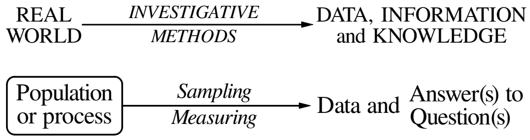
Figure 2.2. DATA-BASED INVESTIGATING: Error Categories and Management

1. Background I – What is Statistics About? [optional reading]



As summarized in the two schemas at the right, statistics is concerned with *data-based investigating* (or empirical problem solving) of the real world, which means investigating some population or process on the basis of *data* to *answer* one or more *questions.* For this introductory discussion, only the investigative methods of *sampling* and *measuring* are shown in the lower schema.

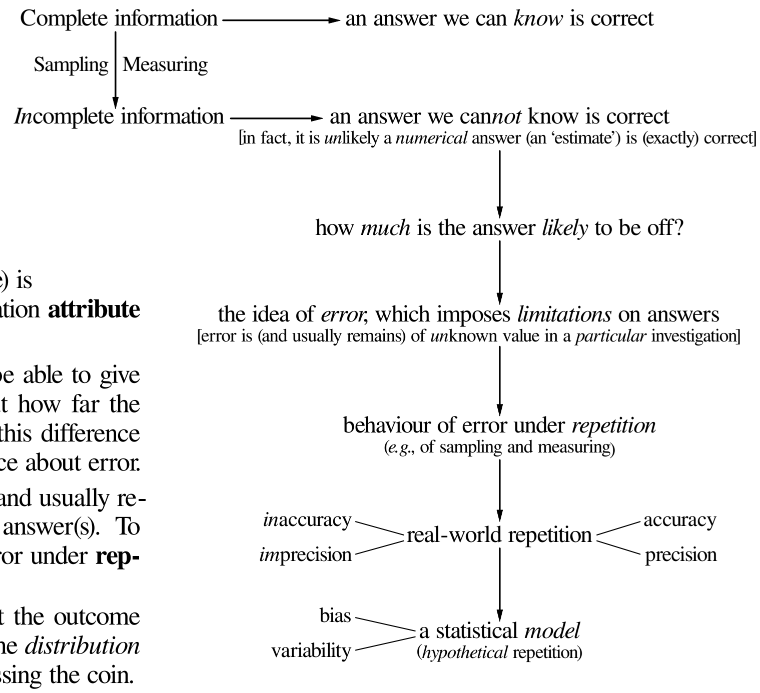
This introduction presupposes initially that data-based investigating is concerned with answer(s) to question(s) about a collection of **elements** which comprise a **population**. For example:

- ○ a computer manufacturer may want to assess warranty claims for one of its products – an *element* would be one item of the product and the *population* is all such items sold over some specified period;
- ○ a government department may want to know the proportion of Canadian adults who have concerns about the level of funding of the health care system – an *element* would be a Canadian adult and the *population* is all such Canadians;
- ○ a financial institution may wish to find people whose profile makes them more likely to accept an unsolicited offer of a credit card – an *element* would again be a person but the *population* is harder to specify; it could be people whose profile puts them at or above an acceptance level deemed likely to make the credit card offering profitable for the financial institution.

The question(s) to be answered may *sometimes* be concerned with an *individual* to be selected in future from the population.

2. Background II – Key Ideas: Uncertainty, Error and Repetition [optional reading]

A central issue in data-based investigating is that external constraints and the type of information we require *impose* sampling and measuring processes on most investigating; we then need to assess the likely 'correctness' of the answer from the investigating in light of the *uncertainty* introduced by these two processes. In essence, this is the problem of **induction**. An overview of this matter is given in the schema at the right below; the main ideas are summarized at the left.


Complete information
an answer we can *know* is correct
Sampling
Measuring
*In*complete information
an answer we can*not* know is correct
[in fact, it is *un*likely a *numerical* answer (an 'estimate') is (exactly) correct]
how *much* is the answer *likely* to be off?
the idea of *error*, which imposes *limitations* on answers
[error is (and usually remains) of *un*known value in a *particular* investigation]
behaviour of error under *repetition*
(*e.g.*, of sampling and measuring)
*in*accuracy
*im*precision
real-world repetition
accuracy
precision
bias
variability
a statistical *model*
(*hypothetical* repetition)

- ∗ If we have *complete* information, we can obtain a *certain* answer; that is, an answer we can *know* is correct.
- ∗ If we have *in*complete information, we can*not* know an answer is correct (an *un*certain answer) – in fact, it is *un*likely a *numerical* answer (an 'estimate') is (exactly) correct;
  - – sampling and measuring yield data (and, hence, information) that are *inherently* incomplete.
- ∗ An answer which is a *number* (like a sample average) is called a **point estimate** of the corresponding population **attribute** (the population average).
- ∗ To make such an answer more useful, we want to be able to give an **interval estimate**, a quantitative statement about how far the point estimate is *likely* to be from the true value – this difference is the **error** of the estimate. **Uncertainty** is ignorance about error.
- ∗ In a *particular* investigation, the size of the error is (and usually remains) *un*known; this is reflected in **limitations** on answer(s). To quantify uncertainty, we turn to the behaviour of error under **repetition** of the sampling and measuring processes;
  - – an analogy is tossing a coin – we cannot predict the outcome of a *particular* toss but *probability* can describe the *distribution* of outcomes under *repetition* of the process of tossing the coin.
- ∗ In the classroom, we can do *actual* repetition (repeating over and over) to demonstrate, for example, the statistical behaviour of the values of sample attributes (*e.g.*, averages) and of the values which arise from measuring processes;
  - – the two characteristics of *sign* and *magnitude* of (numerical) error lead, under repetition, to what we call **inaccuracy** and **imprecision**; the *inverses* of these two ideas – **accuracy** and **precision** – provide more familiar terminology, but we must realize that statistical methods (try to) manage the (undesirable) *former* to achieve needed levels of their (desirable) inverses.
- ∗ *Out*side the classroom, we recognize that *actual* repetition is usually not a viable option – we use instead *hypothetical* repetition based on an appropriate statistical **model** (for the selecting and measuring processes, for example);
  - – *we* help maintain the distinction between the real world and the model by using **bias** and **variability** as the *model* quantities which represent inaccuracy and imprecision in the real world.
- ∗ A statistical model, including its *assumptions*, allows us to achieve our aim of quantifying (some of) the uncertainty in our

(*continued overleaf*)

- ∗ estimate(s); part of identifying the limitations on an Answer is to assess model error arising from the difference between ('idealized') modelling assumptions and the real-world situation. [Model error is pursued in Appendix 2 on page 2.32.]

There is further illustration in Figure 1.1 of these Course Materials of the ideas presented overleaf on page 2.27.

3. Background III – Some Terminology [A more detailed Glossary is given in Figure 2.17.] [optional reading]

∗ **Aspect:** a binary categorization of the primary concern of a Question, identified in the Formulation stage of the FDEAC cycle.
- **Descriptive:** a Question whose Answer will involve primarily values for *population attributes* (past, present, future).
- **Causative:** a Question whose Answer will involve primarily whether and/or how the focal explanatory variate is *causally* related to the response variate in a population/process.

∗ **Population:** a (finite) well-defined group of *elements* ***other than*** the sample. [An *infinite* 'population' is a (sometimes useful) *model*.]

∗ **Element:** the entities that make up a population; for example, a person is an element of the population of Canadians, but we recognize that many populations in data-based investigating have non-human or *in*animate elements.

∗ **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 page 8.56 in Figure 8.11 in these Course Materials.

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

Thus, in statistics, a process of the first type involves *elements*. WHEREAS:

In *probability*, a process is any set of *operations* from which there are at least two possible *outcomes*; observing *which* outcome occurs when the process is executed generates *data*, which may yield values for probabilities associated with, or other characteristics of, the process. [The population-process distinction is discussed in Statistical Highlight #94..]

∗ **Sampling:** includes the processes of *selecting* and *estimating*.

∗ **Measuring:** the process used to determine the value of a variate.

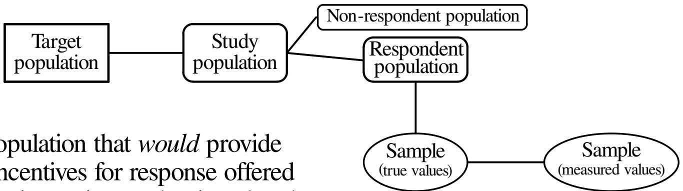
∗ **Induction:** reasoning from *particular* cases, investigations or data, to a more *general* Answer, using logic and relevant theory.

The schema at the right below shows five groups of elements which *we* distinguish for data-based investigating (the **model** is discussed in Appendix 2 on page 2.32); further terminology relevant to this Figure 2.2 follows the definitions of these groups.

∗ **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 – see also Note 1 on the facing page 2.29.

∗ **Sample**: the group of units selected from the respondent population *actually used* in an investigation – the sample is a *subset* of the respondent population (as the vertical line in the schema above reminds us).

∗ **Variate:** a characteristic associated with each *element* of a population/process.
- **Response variate (¥):** a variate defined in the Formulation stage of the FDEAC cycle; an Answer gives (information about attribute(s) of the response variate over the target population/process.
- **Explanatory variate (Ƶ):** a variate, defined in the Formulation stage of the FDEAC cycle, that accounts, at least in part, for changes from element to element in the value of a response variate.
- **Focal (explanatory) variate (Ӿ):** for a Question with a *causative* aspect, the focal variate is the *explanatory* variate whose relationship to the *response* variate is involved in the Answer(s) to the Question(s).

∗ **Attribute:** a quantity defined as a function of the response (and, perhaps, explanatory) variate(s) over a *group* of elements (or units), typically, the: – target population, – study population – respondent population, – non-respondent population, – sample.
Familiar (simple numerical) attributes are averages, proportions, medians, totals and s.d.s; the importance of attributes is that:
+ Answer(s) to Questions(s) are usually given in terms of attributes, often their values;
+ five of the six categories of *error* are defined in terms of attributes – see the upper half of page 2.30.

∗ **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.*

(*continued*)

Figure 2.2. DATA-BASED INVESTIGATING: Error Categories and Management (continued 1)

∗ **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.*

∗ **Error:** *Overall* error in an investigation refers to the net effect of *all* relevant categories of error on the Answer(s) from the investigation – see Section 4 which starts at the bottom of this page 2.29.

∗ **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** (defined below).

∗ **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 on page 8.58 in Figure 8.11.]

∗ **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 (*e.g.*, μ, σ and π) for **parameters** of statistical and probability models.
In *any* situation where an Answer is based, in whole or in part, on a mathematical model, we should bear in mind a maxim of the late Dr. George E. P. Box, a respected U.S. statistician: *All models 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 (**Ƶ**) whose differing distributions of values over groups of elements or units with different values of the focal variate, if taken into account, would meaningfully change an Answer about an **X̶-Ɏ** relationship.

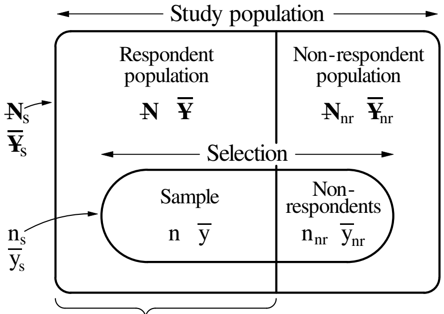
∗ **Confounding:** differing distributions of values of one or more *non*-focal explanatory variate(s) among two (or more) groups of elements or units [like (sub)populations or samples] with different values of the focal variate.

NOTE: 1. 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 N̶s and ns refer to *numbers of elements or units*;
- the $\overline{\text{Ɏ}}$s and $\overline{\text{y}}$s are *averages* of a response variate **Ɏ** of the elements or units.

The relationships among the numbers of elements or units are:

Study population = Respondent population + Non-respondent population
N̶$_{\text{S}}$ = N̶ + N̶$_{\text{nr}}$

Selection = Sample + Non-respondents
$\text{n}_{\text{s}} = \text{n} + \text{n}_{\text{nr}}$


Study population
Respondent population
N̶ $\overline{\text{Ɏ}}$
Non-respondent population
N̶$_{\text{nr}}$ $\overline{\text{Ɏ}}_{\text{nr}}$
N̶$_{\text{S}}$
$\overline{\text{Ɏ}}_{\text{S}}$
Selection
Sample
n $\overline{\text{y}}$
Non-respondents
n_{nr} $\overline{\text{y}}_{\text{nr}}$
n_{s}
$\overline{\text{y}}_{\text{s}}$

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

4. Six Categories of Error in Data-based Investigating [The title matter of this Figure 2.2.]

We identify six categories of error (defined overleaf on page 2.30) which can contribute to overall error defined above:

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

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

(*continued overleaf*)

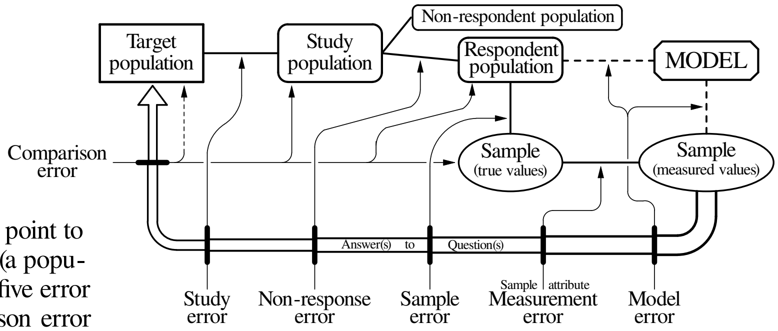
manage the six categories of error are summarized in Table 2.2.1 below and on the facing page 2.31.

- ∗ **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 and 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 **X-Y** relationship that is based on comparing attributes of groups of elements or units with different values of the focal variate(s), comparison error is the difference from the *intended* (or *true*) state of affairs arising from:
  - – differing distributions of lurking variate values between (or among) the groups of elements or units OR – confounding.

  The alternate wording of the last phrase 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 – discussed in Figure 9.8;
  - ○ 'confounding' is more common in the context of comparative Plans (discussed in Section 8 in Figure 9.1), but the variety of useage of 'confounding' can be a source of difficulty (discussed in Figure 9.9).

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 (2.15.1) near the bottom of page 2.71 in Figure 2.15.

The schema at the centre right of page 2.28 has been extended, as shown at the right, to include the model (as a link between the respondent population and the sample) and the effect of our six categories of error in imposing limitations on Answers obtained by data-based investgating.



- ○ The arrow rising from each error category shows its point(s) of impact.
- ○ 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 9.12.2 near the middle of page 9.76 in Figure 9.12.

Table 2.2.1

Plan Component			**Error category**	**Error Management Strategy**
Specifying the study population/process			Study	Specify the study population/process so its attribute(s) can be anticipated to be adequately close (in value) to those of the target population/process. ● *Restricting* values of explanatory variates can reduce variation in the study population/process – this may *de*crease sample error but *in*crease study error. ○ Sample error is preferred because statistical methods to manage it are better defined than the extra-statistical knowledge usually needed to manage study error.
Selecting units	Method of selecting	EPS	Sample	**EPS** is the basis of sampling theory which provides for: ● unbiased estimating of the respondent population average by the sample average; ● quantifying the likely size of sample error under repetition [*i.e.*, quantifying sampling imprecision, which *we* take here as 'quantifying uncertainty'].
		Judgement		**Judgement selecting** aims to make sample error as small as needed in the context of the *particular* investigation. ● It provides no basis for assessing if this aim has been achieved.
	Sample size (Replicating)	EPS		**EPS**: Sampling imprecision *de*creases with *in*creasing sample size (see Figure 2.16).
		Judgement		**Judgement selecting:** increasing sample size usually decreases the difficulty of making sample error as small as needed in the Question context. BUT: ● There is no theoretical basis which relates sample size to sampling imprecision.
	Stratifying the respondent population			Decreases sampling imprecision under EPS from the (properly-chosen) strata. ● Provides attribute estimates for the strata as well as for the respondent population.

(*continued*)

Figure 2.2. Error Categories and Management (continued 2)

Table 2.2.1 (continued)

Plan Component		Error category	Error Management Strategy
Obtaining responses	Incentives: Questionnaire Interviewer Call-backs Other	Non-response	Apart from a possible *legislated* requirement to respond (*e.g.*, to a population census), obtaining responses from units which are humans relies on *incentives* which include: ● a clear, answerable, succinct questionnaire; ○ when feasible, a questionnaire on *one* sheet of paper (or equivalent) is an advantage; ● properly trained interviewers; ● call back to units until those who are *un*available *are* contacted; ● appeal to altruism – respond to provide information that will benefit society; ● offer a material reward for response: ○ give *every* respondent a small item like a pen or a dollar coin; ○ offer respondents a chance to win a substantial prize like a trip. The skill and persistence of interviewers, developed by training, are a component of the incentives – see also Note 4 on the upper half of page 2.73 in Figure 2.15). The clean separation of respondents and non-respondents is an idealization – partial (or ‘item’) non-response is also encountered in practice when sampled units provide some, but *not all*, of the information requested.
Imputing			**Imputing** is the process of assigning values for missing observations – *e.g.*, assigning a value for the response of a non-respondent on the basis of its values for known explanatory variates (like sex, age, location) that (it is hoped) are reasonable ‘predictors’ of the response variate. ● The purpose of imputing is to simplify the data analysis; it *rarely* meaningfully increases the completeness of the information in the data.
Measuring variates	Imprecision Inaccuracy	Measurement	Use a measuring process whose inaccuracy is acceptable in the Question context. ● Inccuracy of a measuring process does *not* necessarily decrease as its *cost* increases. ○ Inaccuracy is managed by using standards (where they exist – see page 2.71 in Figure 2.15) to *calibrate* the measuring process. Use a measuring process whose imprecision is acceptable in the Question context. ● Decreased imprecision for a measuring process usually entails a more *costly* process but the converse is *not* always true.
Estimating attribute values	Simple Ratio Regression	Sample	Under EPS, the sample average and sample standard deviation provide estimates, with defined behaviour under repetition, of the corresponding respondent population attributes. When estimating the respondent population average or total under EPS, ratio and regression estimating improve the (simple) estimate by using the respondent population average or total of an *explanatory* variate with a (strong) positive association with the response variate whose attribute is of interest. ● Ratio estimating decreases sampling imprecision when the standard deviation of the response variate increases linearly with the square root of the explanatory variate. ● Regression estimating decreases sampling imprecision when the standard deviation of the response variate does not change with the value of the explanatory variate. Ratio and regression estimating introduce *estimating bias* but can have smaller rms error than $\overline{Y}$ or $\mathbf{N}\overline{Y}$ as the estimator of the respondent population average or total.
Assessing modelling assumptions	EPS Form of the structural component Normality Probabilistic independence Equal standard deviations	Model	Limitations imposed by model error from two modelling assumptions are managed by: ● ensuring the selecting process for units *is* (equivalent to) *EPS*; ● ensuring variate values are measured *independently.* Assessing how well modelling assumptions appear to be met usually involves graphical displays (*e.g.*, scatter diagrams) of the estimated residuals from the response model; ● use a normal quantile plot (or, sometimes, a histogram) to assess nomality; ○ transforming (*e.g.*, taking logarithms of) the data can help meet this assumption; ● use a plot in the time order of data collecting to assess probabilistic independence. ● use side-by-side dot- or boxplots, or a plot with the explanatory variate from the structural component of the model on the horizontal axis, to assess equality among, or dependence on an explanatory variate of, standard deviation(s).
Question with a causative aspect	Experimental Plan	Comparison	● **Blocking:** forming groups of units with the *same* values of one or more non-focal explanatory variates; the units within a block are then assigned *different* values of the *focal* variate. ● **Equiprobable assigning:** a *probabilistic* mechanism used to assign the value of the focal explanatory variate to the units: – within each block in a blocked Plan; – in the sample in an *un*blocked Plan. ○ **Blinding participants and treatment administrators:** by withholding from participants and treatment administrators knowledge of which group a participant is in, these two blindings try (like *equiprobable assigning*) to manage factors which may promote differences in averages of unknown and unmeasured non-focal explanatory variates in the (treatment and control) groups whose (average) response variate is being compared. [Management of **comparison** error.] ⊙ **Blinding treatment assessors** tries (like making measurements *independent*) to prevent the assessors’ other knowledge from improperly influencing their assessment of participants’ health status. [Management of **measurement** error.]
	Observational Plan		● **Matching:** forming groups of units with the *same* values of one or more non-focal explanatory variates but *different* values of the *focal* variate. ○ **Subdividing:** a form of matching in which the values of the response variate for the units of the sample, and the attribute they are used to estimate, are *subdivided* on the basis of the values of a non-focal *explanatory* variate that may be *confounded* with the focal variate under the Plan (see Table 9.6 on page 9.15 in Figure 9.1).

(*continued overleaf*)

5. Appendix 1 – Overall Error for Questions with a Descriptive Aspect

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

overall error = study error + non-response error + sample error + sample attribute measurement error. -----(2.2.1)

The schema at the lower right shows pictorially, when estimating an *average* to answer a Question with a descriptive aspect, the breakdown of overall error given in equation (2.2.1); symbols are defined in Table 2.2.2 at the right. Licence on two matters improves the clarity of the diagram:

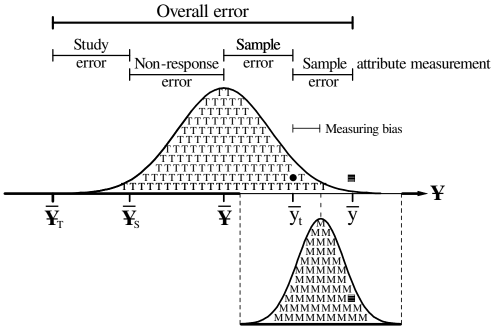
- ○ all four error components are *positive* – in practice, overall error may involve some *cancellation* among error components of *opposite* sign;
- ○ the distribution of *measured* sample attribute values has been moved *down*.

Other matters about the diagram are:

- ⊙ true (T) and measured (M) values of a sample attribute (here, an *average*), under repetition of the selecting, measuring and estimating processes, have each been modelled by a *normal* distribution;
- ⊙ the value of the *true* average of the sample *selected*, from among the set of all possible samples, is represented by the black filled circle (•);
- ⊙ the value of the *measured* sample average, from among the set of all possible such values, is represented by the black filled square (▪);
- ⊙ there is *no* sampling *bias* – the mean of the sampling distribution is $\overline{\underline{\mathbf{Y}}}$;
- ⊙ there *is* measuring *bias* – the horizontal distance between $\overline{y}_t$ and the long-term *average* (the *mean* of the distribution) of its *measured* values;
- ⊙ sampling variability is larger than measuring variability – the standard deviation of the distribution of the Ts is larger than that of the Ms.

Table 2.2.2: SYMBOL DEFINITIONS

$\underline{\mathbf{Y}}$	Response variate
$\overline{\underline{\mathbf{Y}}}_T$	(True) target population average
$\overline{\underline{\mathbf{Y}}}_S$	(True) study population average
$\overline{\underline{\mathbf{Y}}}$	(True) respondent population average
$\overline{y}_t$	True average for sample selected
$\overline{y}_m \equiv \overline{y}$	Measured average for sample selected
T	True value of a sample average
M	Measured value of a sample average

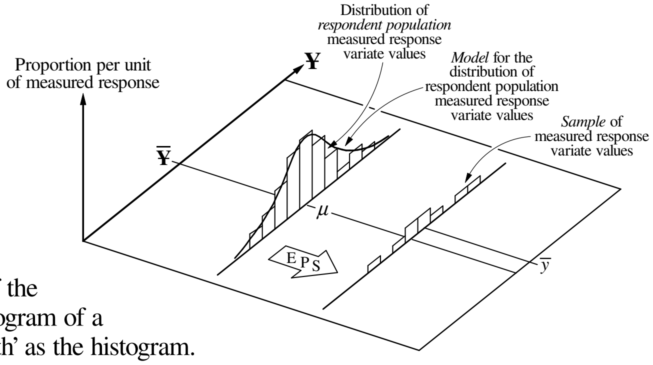

Overall error
Study error
Non-response error
Sample error
Sample attribute measurement error
Measuring bias
$\underline{\mathbf{Y}}$
$\overline{\underline{\mathbf{Y}}}_T$
$\overline{\underline{\mathbf{Y}}}_S$
$\overline{\underline{\mathbf{Y}}}$
$\overline{y}_t$
$\overline{y}$

6. Appendix 2 – Statistical Modelling and Model Error [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 $\overline{\underline{\mathbf{Y}}}$ in the response model (8.11.2) near the middle of page 8.49 in Figure 8.11 – see also the upper diagram at the right below and the discussion to its left.

The four main response models [which include (HL77.2)] discussed in STAT 231 are summarized in Statistical Highlight #71; model symbols are defined in Highlight #72, an overview of least square estimating of model parameters is given in Highlight #73.

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 at the bottom of page 2.28). 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.


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

All mathematical models are idealizations and are products of the intellect and the imagination; as shown in the schema at the right, *we* think of the model as a *link* between the *sample* and the *respondent population*. Also, Table 2.2.3 below gives terminology which helps maintain the real world-model distinction for measures of location and of variation and variability (see also Section 6 in Statistical Highlight #94).

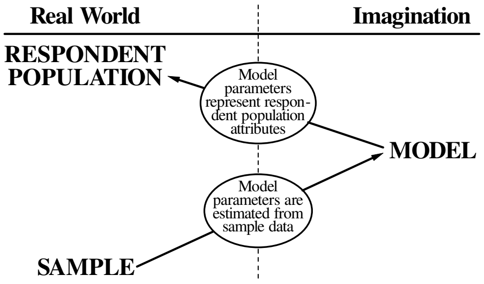

Real World
Imagination
RESPONDENT POPULATION
Model parameters represent respondent population attributes
MODEL
Model parameters are estimated from sample data
SAMPLE

Table 2.2.3

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