but *un*known) state of affairs, the confounding effect of equation (10.8.1) above is:

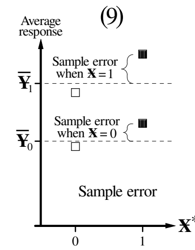
+ under schema (H), the (main) effect of **Z** on **Y** plus, *if* there is an **X**-**Z** interaction, the **X**-**Z** interaction effect;
+ under schemas (I) and (J), the effect of **Z** on **Y**.

(H)  (I)  (J) 

[Limitations *inherent* in observational Plans were discussed in Appendix 15 on pages 5.82 to 5.84 in Figure 5.7].

For a Question with a causative aspect investigated with an observational Plan, the relevant diagrams (8) and (9) below have more in common with diagrams (2) and (3) on page 10.17 than with diagrams (4) to (7) on the facing page 10.18, except there are now *two* samples selected from the two respondent *sub*populations.





- Under EPS, whether in the 'ideal' case of diagram (8), or diagram (9) where there *are* sample errors but likely of different magnitudes in the two samples, the three benefits from statistical theory are offset by the *inherent* limitation of an observational Plan – there is thus (even under EPS) a *severe* limitation on Answers due to comparison error.
- Under judgement selecting, *lack* of theory and its benefits compounds the already severe limitation from comparison error under EPS to (usually) make *un*acceptable the limitation on Answers imposed by comparison error.

In summary, it may be that, under judgement selecting:

* an $\mathbf{X}^*$-**Y** *relationship* created by **Z** is (relatively) *common*, thus imposing a (usually) *un*acceptable limitation:
  - due to sample error on Answer(s) to Question(s) with a *descriptive*: aspect, AND:
  - due to comparison error (as the manifestation of sample error *and* the confounding effect) on Answer(s) to Question(s) with a *causative* aspect in an *observational* Plan (but taking account of the comments in Figure 5.7 in Note 39 about Case-Control Plans on page 5.40 and in Note 55 on pages 5.51 and 5.52), BUT:
* an **X**-$\mathbf{X}^*$ (*i.e.*, an **X**-**Z**) *interaction* is (relatively) *un*common, thus imposing a (usually) *acceptable* limitation due to comparison error (as the manifestation of sample error) on Answer(s) to Question(s) with a *causative* aspect in an *experimental* Plan.

NOTES: 1. The foregoing discussion in this Figure 10.8 is illuminated by a sampling exercise used over more than a decade in teaching introductory statistics in the 4-year Bachelor of Mathematics program at the University of Waterloo.

A population of 100 'blocks' (irregular polygons cut from 6-mm grey plastic sheet, numbered from 1 to 100) is laid out on a table in the classroom and each of the (50 to 80) students selects a sample of 10 blocks by EPS (using a table of equiprobable digits) and by judgement selecting. From a list of the 100 block weights, each student calculates their two sample averages, which are then used by the instructor to construct, on an overhead projector at the front of the classroom, a bar-graph (in 2-gram intervals) of the averages from each selecting method.

- Under EPS, the bar-graph is usually centred close to the population average block weight (32.4 grams), is roughly Gaussian (or at least symmetrical), and has most of its values within about 10 grams of its centre.
- Under judgement selecting, the centre of the bar-graph is typically at least 40 grams (more than 20% too *high*), the shape is more 'ragged' and the width is appreciably greater than for EPS.

Although this is a *restricted* sampling context, the persistence of sampling *inaccuracy* under judgement selecting is noteworthy – no substantial exception to the characteristics noted above for the judgement-selecting bar-graph was observed in one to two hundred classroom uses of the exercise.

2. An illustration of this Figure 10.8 discussion above and on the facing page 10.18 is provided by the U.S. Physicians'

 (*continued overleaf*)

NOTES: (cont.) 2. Health Study (summarized in Figure 10.2), which investigated the effect of aspirin on the risk of heart attack in males. The sample was 22,071 male doctors, who were assigned to aspirin or placebo under EPA – the treatment and control groups were thus each of size about 11,000. In the notation of this Figure 10.8, the binary variates were:

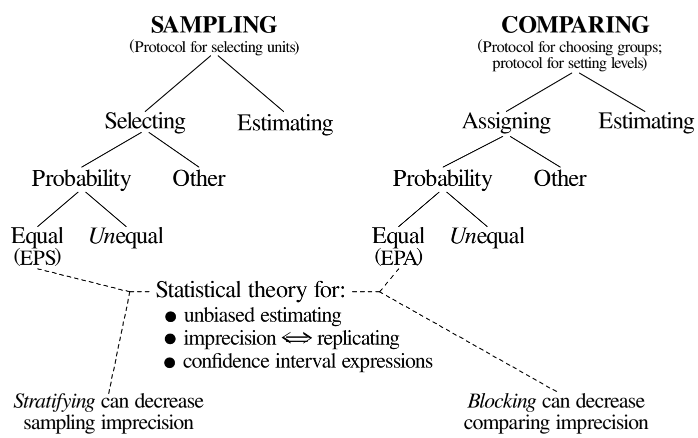
- focal variate $\mathbf{X}$: taking placebo ($\mathbf{X}=0$, 11,034 doctors) or taking aspirin ($\mathbf{X}=1$, 11,037 doctors),
- ‘confounder’ $\mathbf{Z}$: not being a doctor ($\mathbf{Z}=0$) or being a doctor ($\mathbf{Z}=1$),
- response variate $\mathbf{Y}$: not having a heart attack ($\mathbf{Y}=0$) or having a heart attack ($\mathbf{Y}=1$) during the investigation.

Here, the pattern of relationships among the variates is probably more like schema (C) than schema (D) on page 10.18 – it seems reasonable to assume that the effect of aspirin on heart attack risk in males is not (or is only *weakly*) related to whether a person is a doctor. Also, under EPA, the large sample size reduces the limitation imposed by comparison error from its *first* source (given in the middle of page 10.18). Thus, despite the use of judgement selecting to obtain the sample, there should be acceptable limitation due to comparison error on the Answer from this investigation.

3. The discussion of judgement selecting in this Figure 10.8 reminds us how statistics deals with uncertainty [and the resulting limitation on Answer(s)] due to sample (and comparison) error – that is, how statistics deals with *inductive* reasoning from the sample (or the treatment and control groups) to the respondent population.
   - In contrast to *predictable* benefits and acceptable limitation due to sample (or comparison) error under *probability* selecting, *judgement* selecting (usually) imposes an *un*acceptable limitation on Answer(s) primarily because of *lack of predictability* of its behaviour under repetition.
     - Judgement selecting *might*, in a particular investigation, yield sample (or comparison) error of *smaller* magnitude than EPS but there is no *theory* to identify *when* this is likely to be the case.
     - This matter is a *statistical* version of the precept that *knowledge is more useful than ignorance*.

   An Answer (*e.g.*, from judgement selecting), no matter *how* severe its limitation, *may* be ‘correct’ – for instance, early (before 1940) investigations with a Case-Control Plan *correctly* identified cigarette smoking as an explanatory variate associated with the difference between surgery patients admitted to hospital because they had lung cancer and those admitted for *other* diseases – recall also Notes 51 and 52 on pages 5.47 and 5.48 in Figure 5.7.
     - Likewise, an Answer with *acceptable* limitation will sometimes be ‘wrong’ – too far from the ‘truth’ to be useful.

4. Some discussion in this Figure 10.8 reminds us of the *common* ground in dealing with ‘confounder(s)’ by EPS in sampling and by EPA in assigning, as indicated in the schema at the right (from Note 53 on page 5.48 in Figure 5.7 but with*out* its accompanying comments on pages 5.48 and 5.49), although the useful statistical insight this schema provides is incidental to the main discussion of this Figure 10.8.



5. The parenthetical [] comment near the end of the second paragraph on the upper half of page 10.17 of this Figure 10.8 – that the imposed value of $\mathbf{X}^*$ does not (usually) change a unit’s $\mathbf{Y}$ value – may not apply in some samples selected from human populations: being in the sample may *change* a unit’s response(s). Illustrations of this phenomenon (from the discussion of the ‘unit measured’ on the lower half of page 5.61 in Appendix 5 in Figure 5.7) are:
   - Maclean’s ranking of Canadian universities might make universities change their operations in ways that would improve their ranking but make no substantive change to the quality of the educational experience they offer students.
   - Households selected for a panel used to obtain Nielsen ratings of TV programs might change their TV viewing habits as a consequence of *knowing* their viewing habits are being monitored (*e.g.*, when the household keeps a diary of programs watched).
   - The interviewer administering a questionnaire (the ‘operator’) might (unintentionally) influence the person responding.
   - A *slanted* question on a questionnaire may have a different effect on different (types of) respondents.

   An extreme case is when measuring *destroys* the unit (*e.g.*, in quality assurance, firing shotgun cartridges, measuring cigarette tar and nicotine levels or the bursting pressure of plastic bags and condoms); destructive measuring precludes the statistical benefits from repeated measuring on the same unit. [This is the same *statistical* issue as attempting repeated measuring when a questionnaire is involved – see the comment (+) on the lower half of page 5.60 in Appendix 5 in Figure 5.7].