University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview

1. Introductory Overview

This Highlight #88 develops the idea from Statistical Highlight #87 that statistics is concerned with data-based investigating. *Successful* data-based investigating means obtaining a 'correct' *answer* to a *question* without unnecessary expenditure of resources (money, time, etc.). Although a 'correct' answer may occasionally be obtained by guessing or some other short-cut, such methods do not have the potential to succeed in *repeated* applications. For *long-term* success, a *structured process* is needed.

∗ A *'correct'* answer is one which is close enough to the *actual* state of affairs (the 'truth') to be useful in the question context.
- An answer which is a *number* (like an average, a standard deviation or a proportion) will rarely be *equal* to the true value, because it comes (by *inductive* reasoning) from *in*complete data; however, such an answer can still be 'correct'.
- An answer which is one of two or more options like 'Yes' and 'No' (*e.g.*, to a Question like: *Is* **X** *a cause of* **Y**?) is more obviously 'correct' [the (actual) 'truth'] or 'wrong' [too far from the 'truth' to be useful in the question context].

In statistics, 'correct' and 'wrong' (numerical) answers *both* involve **error** – the difference between what is stated [*e.g.*, in an answer] or assumed [*e.g.*, in a response model] and the *actual* state of affairs; the distinction is that the likely degree of error imposes *acceptable* limitation on a 'correct' answer, *un*acceptable limitation on a 'wrong' one. Error is important because:
- six *categories* of error guide statistical practice and inform development of statistical theory;
  + error leads to recognizing the ideas of imprecision, inaccuracy and uncertainty and to their succinct definitions;
    ⊙ we then see *why* statistical methods aim to *manage imprecision (by managing variation) and inaccuracy*;
- error is the *source* of limitations imposed on Answer(s).

∗ Our concern with *long-term* behaviour has a parallel in the widely-acknowledged success of Japanese *manufacturing* processes – one of the reasons cited for this success is the stress the Japanese place on *long*-term thinking, rather than on the *short*-term performance of manufacturing processes that is a North American preoccupation.
- W. Edwards Deming, in his *fourteen points*, also emphasizes the importance of *long*-term thinking in his idea of *constancy of purpose*, and he cites emphasis on short-term profits and short-term thinking as 'diseases that stand in the way of the transformation' he advocates for management style (see Appendix 1 on page HL88.16).
- Long-term behaviour is familiar from the numerical evaluation of probabilities by *long-run* proportions.

A 5-stage structured process for data-based investigating is the Formulation-Design-Execution-Analysis-Conclusion (abbreviated FDEAC) cycle. We refer to a *cycle* because *one* pass through the five stages may not be enough – sub-investigations (and, hence, sub-FDEAC cycles) are sometimes needed; an example is assessing the measuring process(es) that will be used in an investigation. An overview of the five stages of the FDEAC cycle is:

- **F**ormulation: identifying: **what** population or process is to be investigated
  what would constitute a 'correct' Answer(s) to the Question(s);
- **D**esign: drawing up a Plan for **how** to carry out the investigation defined in the Formulation stage;
- **E**xecution: collecting the **data** according to the Plan developed in the Design stage;
- **A**nalysis: summarizing and analyzing the data from the Execution stage – going from *data* to **information**;
- **C**onclusion: using the information from the Analysis stage to give a 'correct' Answer(s), albeit with limitations, to the Question(s) – going from *information* to **knowledge**.

The components of the five stages are summarized in Table HL88.1 below; elaboration of this overview occupies the rest of this Highlight #88. Unfortunately, the presentation cannot be strictly *sequential* – at times, a definition or description needs to use a term not yet defined; the Glossary in Statistical Highlight #91 can then be helpful as a *dictionary* when studying this Highlight #88.

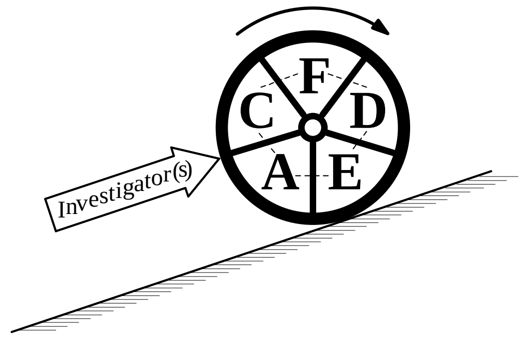
Table HL88.1: The FDEAC cycle: a *structured process* for data-based investigating

Stage	Formulation stage	Design stage	Execution stage	Analysis stage	Conclusion stage
Input	***Question(s)***	***clear Question(s)***	***a Plan***	***Data***	***Information***
Components	Target element Target population/process Variates: • Response • Explanatory Attributes Fishbone diagram Aspect: • Descriptive • Causative	Study element/unit Study population/process Respondent population/process Refine response variate(s) Deal with explanatory variates Protocol for: • Selecting units • Choosing groups • Setting levels Measuring process(es) Plan for the: • Execution stage • Analysis stage	Execute the Plan Monitor the data Examine the data Store the data	Informal analysis: • Numerical attributes • Graphical attributes • Other informal methods Assess modelling assumptions Formal analysis: • Confidence intervals ○ Prediction intervals • Significance tests • Other formal methods	In the language of the Question context: Answer(s) Limitations Recommendations ['*Evidence*-based decisions, improvements, ...' means using Answers from *data*-based investigating with an *adequate* Plan.]
Output	***clear Question(s)***	***a Plan***	***Data***	***Information***	***Knowledge***

 (*continued overleaf*)

To emphasize the *onerous* nature of data-based investigating, the diagram at the right conveys three images relevant to using the FDEAC cycle to obtain Answer(s), with acceptable limitation in the investigation context, to substantive (statistical) Question(s):

- the *effort* of pushing a (heavy) object *up*hill;
- potential *waste* of resources by *premature* cessation of effort;
- the *circular* object is a reminder of the FDEAC *cycle*.



Despite the 'obviousness' of matters like formulating *clear* Question(s) and using measuring processes of acceptable imprecision and inaccuracy in the investigation context, often-poor implementation of these and other components of the FDEAC cycle (*un*necessarily) yields 'wrong' Answer(s).

2. Projects, Investigations and Problems

To develop the ideas encompassed by the FDEAC cycle, we distinguish between:

∗ a **project**, which is *broad* and involves *many* questions [one goal of project *formulation* is to *prioritize* these many questions];

∗ an **investigation**, which is *narrower* and involves *one* (*or a few*) question(s) – this question(s) may arise from a project or it may be of interest in its (or their) own right.

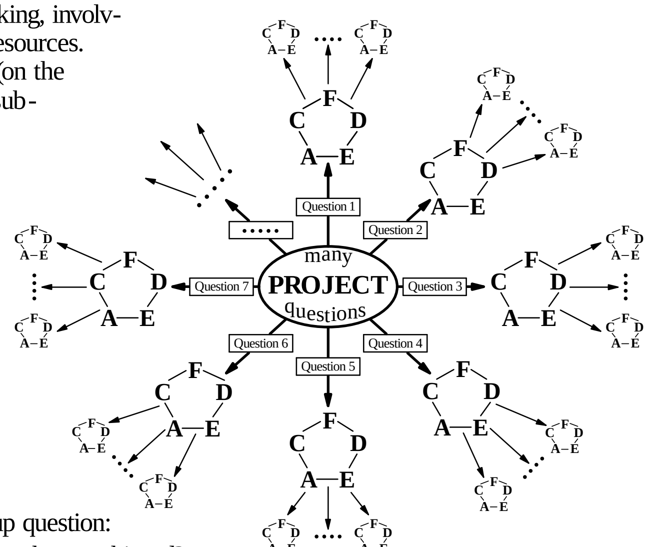
- The question(s) to be answered are the *input* to the Formulation stage of the FDEAC cycle.
- With*in* the FDEAC cycle, the *output* of each stage is the *input* to the next stage, except that the *knowledge* output from the Conclusion stage *answers* the input *question(s)* to the Formulation stage – to portray this visually, Table HL88.1 overleaf at the bottom of page HL88.1 could be formed into a cylinder by joining the respective left- and right-hand edges of the Formulation and Conclusion columns.

  Further discussion of dependencies between stages of the FDEAC cycle, and between four of the stages and the model (which is needed as a basis for *formal* methods of data analysis) is given in Appendix 2 on page HL88.16.

In both projects and investigations, there will inevitably be matters of formulation that require subject-matter expertize – that is, *extra*-statistical knowledge. Examples of projects (phrased as ***How can*** .... questions) are:

- How can the amount of scrap produced by a manufacturing process be reduced?
- How can the quality of drinking water in a province be improved?
- How can the performance on standardized tests of students in a province be improved?
- How can transparent decision-making processes in a large organization be achieved?
- How can a large software system be made less prone to failure or easier to maintain?
- How can satisfaction for the customers of a company providing goods or services be increased?

As indicated in these examples, a project is a *broad* undertaking, involving answers to *many* questions and substantial commitment of resources. The tasks of *formulating* a project and *prioritizing* its questions (on the basis of factors like importance, cost, logical necessity) require subject-matter expertize. The display at the right shows the prioritized questions from a project as inputs to a sequence of FDEAC cycles; each 'main' FDEAC cycle (shown larger) may involve one or more (smaller) sub-cycles to answer questions that arise in answering each 'main' question.



For a question arising from a project or of interest in its own right, we distinguish between the need for:

● extra-statistical knowledge, AND ● data,

to answer it. For example, a quarterback may ask:

∗ How can I increase my proportion of completed passes?

Obviously, *extra*-statistical knowledge is needed to develop a strategy that may produce improvement in passing; THEN: an investigation using the FDEAC cycle can answer the follow-up question:

∗ Has the desired increase in the proportion of completed passes been achieved?

To provide a 'correct' answer, such an investigation would likely need to go beyond the simple-minded comparison of two proportions calculated before and after implementing the improvement strategy. If the desired increase has *not* been achieved, another strategy could be developed, again using extra-statistical knowledge, and its results assessed using the FDEAC cycle.

The defining characteristic of a (statistical) *question* is the requirement for *data* to answer it – see also Statistical Highlight #90.

As with answering a statistical question using the FDEAC cycle, a question requiring *extra*-statisatical knowledge is usually

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 1)

also best answered using a *structured process* – for example, the five-stage DefineMeasureAnalyzeImproveControl of Six Sigma. *Statistical* questions arise most obviously in the I stage of DMAIC, but the FDEAC cycle is also relevant to the tasks of project description and question prioritizing in the D stage; statistical considerations of measuring processes, introduced in Section 9 on page HL88.14 and pursued in Statistical Highlight #38, are relevant in the M stage.

- To help distinguish contexts which require *distinct* structured processes like the FDEAC cycle and DMAIC, we can say:
  - the FDEAC cycle deals with *answering (statistical) Questions*,
  - DMAIC deals with *(extra-statistical) problem solving*.

  We *avoid* equivocal phrases like *a task to be done* or *a conclusion to be drawn*.

3. The Formulation Stage

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 elements *other than* a sample; a **process** is defined overleaf in Note 1 on page HL88.4.

∗ An **element** is the populationc entity of interest to the Question(s) to be answered and for which variate values could be obtained; *target* elements make up the *target* population. [***Informally***, an element is an ‘individual’.]

∗ 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 response variate is $\mathbf{Y}$.

∗ **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 an explanatory variate is $\mathbf{X}$ or $\mathbf{Z}$ (or $\mathbf{Z}_i$).

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 are characteristics called *attributes*.

∗ **Attribute:** a quantity defined as a function of the response (and, perhaps, explanatory) variate(s) over a *group* of elements, like:
  – the target population, – the study population, – the respondent population, – the non-respondent population, – the 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 like **measurement**, **person**, **environment**, **method**, **material** and **machine**.
  – Fishbone diagrams are discussed in Appendix 4 on pages HL88.17 and HL88.18.

∗ **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. As indicated in Note 6 in the middle of page HL88.5, a Question with a *causative* (as distinct from a *descriptive*) aspect raises additional matters to consider in the Formulation and Design stages of the FDEAC cycle.

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*.

Target element -------- Variates < response ←; explanatory → Fishbone diagram   Aspect
Target population ----- Attribute(s)

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 error would impose *acceptable* limitation on the Answer. There are three complications in addressing this matter:

- error has *six* categories (listed overleaf at the top of page HL88.4);
- methods for assessing likely degree of error/severity of imposed limitation:

(*continued overleaf*)

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

A summary of assessment methods by error category (which are defined on the lower half of page HL88.7) is:

- **Study error:** assessment is based mainly on *extra*-statistical knowledge and is usually difficult to quantify.
- **Non-response error:** assessment is based on statistical theory seldom covered in introductory courses (see Note 8).
- **Sample error:** in introductory courses, assessment is based on statistical theory which yields a *confidence interval*.
- **Measurement error:** assessment is similar to the method for sample error.
- **Comparison error:** assessment is similar to the method for sample error.
- **Model error:** specific (often graphical) methods are used to assess (somewhat subjectively) modelling *assumptions*.

In keeping with their central roles in data-based investigating, the themes of categories of error, their assessment and their management recur throughout these Statistical Highlights and Course Materials. An illustration of limitation imposed by *sample* error on an Answer which is a *proportion* is given in the first bullet (●) on the upper half of page HL88.11,

The (unfamiliar) concept of 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.

Systematic discussion of the *management* of the six categories of error in statistics, together with illustrations, is provided in Statistical Highlights #6 to #22 – #6, #9, #10, #18 and #22 are expository.

∗ **Uncertainty:** incomplete information about error [usually about its size or magnitude].

– *Limitations* on Answers remind us of ever-present uncertainty – an Answer we anticpate will be 'correct' may actually be 'wrong' and an Answer expected to be 'wrong' may be 'correct', although disciplined and correct use of statistical methods can make the first of these possibilities *un*likely.

NOTES: 1. The discussion of the Formulation stage in Section 3 overleaf on page HL88.3 and above starts by defining 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 elements, 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 elements 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.

2. Our variate qualifiers 'response' and 'explanatory' are the ordinary English words that best evoke relevant statistical issues for variates; qualifier pairs used elsewhere (usually with 'variable') include:

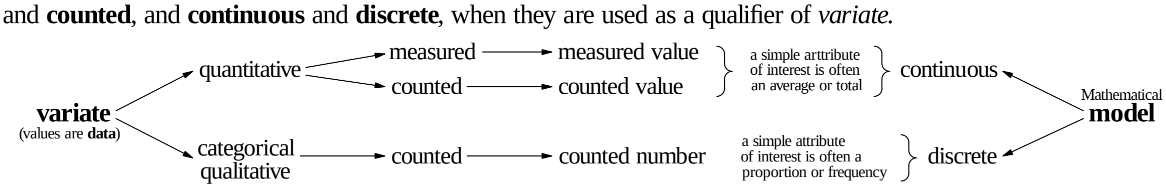
● *dependent* and *independent*; ● *outcome* and *predictor*; ● *outcome* and *exposure*.

No 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 – see, for instance, the entries for 'Independence' in the right-hand column of page HL91.9 of Statistical Highlight #91.

– Our definition of 'explanatory' makes it clear that a variate Ŧ whose change does *not* account, at least in part, for change in ¥ is *not* an *explanatory* variate and so need not be considered in the (already complicated) discussion of statistical relationships.

– *We* use 'variate' rather than 'variable' to avoid connotations that come with the latter from other disciplines.

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



● The quantitative *measured* and quantitative *counted* distinction blurs progressively as counted values become larger; it is also affected by the limited resolving power of real measuring processes.

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

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

Quantitative and qualitative are used more broadly elsewhere as adjectives to categorize 'research methods'.

(*continued*)

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 2)

NOTES: (cont.) 4. Elsewhere, our population attribute may be called a population **parameter**; *we* reserve 'parameter' for one of the components of a *statistical* (*mathematical*) *model* – see Appendix 3 on pages HL88.16 and HL88.17.

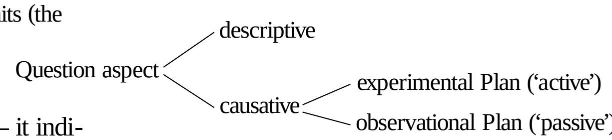
5. Equations (HL88.1) to (HL88.3) at the right below give symbolic expressions for three simple numerical attributes: an average, a (data) standard deviation and a proportion, where:
   - the (optional) subscript **¥** on an attribute indicates to which response variate the attribute refers – *generic* attribute notation is $a_{\text{¥}}(\mathbf{P}_{\text{Target}})$, where $\mathbf{P}_{\text{Target}}$ denotes the target population;
   - **¥**(*u*) is the value of the response variate for element *u*;
   - $\mathbf{N}_{\mathrm{T}}$ is the number of elements in the target population;
   - the summations run over all elements of the target population;
   - the proportion refers to *two* categories of units, C [for which **¥**(*u*) = 1] and $\overline{\mathrm{C}}$ [for which **¥**(*u*) = 0]. [Such use of a (binary) **indicator variate** makes the expression (HL88.3) for a proportion in two categories a special case of the expression (HL88.1) for an average.]

$$\text{average}_{\text{¥}}(\mathbf{P}_{\text{Target}}) = \frac{\sum_{\text{all } u} \text{¥}(u)}{\mathbf{N}_{\mathrm{T}}} \quad \text{-----(HL88.1)}$$

$$\text{standard deviation}_{\text{¥}}(\mathbf{P}_{\text{Target}}) = \sqrt{\frac{\sum_{\text{all } u} [\text{¥}(u) - \text{average}_{\text{¥}}]^2}{\mathbf{N}_{\mathrm{T}} - 1}} \quad \text{-----(HL88.2)}$$

$$\text{proportion}_{\mathrm{C}}(\mathbf{P}_{\text{Target}}) = \frac{\sum_{\text{all } u} \text{¥}(u)}{\mathbf{N}_{\mathrm{T}}} \quad \text{-----(HL88.3)}$$

6. The Question aspect is introduced as early as the *Formulation* stage because of the following dichotomy.
   ∗ **Experimental:** to be contrasted with *observational* – it indicates a comparative Plan where the *investigators* (*actively*) assign the value of the focal (explanatory) variate to each unit in the sample/blocks. [**Blocking** in an experimental Plan is forming groups of units (the **blocks**) with the *same* (or similar) values of one or more non-focal explanatory variates; units within a block are then assigned *different* values of the focal variate.]



   ∗ **Observational:** to be contrasted with *experimental* – it indicates a comparative Plan where, for each unit selected, the focal explanatory variate (*passively*) takes on its 'natural' value *uninfluenced* by the investigator(s). [See Section 8 on page HL88.13.]

   To answer a Question with a *causative* aspect, an *experimental* Plan is used, where feasible, to reduce the limitation on an Answer imposed by *comparison error*; what is *feasible* is determined by:
   - the nature of the focal variate – sex and age, for example, can*not* be assigned.
   - the resources available for the investigating – it is *feasible* to change people's diets but a *particular* investigation may not have the resources to do so;
   - what is *ethical* – cigarette smoking or a high-fat diet, for instance, can no longer ethically be assigned to participants.

   Comparison error and the experimental/observational Plan distinction are discussed in Statistical Highlight #9.

   Two words which evoke the essential difference between the two types of comparative Plan are the **active** assignment by the investigator(s) of focal variate values compared with the **passive** acceptance of its 'natural' values.

7. When using the FDEAC cycle in data-based investigating, it may be convenient to think of:
   - Formulation stage components as addressing matters dealing with *What .....?*
   - Design stage components as more typically addressing matters dealing with *How .....?*

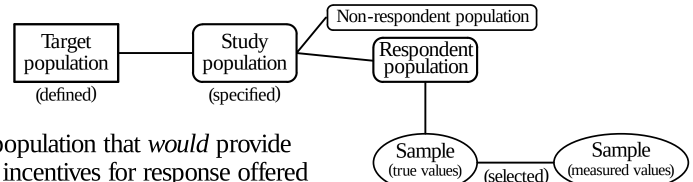
4. The Design Stage I: Three Populations and the Sample

The task of the Design stage is to develop a **Plan** for the data-based investigating that will answer the clear Question(s) that are the output of the Formulation stage. A danger to the Plan is that a desire to proceed with an investigation (*i.e.*, eagerness to undertake the Execution stage) makes it easy for too few resources to be committed to the Design stage.

The schema at the right below shows five groups of elements which *we* distinguish for data-based investigating; the last four are defined in the Design stage. [For convenience, the definition of the *target* population is repeated here.]

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

∗ **Study population**: a group of elements *available* to an investigation.



∗ **Respondent population:** those elements of the study population that *would* provide the data requested under the incentives for response offered in the investigation [such incentives arise predominantly when the elements are people, but *missing data* may also arise when elements are *in*animate].

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

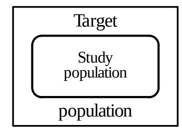
∗ **Sample:** the group of units/blocks selected from the respondent population *actually used* in an investigation – the sample is a *sub*set of the respondent population (as the vertical line in the schema overleaf at the bottom of page HL88.5 reminds us). [A **census** uses *all* the respondent (or study) population elements/units.]

∗ **Unit:** the entities *selected* for the sample; a unit may be one element (*e.g.*, a person) or *more than* one (*e.g.*, a household).

The element-unit distinction is discussed in more detail in Appendix 1 on pages HL77.8 and HL77.9 in Statistical Highlight #77.

The three terms *defined*, *specified* and *selected* in brackets () on the schema overleaf on page HL88.5 remind us of the (very) different processes of origin of two populations and the sample. Further, we recognize that the respondent and non-respondent populations, and differences between true and measured values, originate in human and measuring instrument falibility.

Just as target elements make up the target population, so study elements make up the study population. Target elements and study elements are often the same but *may* differ; for instance, when assessing drug efficacy and side effects using laboratory animals, target elements are humans but study elements are laboratory animals. The possibility of different target and study elements may be overlooked by showing the study population as a *subset* of the target population in a (misleading) diagram as at the right.



!

- Such a diagram may also show the *sample* as a subset of the study population (but see the schema at the bottom right of page HL88.6 overleaf and its extension at the top right of page HL88.8).

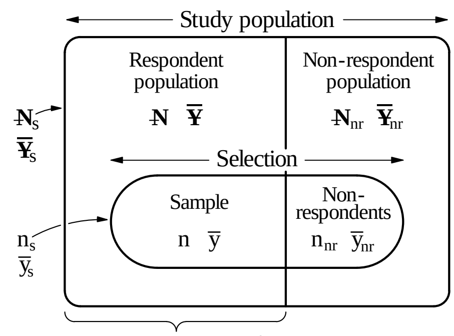
An investigation with a target *population* will have a study *population*; an investigation with a target *process* that is a set of operations will also have a study *population* – the available elements produced or affected by the process. An investigation with a target *process* that flows will have a study *process*, usually the target process over a restricted time period (see Table HL88.2 at the right).

Table HL88.2
Populations and Processes

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

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

– the $\mathbf{N}$'s and n's refer to *numbers* of elements/units;
– the $\overline{\mathbf{Y}}$'s and the $\overline{y}$'s are *averages* of an element/unit response variate $\mathbf{Y}$.



Statistical theory, particularly of sample surveys, is developed mainly in the context of the *respondent* population, often with*out* recognizing it explicitly

The relationships among the numbers of elements/units are:

Study population = Respondent population + Non-respondent population
$$\mathbf{N}_s = \mathbf{N} + \mathbf{N}_{nr}$$
Selection = Sample + Non-respondents
$$n_s = n + n_{nr}$$

When using the average, represented by the random variable $\overline{Y}$, of the sample selected by EPS as the estimator of the study population average, $\overline{\mathbf{Y}}_s$, the **non-response bias**, the model quantity representing non-response inaccuracy, is:

$$E(\overline{Y}) - \overline{\mathbf{Y}}_s \equiv \overline{\mathbf{Y}} - \overline{\mathbf{Y}}_s = \overline{\mathbf{Y}} - \frac{\mathbf{N}\cdot\overline{\mathbf{Y}} + \mathbf{N}_{nr}\cdot\overline{\mathbf{Y}}_{nr}}{\mathbf{N} + \mathbf{N}_{nr}} = \frac{\mathbf{N}_{nr}}{\mathbf{N} + \mathbf{N}_{nr}}(\overline{\mathbf{Y}} - \overline{\mathbf{Y}}_{nr}). \qquad \text{-----(HL88.4)}$$

[$E(\overline{Y}) = \overline{\mathbf{Y}}$ is established on page HL77.4 of Statistical Highlight #77.] To manage non-response error, the Plan for an investigation needs to include incentives for response that try to reduce one or both of the terms on the right-hand side of equation (HL88.4):

- the non-response rate, $\frac{\mathbf{N}_{nr}}{\mathbf{N} + \mathbf{N}_{nr}}$ (*i.e.*, the non-respondent population size as a proportion of the study population size), AND/OR:
- the difference in attribute values (*e.g.*, the averages) for the respondent and non-respondent populations.

NOTES: 8. Which elements of the study population fall in the respondent and non-respondent populations depends on the *incentives* offered for response – different incentives will presumably, in general, result in *different* sets of the study population elements in the two populations. For *given* incentives for response (as specified in the protocol for selecting units in a *particular* investigation), statistical theory to manage non-response error can be based on:

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

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

Complexities of modelling non-response is a reason why, as aluded to at the top of page HL88.4, trying to quantify non-response error is rarely covered in introductory statistics courses.

9. Incentives for response offered in sample surveys of human populations include:

- appealing to altruism by pointing out the benefits to the population of providing the data requested;

(*continued*)

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 3)

NOTES: (cont.) 9 • offering specific rewards such as a chance to win a substantial prize like a trip, or giving *all* units selected for the sample a small gift like a pen or a dollar coin.

10. The clean separation of respondents and non-respondents is an idealization – partial (or 'item') non-response is also encountered in practice when the elements/units investigated provide some, but *not all*, of the data requested.
    - The limitation imposed on Answer(s) by non-response error arises most commonly with a *sample* (*e.g.*, in a sample survey) but it is equally of concern in the less common type of investigation which tries to gather data from a **census** of *all* the study population elements/units.
      – Thus, 'sample' and 'census' *both* refer (strictly) to the elements/units of the *respondent* population.

11. The emphasis in Section 4 is on non-response among elements which are *people*; extra-statistical knowledge to deal with it includes questionnaire design and interviewing techniques (which may involve expertize in psychology).
    – We largely omit discussion of **equipment malfunction** as a source of missing data, where extra-statistical knowledge of the relevent measuring process(es) is likely to be needed.

12. Most presentations of introductory statistics involve our *respondent* (*not* study) population because of an *im*plicit assumption that the units selected for investigation (the 'sample') *all respond*.
    - The 'selection', denoting the sample plus the non-respondents, is a term *unique* to these Materials.

5. The Design Stage II: Error Sources as Backgruond to Developing the Plan

The three populations and the sample defined early in the Design stage of the FDEAC cycle reflect constraints that real-world conditions impose on data-based investigating.

- The study population/process *available* to investigators is rarely a perfect match to (the ideal of) the target population/process.
- Measuring equipment malfunction or human imperfection (in knowledge, cooperation, truthfullness) mean that the respondent population is routinely a source of *missing data/non-response* compared to what the study population, in principle, contains.
- Resource constraints commonly dictate using a *sample* from, rather than a *censu* of, the (respondent) population.
- Measuring processes for 'continuous' variate values seldom yield (exactly) correct results.

The last two matters, as *un*avoidable sources of *sample error* and *measurement error*, were foreshadowed as early as the introduction of Statistical Highlight #87 (and of Figure 1.1 of both the STAT 220 and STAT 231 Course Materials).

The difinitions of our six error categories, which formalize the constraints of real-world investigating, are as follows.

∗ **Study error:** the difference between [the (true) values of] the study population/process and target population/process attributes.

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

∗ **Sample error:** the difference between [the (true) values of] the sample and respondent population attributes.

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

∗ **Model error:** the difference between the model, together with its modelling assumptions, and the actual state of affairs in the real world; **modelling assumptions** in introductory courses are typically restricted to:
  ○ equiprobable selecting of units for the sample; ○ the form of the structural component of the response model;
  ○ the normality of each residual; ○ probabilistic independence of the residuals;
  ○ equal standard deviations of (response) variate values among different groups of elements or units.

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

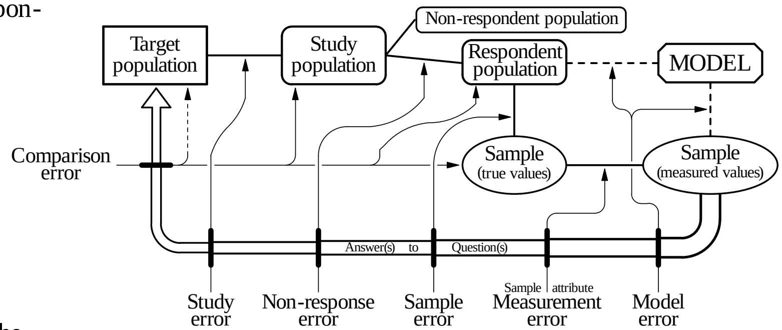
  The alternate wording of the last phrase of the definition of comparison error accommodates the equivalent terminologies of lurking variates and confounding; in a particular context, we use the version of the definition appropriate to that context:
  ○ 'lurking variates' can more readily accommodate phenomena like Simpson's Paradox – see Statistical Highlight #50;
  ○ 'confounding' is more common in the context of comparative Plans (see Statistical Highlight #63), but the variety of usage of 'confounding' can be a source of difficulty (see Statistical Highlight #3).

Study error, sample error, non-response error and comparison error are defined in terms of attributes of *groups* of elements whereas measurement error involves *individual* measurements – this is why the additional (sub)category of *attribute* measurement error is needed – see also the discussion involving equation (HL38.1) near the bottom of page HL38.5 in Statistical Highlight #38.

Restricting the discussion of model error for an introductory course is because model error generally is a large and difficult topic; recognizing the mofelling assumptions underlying a particular data-based investigation and assessing how well they are met is an *onerous* (and vital) task. Further discussion of modelling is given in Appendix 3 on pages HL88.16 and HL88.17.

(*continued overleaf*)

The schema at the lower right of page HL88.5 is given at the right below with four extensions.

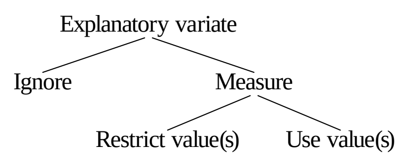


- The model is shown as a link between the respondent population and the sample.
- The six error category names are shown at the bottom and left, although 'measurement error' is really 'sample attribute measurement error.'
- The arrow rising from each error name shows the point of impact in the schema of that category of error;
- The broad arrow from the sample ellipse of measured values back to the target population represents Answer(s) to the Question(s) about the target population that sre *inferred* from measured sample data on response (and usually explanatory) variates.
  - The thick lines crossing this broad arrow at the arrows rising from each error category represent *limitations* inposed by error on Answer(s); the progressive *de*crease in width of the broad arrow after each error category reinforces this idea.
- ○ The four arrows arising from comparison error point to *boxes* representing *groups* of elements or units (a population or a sample) rather than, as for the other five error categories, to *lines joining boxes*; the comparison error arrow at the right is to be taken as pointing to *both* sample ellipses.
- ○ *Multiple* comparison error arrows are a consequence of its different manifestations in different Question contexts – for example, as summarized in Table HL60.1 at the upper right of page HL60.4 in Statistical Highlight #60.

Refining *response* variate(s) from the Formulation stage may be needed in the Design stage as a consequence of:

- ∗ turning a Question into a *clear* Question – when deciding on the 'best' web browser to buy, is the response variate the browser cost or the number of hits or the number of *relevant* hits obtained in response to a set of 'standard' search requests?
- ∗ considering what can practically be *measured* – what should be classified as a 'defect' in the paint of a new car or how can the change in taste of a supermarket food item be measured when investigating shelf life?

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 Statistical Highlights #89 (and #63).

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 with unacceptably limited overlap (*e.g.*, due to *missing* target population elements, *different* target and study population elements, too short a *duration* for the study *process*) with the target population, resulting in an *un*acceptable level of limitation imposed by study error.

NOTE: 13. The experimental/observational Plan distinction has implications for the expressions for *study* (or respondent) population *attributes* like those for the *target* population in equations (HL88.1) to (HL88.3) on page HL88.5 in Note 5. For an **average**, for instance:

- the (binary) focal variate $\mathbf{X}$ takes values x of 0 or 1, representing the two 'treatments';

$$\text{average}_{\mathbf{Y}|\mathbf{X}=\text{x}}(\mathbf{P}_{\text{Study}}) = \frac{\sum_{\text{all } u} \mathbf{Y}(u) \,|\, \mathbf{X}(u) = \text{x}}{\mathbf{N}_{\text{S}}} \quad \text{-----(HL88.5)}$$

(*continued*)

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 4)

NOTES: (cont.) 13.
- the vertical line is conditional probability notation and means *given that*;
- $\text{¥}(u)$ is the value of the response variate for element u when $\text{Ӿ}(u) = \text{x}$;
- ₦_{S} is the number of elements in the study population;
- summations run over all elements in the study population but:

$$\text{average}_{\text{¥}|\text{Ӿ}=0}(\mathbf{P}_{\text{Study}}) = \frac{\sum_{\text{all}\,u} \text{¥}(u)\cdot[1-\text{Ӿ}(u)]}{\text{₦}_{\text{S}} - \sum_{\text{all}\,u} \text{Ӿ}(u)} \quad \text{-----(HL88.6)}$$

$$\text{average}_{\text{¥}|\text{Ӿ}=1}(\mathbf{P}_{\text{Study}}) = \frac{\sum_{\text{all}\,u} \text{¥}(u)\cdot\text{Ӿ}(u)}{\sum_{\text{all}\,u} \text{Ӿ}(u)} \quad \text{-----(HL88.7)}$$

  - in equation (HL88.5) for an *experimental* Plan, *all* ₦_{S} elements contribute to *both* numerator and denominator and so, when *estimating* the two study population averages for $\text{Ӿ} = 0$ and $\text{Ӿ} = 1$, each sample provides infor- mation about the *entire* study population for the relevant value of the focal variate;
  - in equations (HL88.6) and (HL88.7) for an *observational* Plan, only the study *sub*population elements with a given value of Ӿ contributes to the relevant expression and so *neither* average (or its estimate) applies to the *entire* study population (or, at one stage removed, to the entire target population).

6. The Design Stage III: Further Background Terminology

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

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

∗ **Estimator:** a *random variable* whose distribution *represents* the possible values of the corresponding *estimate* under repetition of the selecting, measuring and estimating processes.

∗ **Limitations:** apply to Answer(s) to the Question(s) and must: + assess the likely importance of each category of error; + be expressed in the language of Question *context*.

∗ **Overall error:** the net effect of *all* relevant categories of error on the Answer(s) from an investigation.
For a Question with a **descriptive** aspect, the overall error is the sum of four error categories (see Statistical Highlight #18):
overall error = study error + non-response error + sample error + sample attribute measurement error. -----(HL88.8)

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

∗ **Variability:** the model quantity representing imprecision, which *we* distinguish from **variation**.

∗ **Inaccuracy:** *average* error (*i.e.*, its *systematic* component) under *repetition*.
- **Sampling inaccuracy:** average sample error under repetition of selecting and estimating.
- **Measuring inaccuracy:** average measurement error under repetition of measuring the *same* quantity.

∗ **Precision:** the inverse of *im*precision. ∗ **Accuracy:** the inverse of *in*accuracy.

∗ **Bias:** the model quantity representing inaccuracy. [Bias is discussed in more detail in Statistical Highlight #7.]

∗ **Replicating:** selecting more than one unit/block from the study population for the sample.
- **Adequate replicating:** selecting *just* enough units/blocks from the study population to make the likely magnitude of *sample* error [and, hence, the limitation it imposes on Answer(s)] *acceptable* in the Question context.

∗ **Repetition:** repeating over and over (usually *hypothetically*) one or more of the processes of selecting, measuring and estimating.
Repetition is inherent in our definitions of a confidence interval, an estimator, inaccuracy and imprecision.

∗ **Model:** a provisional idealized symbolic description of a real-world phenomenon.
We use Greek letters for **parameters** of statistical and probability models (*e.g.*, μ, σ and π).
In *any* situation where an Answer is based, in whole or in part, on a statistical model, we should bear in mind a maxim of the late Dr. George E. P. Box, a respected U.S. statistician: *All model are wrong, some are useful.*

∗ **Variation:** differences in (variate or attribute) values:
- across the individuals (*e.g.*, elements or units) in a group, such as;
  ○ a target population/process, ○ a study population/process, ○ a respondent population,
  ○ a non-respondent population, ○ a sample, and: ○ repeated measurements;
- arising under repetition [*e.g.*, for error or a sample average].

∗ **Lurking variate:** a non-focal explanatory variate whose differing distributions of values over groups of elements with different values of the focal variate, if taken into account, would meaningfully change an Answer about an Ӿ-¥ 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.

(*continued overleaf*)

7. The Design Stage IV: The Protocol for Selecting Units

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

Sampling (comprising the processes of selecting and estimating) is involved in most data-based investigations for two reasons:

∗ primarily because resources are seldom available to gather data from *all* the elements of the population of interest:
  - error associated with sample attributes as estimates of population attributes is usually an acceptable trade-off for the reduced cost of using a sample;
  - greater timeliness of data and Answer(s) from a (smaller) sample than from the (larger) population may also be advantageous;

∗ secondarily, sampling is the only option in the *un*common situation where the measuring process *destroys* the element being measured (*e.g.*, firing shot gun cartridges to assess the quality of their manufacturing process).

Three important matters about the sample used in any data-based investigation are:

○ how the sample was selected; ○ the sample size; ○ the non-response rate.

Our primary concern here is with the *first* of these matters; the usual terminology for the statistically desirable process is 'random selecting' but *we* use the more evocative **equiprobable selecting** (abbreviated '**EPS**'); we also discuss the *second* matter here.

There are many processes used in practice to select samples; three of them are discussed in this Highlight #88:

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

∗ **Equiprobable (simple random) selecting [EPS (SRS)]:** all samples of size n units from a (study) population of size $\mathbf{N}_S$ units have probability $1/\binom{\mathbf{N}_S}{n}$ of being selected.
  - What we call *equiprobable* selecting is likely to be called *simple random* (or *random*) selecting (or sampling) elsewhere.
  - *Equiprobable* refers to a *process*; we should be careful to avoid referring to an equiprobable (or random) *sample*.
  - The *definition* of EPS is in terms of *sample* selecting probabilities, not *unit* inclusion probabilities; consequences of this distinction for a sample of size n are:
    + under EPS, the inclusion probability is $n/\mathbf{N}_S$ for each unit in the study population;
    + even if the inclusion probability is $n/\mathbf{N}_S$ for each unit in the study population, the selecting process is *not necessarily* EPS – see Appendix 6 on pages HL88.19 and HL88.20;
    + the sample selecting process is *not* EPS if, for each study population unit, the inclusion probability is:
      ⊙ not equal to $n/\mathbf{N}_S$ OR: ⊙ not equal to that of all other unit(s).

  Refinement of the useage of 'EPS' and 'unit' are discussed in Note 28 on the lower half of page HL88.20 in Appendix 6.

∗ **Systematic selecting:** one unit is selected by EPS from the first k units of the study population ($k \ll \mathbf{N}_S$) and then every k*th* unit is selected.
  - Referring to the *first* k units of the study population implies an ordered (*e.g.*, alphabetic or numeric) list of these units; such a list (called a **frame**) may be real or conceptual (*e.g.*, a rule that would, if implemented, generate the list).
  - For convenience, it is usually assumed that $\mathbf{N}_S = nk$ so *all* 1-in-k samples selected systematically are of the *same* size n.

∗ **Judgement selecting:** human judgement is used to select n units from the $\mathbf{N}_S$ elements/units of the study population.

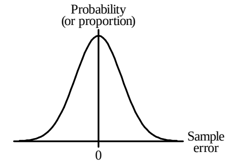
NOTE: 14. $\mathbf{N}_S$ is the number of *elements* in the study population; its use above for the number of *units* implies that sample units contain *one* population element, a common simplification for convenience in introductory courses.
  - In *cluster* sampling, when units contain *more than* one element, additional notation (beyond our current concern) is needed (see, for example, Appendix 1 on pages HL77.8 and HL77.9 in Statistical Highlight #77).

Statistics emphasizes EPS (or its equivalent), particulary in introductory courses, because it is the basis of statistical theory which provides (under repetition):

- an (inverse) relationship between sampling imprecision and *sample size* (or degree of *replicating*);
- *unbiased* estimating (*i.e.*, *zero* sampling inaccuracy) of a population *average* (an attribute commonly of interest).
- an expression for a *confidence interval* (CI) for a population average – such an interval, under suitable modelling assumptions, *quantifies* sampling and measuring imprecision – see Statistical Highlight #2.

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

- Also, a result from probability theory (the Central Limit Theorem) makes approximately *normal* (as illustrated at the right) the distribution of the averages of the set of all possible samples of a given size from the respondent population; as a consequence, *under EPS* there is a *higher* probability of selecting a sample with sample error of *smaller* magnitude, a *lower* probability of selecting one with *larger* sample error – see Statistical Highlight #74.



  - The *centre* (or 'average') of the (symmetrical) normal distribution in the diagram being at *zero* sample error is what is meant above by *unbiased* estimating – this matter is illustrated in Appen-

(*continued*)

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 5)

dix 1 starting at the bottom of page HL21.2 in Statistical Highlight #21 and also in Note 3 on page HL74.2 of Statistical Highlight #74.

EPS does *not*, of itself, *reduce* sample error or sampling imprecision, as implied in (wrong) statements such as:

- ○ EPS generates a *representative* sample – see Section 9 on page HL77.9 in Statistical Highlight #77; !
- ○ EPS generates a sample which provides a proper basis for *generalization*;

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

EPS provides for quantifying sampling imprecision and so, *in conjunction with adequate replicating* (or an *adequate sample size*), allows an Answer to be obtained with acceptable limitation imposed by sample error in the context of a particular investigation.

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

Experience showns that EPS is *the* process for selecting the sample to answer a Question with a *descriptive* aspect, and that sample error under judgement selecting usually imposes an *un*acceptable limitation on an Answer.

- For a Question with a *causative* aspect where EPS is often not feasible, as discussed in Statistical Highlight #83 there are circumstances under which there *can* be acceptable limitation imposed on an Answer by sample error from the judgement (or other) selecting process that is the *feasible* alternative to EPS in the investigation context.

There is further discussion in Statistical Highlight #21 of the reasons for the emphasis on EPS in introductory statistics courses.

NOTES: 15. A simple image of how EPS is implemented is to have, in a box, a slip of paper labelled for each element/unit in the study population; the $\mathbf{N}_S$ slips are *thoroughly* mixed and then n of them are selected without replacement – the labels of these n slips specify the units that will comprise the sample (or, in *our* terminology, the *selection*).

- In practice, EPS would usually be implemented with computer software that makes use of an *equiprobable digit* (or *random number*) *generator* – a source that is equally likely to generate any of the digits 0 to 9 at any position of a string of digits of specified length. Equiprobable digits are also available in printed tables. In this approach, the elements/units of the study population are usually thought of as being numbered (labelled) from 1 to $\mathbf{N}_S$.

16. The result of statistical theory (given on the lower half of the facing page HL88.10) which inversely relates (under repetition) sampling imprecision to (the square root of) sample size appears to be widely recognized, perhaps in part because it accords with intuition that an Answer from a 'large' sample is more likely to reflect the state of affairs in the population than an Answer from a 'small' sample. However, less widely appreciated (or more easily overlooked) is that the statistical theory is based on *EPS* of the sample. Hence, in proper statistical practice:

- investigator(s) must make clear, and users take note of, *how* a sample was selected;
- with a *non*-probability selecting process (*e.g.*, judgement selecting), we must forego invoking the sampling imprecision-sample size relationship;
- we should recognize that sampling *inaccuracy* has *no* necessary relationship to sample size – inaccuracy in 'large' samples may thus be more dangerous statistically than in 'small' samples, regardless of the selecting process;
  - *lack* of a 'number of instances-inaccuracy' relationship is also characteristic of the *other* five categories of error – for example, a ruler missing its first centimetre will yield length measurements one centimetre too high regardless of how many times it is used.

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

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

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

 (*continued overleaf*)

NOTES: (cont.) 17. who had the most recent birthday; this process implements (roughly) EPS of the eligible household members.

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

- ○ Because households have differing numbers of members, unit inclusion probabilities are *un*equal at the second stage; these *two* stages of equiprobable selecting therefore do *not* achieve EPS overall.
- ○ Because of non-response, many more (typically about *four* times as many) households need to be selected at the first stage as are required for the final sample size; for example, a national poll of 1,500 people may require around 6,000 telephone numbers to be generated, and some (or many) of these may have to be called multiple times to reach the eligible household member.

18. Limitation imposed by sample error on Answer(s) to Question(s) based on data from a sample selected by EPS can usually be *reduced* if there is prior information, exploited appropriately, about the study population. Examples involving the third option (see the tree diagram at the centre right on page HL888.8) for dealing with explanatory variates are:

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

Stratifying is used for two reasons.

- To make Answer(s) more useful by subdividing them by stratum; for instance, in Canada, the *national* unemployment rate needs also to be available by province and territory.
- To manage sample error (by decreasing sampling imprecision), provided the prior information about the study population allows 'homogeneity within strata, heterogeneity among strata' to be achieved.

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

Using prior information about the study population to reduce limitations imposed by error on Answer(s) is pursued in Appendix 5 on pages HL88.18 and HL88.19; more detailed discussion of ratio and regression estimating is taken up in later courses on survey sampling (*e.g*., STAT 332).

19. Systematic selecting is introduced in this Highlight #88 because it is commonly used in practice; however, *we* think of it as being *equivalent* to EPS by applying the restrictive assumption that the frame (from which every k*th* unit is selected for the sample) has the units arranged so any value of the response variate is equally likely to be anywhere on the list (an *equiprobably ordered frame* for a given response variate). Three illustrations are:

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

Thus, in this highlight #88, we consider two approaches to achieving equiprobability for selecting a sample:

○ via an equiprobable **selecting process**, applied to a frame in *any* order;

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

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

20. Other *named* methods of selecting units for the sample, which are largely omitted from our discussion, include:

- **accessibility selecting:** selecting units (easily) *accessible* to the investigator(s) – for instance, the *top* layer in a basket of fruit or a truckload of potatoes or the *front* pallets or cartons in a large stack in a warehouse;
- **convenience selecting:** selecting units that are *conveniently* available to the investigator(s) – for instance, people with a medical condition of interest who are at a hospital or clinic nearby to the investigator(s);
- **haphazard selecting:** selecting units with*out* (concious) preference by the investigator(s) – shoppers who pass

University of Waterloo W.H. Cherry

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 6)

NOTES: (cont.) 20. the location of an interviewer in a mall or rats in a cage which are more easily caught for a laboratory test;

- **quota selecting:** selecting units according to values of specified explanatory variates (like sex, age, income for human units) so the sample distribution of each variate will (approximately) match that of the study population;
- **volunteer selecting:** asking for (human) volunteers, usually after a brief explanation of what being in the investigation will entail for units of the sample.

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

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

Quota selecting is the same idea as *covering* defined in Note 22 below.

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

21.
- In *judgement selecting*, the *investigator's* judgement determines whether a unit is selected for the sample (or is in the group of units *not* selected);
- in *non-response*, a (human) *unit's* judgement determines whether the unit is in the respondent or the non-respondent population.

22. ∗ **Representative sample:** a sample whose sample error [and corresponding limitation imposed on Answer(s)] is *acceptable* in the Question context. [The representativeness of a sample can *rarely* be known.]

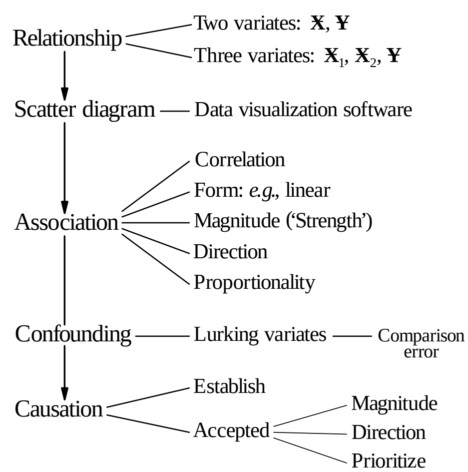
In contrast to this 17-word definition, Kruskal and Mosteller devote 50 pages to discussing the meanings of **representative sampling** in three articles in the *International Statistical Review*, **47**, 13-24, 111-127, 245-265 (1979). [UW Library call number HA 11.I505]

Because if its variety of (sometimes ill-defined) meanings in statistical contexts, use of 'representative' is best avoided in a sampling context in statistics – recall the two (wrong) statements cited near the top of page HL88.11; see also Section 9 on page HL77.9 in Statistical Highlight #77.

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

Covering is a guiding principle in judgement selecting; a greater degree of replicating makes it easier to apply. The benefit that (it is hoped) *may* accrue from covering has no *formal* basis in statistical theory.

8. The Design Stage V: Statistical Language of Relationships



The protocol for selecting units (or blocks) in Section 7 (starting on page HL88.10), is relevant to *any* use of the FDEAC cycle; the protocols for choosing groups and setting levels are relevant only when answering a Question with a *causative* aspect – that is, when investigating a *relationship*. We first define the language of (statistical) relationships (the 'vocabulary of connectedness') – the schema at the right shows a sequence for defining terms, connections among them, and types of (causative) Questions.

∗ A **relationship** in statistics is cast in terms of *variates* – there are *two* variates in the simplest case: an *explanatory* variate **X** (the *focal* variate) and a *response* variate **Y**.

- An **X**-**Y** relationship is the *connection* (if any) between *changes* in **X** and *changes* in **Y** (or in the *average* of **Y**).
  - \+ If (data establish that) **Y** remains *un*changed while **X** changes (or *vice versa*), there is *no* **X**-**Y** relationship, an idea of *un*connectedness captured by one sense of the word **independent**.
    - ○ We should recognize the distinction between the 'behavioural unconnectedness' of *independence* and the 'spatial separateness' captured by *disjoint*, as in 'disjoint events'.
- Data-based investigating of a relationship (to answer a Question with a causative aspect) involves **change** and **comparing** – these activities are therefore inherent in the components of an appropriate Plan.
  - \+ This is why experimental and observational Plans are described as **comparative**.

After *two* variates, the next level of complication is the relationships among *three* variates: *two* (focal) explanatory variates $\mathbf{X}_1$ and $\mathbf{X}_2$ and a response variate **Y**.

To limit the length of the overview of the FDEAC cycle in this Highlight #88, further discusion of the *many* statistical issues involved in investigating relationships is taken up in the following Statistical Highlight #89.

(*continued overleaf*)

9. The Design Stage VI: Measuring Process(es)

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

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

Like the detailed discussion of relationships, that of measuring processes is given elsewhere – see Statistical Highlight #38.

10. The Design Stage VII: Planning for the Execution and Analysis Stages

After the Plan has been developed, investigators turn their immediate attention to the Execution stage but the Design stage should include explicit consideration of *carrying out* the Execution and Analysis stages; logistical issues in managing an investigation include:

- ○ Selecting the sample: **who** will do it and **when**?
- ○ Measuring variates: **who** will do it and **when**?
- ○ How will **missing data** (*e.g.*, non-response in a sample survey) be accommodated?
- ○ What checks are in place to ensure the **Plan is followed**?
- ○ What might **go wrong**? How can such eventualities be managed?
- ○ Who should be **informed about**, and who might be **affected by**, the schedule for carrying out the Execution stage?
- ○ How will the **data be properly recorded** and appropriately **organized in a computer**?
- ○ Will the data and their analysis allow the **Question(s) to be adequately addressed**?
  [In the demanding task of developing the Plan, it is easy to lose sight of the Question(s)].
  - – It is useful (but unpopular) to make up and analyze, *prior to* the Execution stage, a (small) data set with the same structure as the (anticipated) real data.
  - – In a sample survey, a **pilot survey** (using up to 10%, say, of the final sample size) may be done.
- ○ Is the **model** appropriate, and are its **assumptions** likely to be met, in light of the Plan?
  - – The model may be needed in the Plan for information like the sample size which will meet cost or imprecision constraints.

NOTE: 23. Proper *recording* of data is one of Ishikawa's *Seven Tools* (discussed in Figure 11.18 of the STAT 221 Course Materials) and is commonly given too little attention; one characteristic of properly-recorded data is they are fully intelligible at a substantially later date to an investigator *other than* the person who recorded them. Necessary information *supplementary* to the data themselves includes:

- ● the person who did the measuring;
- ● the measuring instrument or gauge used;
- ● the time, date and place of measuring;
- ● the time order of the measurements;
- ● the measuring protocol used and any departures from it;
- ● documentation of incomplete or missing data.

11. The Execution Stage

The task of the Execution stage is to carry out the Plan *as developed*;

- ○ this is often difficult to accomplish; AND SO:
- ○ investigators should remain alert to *departures* from the Plan. We distinguish four components of the Execution stage:
- ∗ **Execute** the Plan: gather the data by carrying out the Plan as developed.
- ∗ **Monitor** the data as they are collected: data that differ appreciably from the actual value(s) of the variate(s) may arise for a variety of reasons; the best time to identify such data and correct them is *as they are obtained*. Monitoring the data aims to do this by checks and conventions such as:
  - – values out of range: 11.6 metres for a person's height, −2.5 for a *non*-negative variate;
  - – wrong symbol type: letters instead of numbers, numbers instead of letters;
  - – use symbols for *missing* data *different* from the symbols for data.
- ∗ **Examine** the data: more informal and earlier *monitoring* of the data at the time of collection merges into less informal and later *examining* the data (usually *after* return from the field), which in turn merges into informal methods of data *analysis* in the Analysis stage. Examining the data uses the same checks as monitoring, supplemented by numerical and graphical summaries that may indicate unexpected (and possibly erroneous) data; the data summaries start with *one* variate at a time and proceed to two or more variates.

(*continued*)

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 7)

∗ **Store** the data: three logistical issues are:
- Will the data be first recorded on paper or can they be entered directly into a computer?
  - + Direct computer entry, if feasible, reduces the number of data transfers and, hence, opportunities for mistakes.
- Which statisistical software package will be used in the Analysis stage?
  - + Choice of package may effect how the data need to be stored and notation conventions (*e.g.*, for missing observations).
- How will the long-term preservation of the data be ensured and, after publication of Answer(s) from the investigation, how will the data be made available to others for further investigation?

12. The Analysis Stage

Data *analysis* is the major component of most statistics courses; only an overview is provided here. We distinguish:

∗ **Informal** methods of analysis: *numerical* attributes
- measures of *location*: average, median, proportion, quantile, total;
- measures of *variation*: (data) standard deviation, interquartile range;
- measures of *linear association*: correlation, regression line;

graphical attributes
- *uni*variate: stemplot, histogram, dotplot, boxplot;
- *bi*variate: scatter diagram, quantile plot;

other numerical, tabular or graphical attributes.

∗ **Formal** methods of analysis: *confidence interval* (for a model parameter representing a population *attribute*);
prediction interval (for a random variable representing an *individual unit*);
test of significance (also called test of hypothesis);
other formal methods of data analysis.

Useful precepts of data analysis are:
- ○ a proper *Plan* makes data *analysis* more straight forward;
- ○ the *best* analysis *most simply* answers the Question(s) with *acceptable* limitations in the investigation context.

NOTES: 24. Undue emphasis on data *analysis*, particularly in *introductory* statistics courses, is *un*desirable; it can:
- obscure the *vital* role of the Formulation, Design and Execution stages in data-based investigating;
- imply (wrongly) that data *analysis* can compensate for inadequacies in Plan components;
- make understanding basic statistical precepts, which is important for an informed citizenry, seem beyond the capability of anyone but a specialist in the discipline.

25. Assessing **modelling assumptions** (from the perspective of an *introductory* course) is discussed in Table HL6.1 on page HL6.5 in Statistical Highlight #6.

13. The Conclusion Stage

The Answer to a Question must:
- ○ address the Question; ○ be intelligible to as wide an audience as is feasible;
- ○ be expressed in the language of the Question context; AND ALSO: ○ be expressed in statistical terminology.

∗ **Limitations:** apply to Answer(s) to the Question(s) and must:
- assess the likely importance of each category of error (*prioritizing* limitations from most to least serious is useful);
- be expressed in the language of Question *context*.

∗ **Recommendations**, if part of the investigation mandate, give the *broader* implications of the Answer(s) to the Question(s).

NOTE: 26. *Limitations* on an Answer remind us of the uncertainty *inherent* in incomplete information, arising most obviously in statistics from the processes of sampling and measuring. Consequences of this uncertainty are:
- 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*correctly on a sample of *one* unit may happen to be close to the value of the respondent (or even the study or target) population average.
  - A statement like *equiprobable (or 'random') assigning is neither necessary nor sufficient to establish causation* is true but unhelpful because it can obscure these two matters – see also Notes 1 (and 2) on pages 9.3 (and 9.4) of Statistical Highlight #9 for further discussion.

Experience shows it is difficult to develop a mind set in which these two matters are routinely recognized; the

(*continued overleaf*)

NOTE: 26. difficulty is compounded by the care needed to frame English statements that deal with uncertainty in statistics.
(cont.)
- 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.

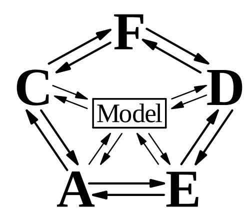
14. Appendix 1: Deming's Fourteen Points

W. Edwards Deming (mentioned in Section 1 on page HL88.1 in this Highlight #88), a poineer in the application of statistical methods to process improvement, summarized his broader management philosophy in his *Fourteen Points*; a version of them is given below. Deming's ideas are discussed in more detail in Figure 11.10 of the STAT 221 Course Materials.

1. Create and publish to all employees a statement of the aims and purposes of the company or other organization. The management must demonstrate constantly their commitment to this statement.
2. Learn the new philosophy, top management and the entire work force.
3. Understand the purpose of inspection, for improvement of processes and reduction of cost.
4. Cease doing business on price tag alone.
5. Improve constantly and forever the system of production and service.
6. Institute training and retraining of workers.
7. Teach and institute leadership.
8. Drive out fear. Create trust. Create a climate for innovation.
9. Optimize, toward the aims and purposes of the company, the efforts of teams, groups, staff areas.
10. Eliminate slogans, exhortations and targets for the work force.
11. Eliminate numerical quotas.
12. Give people a chance to take pride in their work.
13. Encourage education and self-improvement for everyone.
14. Do it!

15. Appendix 2: The FDEAC Cycle and the Model

The diagram at the right makes two *main* points about the FDEAC cycle:



- the *peripheral* arrows remind us each stage has implications for the stages before and after it; for example:
  - the Formulation stage may need to consider the type of Plan (*e.g.*., experimental or observational) and the Plan must have appropriate components to address the Question(s);
  - the Plan must specify the data which are required and the Execution stage must be sure the Plan *as developed* is carried out to generate such data;
  - the Execution stage generates the data for the Analysis stage and the Analysis stage must use modelling assumptions that can be justified in light of the Execution stage;
  - the Analysis stage must obtain from the data the information needed in the Conclusion stage to answer the Question(s) and the Conclusion stage must give Answer(s) that can be justified from (proper) data analysis;
  - the Conclusion stage must give Answer(s) which address the Question(s) with limitations acceptable in the Question context and the Formulation stage must have posed Question(s) for which such Answer(s) *can* be provided;
- the *radial* arrows remind us the (response) *model* has implications for the last *four* stages of the FDEAC cycle; specifically:
  - once we have clear Question(s), part of the Plan to answer these Question(s) is often an appropriate model (*e.g.*, for sampling and measuring processes, including calculating sample size – see Statistical Highlight #74);
  - the Execution stage generates data used to estimate model parameters representing study population attributes, and error in these estimates depends on how the data are generated (*e.g.*, imprecision and inaccuracy of measuring processes);
  - the Analysis stage may use model-based formal methods of data analysis; the model must then be appropriate for the method(s) of analysis that will obtain from the data the information needed to answer the Question(s);
  - the Conclusion stage must consider the model and its assumptions in assessing limitations on Answer(s) and the model must be chosen with Answer(s), and their acceptable limitations in the Question context, in mind.

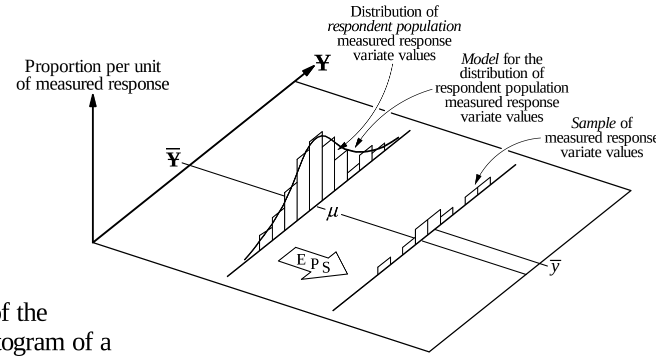
16. Appendix 3: Statistical Modelling [optional reading]

- **Response model:** a mathematical description, including modelling **assumptions**, of the relationship between a response variate and explanatory variate(s); the form of the relationship is contingent, in part, on the Plan for the investigation.
  - The **structural component** models the effect of specific explanatory variate(s) on the response variate.
  - The **stochastic component** models variation about the structural component.
- **Model parameter:** a constant (which *we* denote by a *Greek* letter) in a response model that *represents* a respondent population *attribute*; for example, μ represents $\overline{\mathbf{Y}}$ in the response model (HL88.9) below – see also the diagram at the top right of the facing page HL88.17 and the discussion to its left.

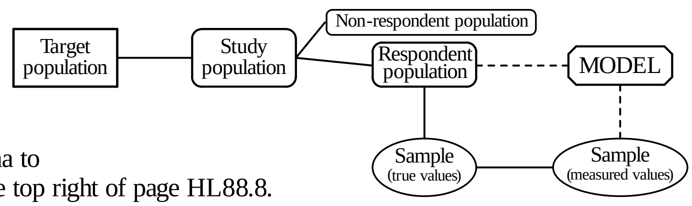
The four main response models discussed in STAT 231 [which include (HL88.9)] are summarized in Statistical Highlight #71; model symbols are defined in Statistical Highlight #72 and an overview of least squares estimating of model parameters is given in Statistical Highlight #73. [As the *simplest* of these models, the sructural component of HL88.9 is just a *constant* and so involves *no* explanatory variates; this model also arises near the middle of page HL77.1 in Statistical Highlight #77.]

$$Y_j = \mu + R_j, \quad j = 1, 2, ..., n, \quad R_j \sim N(0, \sigma), \quad \text{independent}, \quad \text{EPS}, \qquad \text{-----(HL88.9)}$$

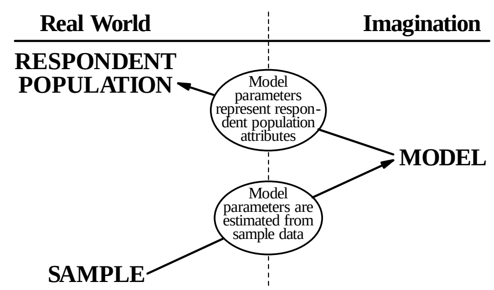
STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 8)



Model-based methods of analysis in statistics use data from a sample to *estimate* values of model parameters which then represent plausible values (in light of the data) for respondent population attributes and, hence, for Answer(s) to Question(s); we distinguish a *point* estimate from an *interval* estimate (defined near the top of page HL88.9). When the normal model is appropriate for the distribution of the response variate values, the model mean μ is estimated by the sample average $\overline{y}$ and σ is estimated by the sample standard deviation s – both *point* estimates. As illustrated at the right, we can think of the process of estimating μ by $\overline{y}$ and σ by s as approximating the histogram of a data set by the normal p.d.f. with the same 'centre' and same 'width' as the histogram.



The schema at the bottom right of page HL88.5 can be extended to include the model, as shown at the right; *our* view of the model as a link between the respondent population and the sample is elaborated further in the schema at the centre right of this page HL88.17 above Table HL88.2. The extension of the schema to clude the model and the six error categories is shown at the top right of page HL88.8.



All mathematical models are idealizations and are products of the intellect and the imagination. Even in the restricted context of an introductory course, the (largely graphical) methods of assessing the modelling assumptions are somewhat subjective, a difficulty that only *in*creases in the wider scope of the topic encountered more generally.

NOTE: 26. To maintain the distinction between the real world (represented by the data) and the model, we use different words – 'average' and 'mean' – for their measures of location; unfortunately, we do not have this option for the two measures of variation, which are both called 'standard deviation'. In the early stages of learning statistics, it is helpful to, at least mentally, add the respective adjectives 'data' and 'probabilistic' to distinguish the two uses of standard deviation. This terminology is summarized in Table HL88.2 at the right.

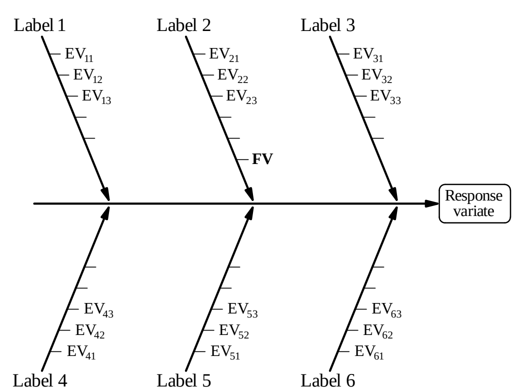
Table HL88.2

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

17. Appendix 4: Fishbone Diagrams for Comparative Plans

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

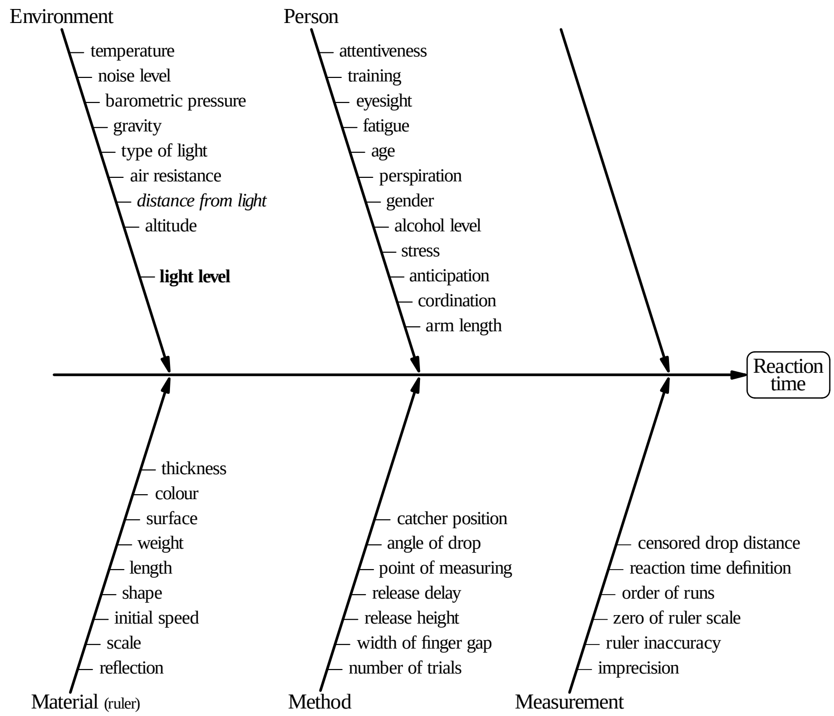
- a box on its right containing the name of the response variate,
- a central horizontal arrow pointing at this box,
- subarrows slanted from left to right and pointing either down or up at the central arrow;
  - each subarrow has a label evocative (in the investigation context) of a category of explanatory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



An example of a fishbone diagram is given overleaf at the upper right of page HL88.18; 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

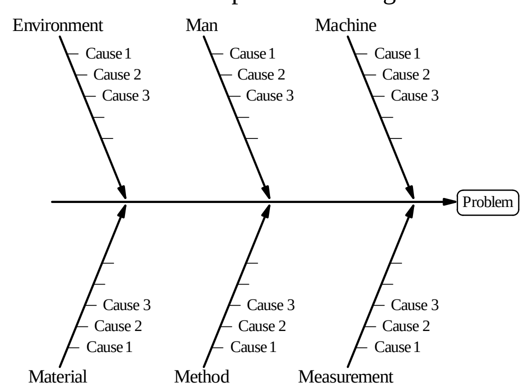
 (*continued overleaf*)

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: 27. 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.

18. Appendix 5: Broader Perspectives on the Protocol for Selecting Units – Clustering and Stratifying [optional reading]

Equiprobable (or simple random) selecting of units consisting of *individual elements* from an *un*stratified (respondent) population is useful for modelling the selecting process but, in practice, more complex sampling protocols are used. Two such protocols are:

∗ **cluster selecting:** selecting equiprobably units from the (respondent) population that are *groups* of elements – clusters may be of *equal* size (*e.g.*, cardboard boxes of 24 cans of soup) or *un*equal size (*e.g.*, households);

∗ **stratified selecting:** subdividing the (respondent) population into groups (called **strata**) so that elements with*in* a stratum have *similar* response variate values ('homogenity of strata') and elements in different strata *differ* as much as practicable from

continued)

STATISTICS and STATISTICAL METHODS: The FDEAC Cycle – An Overview (continued 9)

each other; the sample is obtained by equiprobable selecting of units consisting of individual elements from *each* stratum. [The element-unit distinction is discussed in Appendix 1 on pages HL77.8 and HL77.9 in Statistical Highlight #77.]

Example HL88.1 below illustrates the effects of *clustering* and *stratifying* on *sampling imprecision*, also bearing in mind that:

- *clustering* is commonly used because of the availability of a clustered *frame*, thereby avoiding the cost of generating a frame as part of the investigating – a **frame** can be thought of as a *list* of the (respondent) population sampling units (here, clusters);
- *stratifying* is commonly used because it provides (the often useful additional) *subdivision* of Answers by stratum.

Example HL88.1: A respondent population of $\mathbf{N} = 4$ elements (or units) has the following integer $\mathbf{Y}$-values for its response variate:

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

we examine the *sampling imprecision*, under equiprobable selecting (EPS) with a sample size of $n = 2$, of the random variable $\overline{Y}$, whose values are the sample average $\bar{y}$, as an estimator of $\overline{\mathbf{Y}}$, using three sampling protocols:

- EPS of two units, each consisting of one element, from the *un*stratified population;
- EPS of one *cluster*, of size $L = 2$ elements, from the *un*stratified population;
- EPS of one unit, consisting of one element, from each of two *strata* of size $\mathbf{N}_1 = \mathbf{N}_2 = 2$.

Note that *each* estimator is *un*biased, because $E(\overline{Y}) = \overline{\mathbf{Y}}$ or $E(\overline{Y}) - \overline{\mathbf{Y}} = 0$ [the *average* sample error is *zero*].

Table HL88.3
Unstratified population
EPS of two *elements*

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

Designation of sample error as *large*, *medium* or *small* is *only* for convenience in the context of Example HL88.1.

Table HL88.4a
Unstratified population
Clusters: [1, 2], [4, 5]
EPS of one *cluster* ($L = 2$)

Sample	$\bar{y}$	Error	
(1, 2)	1½	–1½	large
(4, 5)	4½	1½	large

Table HL85.4b
Unstratified population
Clusters: [1, 4], [2, 5]
EPS of one *cluster* ($L = 2$)

Sample	$\bar{y}$	Error	
(1, 4)	2½	–½	medium
(2, 5)	3½	½	medium

Table HL88.4c
Unstratified population
Clusters: [1, 5], [2, 4]
EPS of one *cluster* ($L = 2$)

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

Table HL88.5a
Stratified population
Strata: [1, 2], [4, 5]
EPS of one *element* per stratum

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

Table HL88.5b
Stratified population
Strata: [1, 4], [2, 5]
EPS of one *element* per stratum

Sample	$\bar{y}$	Error	
(1, 2)	1½	–1½	large
(1, 5)	3	0	small
(2, 4)	3	0	small
(4, 5)	4½	1½	large

Table HL88.6c
Stratified population
Strata: [1, 5], [2, 4]
EPS of one *element* per stratum

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

Example HL88.1 illustrates that:

- The effect on (sampling) imprecision of clustering and of stratifying depends on how each is *implemented* in the Plan – that is, it depends on this component of the *sampling protocol*.
- Clustering and stratifying affect imprecision by determining which of the *possible* samples of size n have non-zero selecting probabilities.
- Decreased imprecision is favoured by *heterogeneity of clusters* but by *homogeneity of strata* with respect to the response(s) of interest.
  - In the middle and right-hand columns of Example HL88.1, heterogeneity increases *down* the three clustered sampling protocols, homogeneity increases *up* the three stratified protocols.

[**As an exercise**, quantify the sample error *variation* by calculating the relevant (data) *standard deviation* for each of the seven sampling protocols; comment on what is illustrated by the values obtained.]

- There is a sense in which clustering is *passively* accepted in the interests of reducing investigation cost, whereas stratifying may be *actively* imposed by the investigator(s) on [or may be a natural feature of] the study (or respondent) population.
- While EPS from an *un*stratified population implies *equal* inclusion probabilities for all population *elements*, the converse does *not* hold – in the three clustered and three stratified sampling protocols, all *elements* have equal inclusion probabilities but all six *samples* of size 2 are *not* equally probable [four samples and two samples (respectively) have *zero* selecting probability].

19. Appendix 6: Sample Selecting and Unit Inclusion Probabilities – An Illustration [optional reading]

To illustrate the distinction (introduced in the discussion on page HL88.10 of EPS) between sample selecting and unit inclusion probabilities, and also ideas like clustering and stratifying, we consider a (respondent) population of $\mathbf{N} = 10{,}000$ elements and a sample of $n = 100$ elements, obtained using six protocols for selecting units (listed roughly in order of increasing complexity):

- equiprobable selecting of 100 units (elements) from the *un*stratified population;

(*continued overleaf*)

- ○ systematic selecting: selecting equiprobably 1 unit from the *first* 100 population units (elements) and then every 100*th* unit;
- ○ equiprobable selecting of 10 *clusters* of 10 elements from the population of 1,000 such clusters;
- ○ equiprobable selecting of 10 units (elements) from each of the 10 population *strata* each of $\mathbf{N}_h = 1{,}000$ elements $(h = 1, 2, ..., 10)$;
- ○ two-stage selecting: selecting 100 clusters equiprobably and then selecting 1 unit (element) equiprobably from each cluster;
- ○ two-stage selecting: selecting 2 strata equiprobably and then selecting 50 units (elements) equiprobably from each stratum.

Table HL88.6 below uses the symbol $\binom{\mathbf{N}}{n}$, the number of ways n items can be selected from $\mathbf{N}$ items if order of selecting is *un*-important. [This symbol and its use are discussed in Figure 7.5 of the STAT 220 Course Materials].

Relevant calculations for this illustration are summarized in Table HL88.6 below, where the six protocols are now listed in order of *de*creasing number of possible samples. The *short* names for the protocols in the second column of Table HL88.6 should generally be avoided because their brevity can (temporarily) obscure the nature of, and differences among, the protocols.

Table HL88.6

Protocol for selecting units	**Short name**	**Number of samples**	**Ratio to EPS**	**Selecting or inclusion probability: Sample**	**Selecting or inclusion probability: Unit**
EPS from an unstratified population	EPS	$\binom{10,000}{100} \simeq 6.5\times10^{241}$	1	1.5×10^{-242}	$\binom{1}{1}\binom{9,999}{99}/\binom{10,000}{100} = \frac{1}{100}$
2-stage EPS from a population in equal-sized clusters	2-stage cluster selecting	$\binom{1,000}{100}\binom{10}{1}^{100} \simeq 6.4\times10^{239}$	$\sim10^{-2}$	1.6×10^{-240}	$\binom{1}{1}\binom{999}{99}/\binom{1,000}{100}\bullet\binom{1}{1}\binom{9}{0}/\binom{10}{1} = \frac{1}{10}\bullet\frac{1}{10} = \frac{1}{100}$
EPS from a stratified population	Stratified selecting	$\binom{1,000}{10}^{10} \simeq 1.6\times10^{234}$	$\sim10^{-7}$	6.2×10^{-235}	$\binom{1}{1}\binom{999}{9}/\binom{1,000}{10} = \frac{1}{100}$
2-stage EPS from a stratified population	2-stage stratified slecting	$\binom{10}{2}\binom{1,000}{50}^{2} \simeq 4.0\times10^{171}$	$\sim10^{-70}$	2.5×10^{-172}	$\binom{1}{1}\binom{9}{1}/\binom{10}{2}\bullet\binom{1}{1}\binom{999}{49}/\binom{1,000}{50} = \frac{1}{5}\bullet\frac{1}{20} = \frac{1}{100}$
1-stage EPS from a population in equal-sized clusters	Cluster selecting	$\binom{1,000}{10} \simeq 2.6\times10^{23}$	$\sim10^{-218}$	3.8×10^{-24}	$\binom{1}{1}\binom{999}{9}/\binom{1,000}{10} = \frac{1}{100}$
1-in-100 systematic selecting from an unstratified population	Systematic selecting	$100 = 10^2$	$\sim10^{-240}$	$\frac{1}{100} = 10^{-2}$	$\frac{1}{100}$

The last four columns of (sometimes approximate) *numerical* table entries are, for each of the six protocols:

- ⊙ the number of samples that can be selected; *i.e.*, the size of the set of all possible samples;
- ⊙ the ratio of the number of samples a protocol can select to the number for EPS from an unstratified population;
- ⊙ the probability any *sample* is selected; here, the *reciprocal* of the number of samples [but see the first comment (+) below];
- ⊙ the probability any *unit* is included in the sample.
  - – In contrast to the *extreme* variation (over nearly 240 orders of magnitude) of the *sample* selecting probabilities among the protocols, the six *unit* inclusion probabilities are *all* 1 in 100 in this illustration (the *equality* of these six probabilities is a characteristic of this illustration, *not* a general result).
    - + Use of EPS at the one or both stages of each protocol means that, in *this* illustration, all *samples* the protocol can select are *equally* likely; as a consequence, the *sample* selecting probability for each protocol in the fifth ('Sample') column of Table HL88.6 above is the *reciprocal* of its number of samples [but see the first comment (○) in Note 17 near the top of page HL88.12]. Thus, from the perspective of *samples*, the protocols are:
      - ○ *alike* in having their possible samples *equi*probable; ○ *different* in their numbers of possible samples.
    - + Although *calculations* for *unit* inclusion probabilites for the one-stage and the two-stage clustered and stratified protocols have the *same* structure, they yield vastly different numbers of samples; there are also other important *statis*--*tical* distinctions between clustered and stratified protocols – see Example HL88.1 overleaf on page HL88.19.

NOTE: 28. EPS from an unstratified population yields the (exhaustive) set of all *possible* samples of a given size from a population of a given size; this set contains about 6.5×10^{241} samples when $\mathbf{N} = 10{,}000$ (elements or) units and $n = 100$ units.

- ● Each of the other five sampling protocols can select only a *sub*set of this (exhaustive) set of samples.
  - – These five protocols are useful because EPS from an unstratified population can rarely meet Plan requirements.
  - – When these protocols are properly implemented, they *preferentially* exclude samples with an extreme value for an attribute like an average, thus *de*creasing sampling imprecision (*e.g.*, see Statistical Highlight #21).
- ● As well as yielding all possible samples, EPS is emphasized in introductory discussions because it is:
  - – involved in more practically useful protocols like the last five in Table HL86.6 above on this page HL88.20;
  - – the basis of sampling theory for quantifying the behaviour of sample error under repetition, *i.e.*, for quantifying sampling imprecision – see Statistical Highlight #21.

Thus, we need to distinguish:

- ∗ **EPS from an unstratified population**: a protocol for selecting units which is seldom used in practice but which is the basis of sampling theory; FROM:
- ∗ **EPS** (unqualified): *part* of a protocol for selecting units which involves *other* statistical ideas like stratifying and/or clustering and/or systematic selecting – this is the more *common* usage of 'EPS'.

SOURCE: MacKay, R.J. *Experimental Design and Sampling.* Course Notes for Statistics 332/362, University of Waterloo, Fall, 2005, page VII – 1.