University of Waterloo W. H. Cherry

MEASURING: Measuring Processes

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

REAL WORLD $\xrightarrow[\textit{METHODS}]{\textit{INVESTIGATIVE}}$ DATA, INFORMATION and KNOWLEDGE

$\boxed{\text{Population or process}}$ $\xrightarrow[\textit{Measuring}]{\textit{Sampling}}$ Data and Answer(s) to Question(s)

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 Highlight #38, only the (ubiquitous) investigative methods of *sampling* and *measuring* are shown in the lower schema – 'measuring' is our present focus and 'sampling' is pursued in Statistical Highlight #21.

Measuring processes are used to obtain *variate values* (*i.e*., *data*); such processes 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 this Statistical Highlight #38, and these Course Materials more generally, 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 a degree that enables them to assess properly the limitation(s) imposed on Answer(s) by measurement (and, of course, other categories of) error.
  - – Assessing measurement error will usually be done in collaboration with other investigators who have relevant extra-statistical knowledge. (This may also be true of *other* categories of error – *e.g*., study error).

Discussing measuring processes in this Highlight #38 is useful also because it allows the (*un*familiar) key idea of *error*, and its consequences for statistical methods, to be presented in a context (namely, measuring) familiar to most readers.

- ○ An opportunity for *practical experience* with a measuring process is provided in the following Statistical Highlight #39; there is other discussion of measuring in Statistical Highlights #33 to #37 and #40 to #45, and of measurement *error* in #11 to #15.

2. Background II – Terminology and Related Matters [A more detailed Glossary is given in Statistical Highlight #91.] [optional reading]

- ∗ **Population:** a well-defined (finite) 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).

  In the STAT 231 Course Materials, the element-unit distinction is largely ignored and 'unit' is used throughout (for consistency with the STAT 231 Couse Notes). In the STAT 220 Materials, attention is generally restricted to units which are elements so the distinction does not arise but, more broadly, when units are *groups* of elements (as in *cluster* sampling, for instance), the distinction is essential; the matter is illustrated in more detail in Appendix 1 on pages HL77.8 and HL77.9 in Statistical Highlight #77.

  - ○ Materials like Statistical Highlight #77, concerned with survey *sampling*, refer in most places to units, but respondent population attributes of interest [like **₦** (size), **¥̄** (average), $_T$**¥** (total) and **Ꞩ** (standard deviation)] refer to *elements.*
- ∗ **Process:** ● a set of *operations* that produce or affect elements, OR:
  ● the *flow* of an entity (like water or electrons).

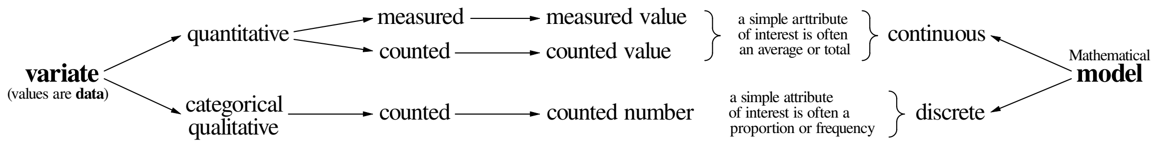
  Thus, in statistics, a process of the first type involves *elements.* WHEREAS:

  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 – see also the discussion under 'Independence' near the middle of page HL38.4.

  The term 'process', as used in the lower schema above at the upper right of Section 1, is to be set against 'population' as *we* use these two terms in a statistical context. In ordinary English, 'process' also encompasses *a series of actions that bring about an end or result*; this meaning (which covers the first case of 'process' defined above) is why we *also* refer to the investigative techniques of selecting, measuring and estimating as processes – sampling is then a process too because it includes selecting and estimating.

  The population-process distinction and is discussed in more detail in Statistical Highlight #94.
- ∗ **Sampling:** includes the processes of *selecting* and *estimating.*
- ∗ **Measuring:** the process used to determine the value of a variate.
- ∗ **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 some 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).

  Distinctions associated with *other* adjectives which qualify 'variate' – **quantitative**, **categorical** (or **qualitative**), **measured** and **counted**, **continuous** and **discrete** – are shown in the following schema.



- ○ An illustration of the distinction between a counted *value* [a real number, modelled by a **continuous** variate] and a counted *number* [an integer, modelled by a **discrete** variate] is:
  - **–** to assess the effectiveness of an insecticide, the number of insects could be counted on a defined part of each of a sample of plants from *un*treated and treated crops – a measure of effectiveness would be the decrease in the *average* number of insects per plant;
  - **–** to assess gender balance in an area of employment, the number of men and women employed in the area could be counted – a measure of the balance would be the *proportion* of each sex in the area.

  The quantitative *measured* and quantitative *counted* distinction blurs progressively as counts become larger in magnitude; it is also affected by the limited resolving power (finite precision) of real measuring processes.

  Continuous and discrete variates are *model* concepts because real measuring instruments with finite precision yield only *discrete* values.
- ○ *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.

∗ **Selecting:** the process by which the units of the sample are obtained from the respondent population – it is described in the **protocol for selecting units** (see Appendix 2 in Statistical Highlight #21). [Our *equiprobable* selecting is abbreviated EPS.]

∗ **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.

∗ **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).

∗ **Attribute**: a quantity defined as a function of response (and, perhaps, explanatory) variates over the five groups of elements defined below.

Familiar (simple) 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 the facing page HL38.3.

∗ **Repetition:** repeating over and over (usually *hypothetically*) one or more of the processes of selecting, measuring, estimating.

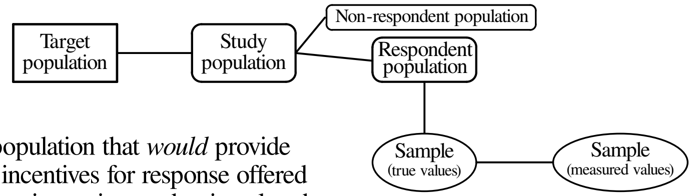
∗ **Error:** *Overall* error in an investigation refers to the net effect of *all* relevant categories of error on the Answer(s) from the investigation – see, for example, Appendix 1 on the lower half of page HL38.7 and also see Statistical Highlight #6.

∗ **Uncertainty:** ignorance (incomplete knowledge) of error; for example:
- **–** for a numerical Answer, ignorance of the magnitude and/or the sign/direction of error;
- **–** for a categorical Answer (like *Yes* or *No*), ignorance of whether the Answer is the correct category.

∗ **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.*

The schema at the right below shows five groups of elements which *we* distinguish for data-based investigating; a key feature for this Highlight #38 is the distinction between the true and measured values of variates of units in the sample.

∗ **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 at the bottom of page HL38.4.

∗ **Sample**: the group of units 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 above reminds us).

(*continued*)

University of Waterloo W. H. Cherry

MEASURING: Measuring Processes (continued 1)

- **Selection:** the units selected from the study population and which comprise the respondents (the sample) *and* the non-respondents – also see Note 1 at the bottom of page HL38.4. ['Selection' is terminology *specific* to these Statistical Highlights.]
- **Census:** using *all* the respondent population elements in an investigation; *i.e*., in a census, the 'sample' size is n = **N**.

The extensive discussion of sampling and (its consquence of) estimating (based on a probability model) in introductory statistics teaching can give the (wrong) impression that these processes are the *primary* source of limitations on Answers from data-based investigating. As indicated by the two lists at the right of the processes that lead to our six error categories, *four* categories – study error, non-response error, measurement error, comparison error (and possibly model error) – must be managed *regardless* of whether the Plan involves a sample or a census – see also Appendix 3 on pages HL21.4 and HL21.5 in Statistical Highlight #21.

A Plan involving a .....

sample:	**population:**
Specifying	Specifying
Selecting	Responding
Responding	Measuring
Measuring	Comparing
Estimating	
Comparing	

∗ **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.

We need *both* true values and long-term average values in *these* two error definitions because:
- ○ 'true' values for quantities like length, mass and time (and the many quantities derived from them) can be invoked because there are **standards** (*i.e*., certified *known* values) for such quantities – see also Section 3 on page HL38.5;
- ○ long-term average values may be all we have available when, for instance, investigating the distribution of responses to a questionnaire with particular question wording and/or question order.

∗ **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.

Model error is discussed further in Appendix 2 on page HL38.8, although it is peripheral to the main concern of this Highlight #38.

∗ **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.

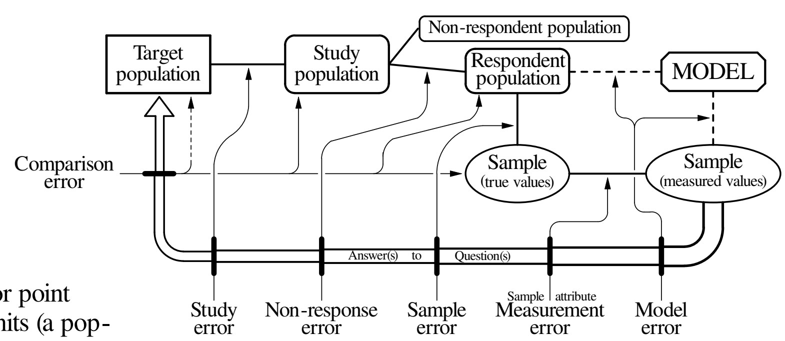
The alternate wording of the last phrase of the definition of comparison error accommodates the equivalent terminologies of lurking variates and confounding; in a particular context, we use the version of the definition appropriate to that context:
- ○ 'lurking variates' can more readily accommodate phenomena like Simpson's Paradox – see Statistical Highlight #51;
- ○ '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).

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

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

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) in Section 3 near the bottom of page HL38.5.

The schema at the lower right of the facing page HL38.2 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 HL60.1 at the upper right of page HL60.4 in Statistical Highlight #60.

∗ **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].

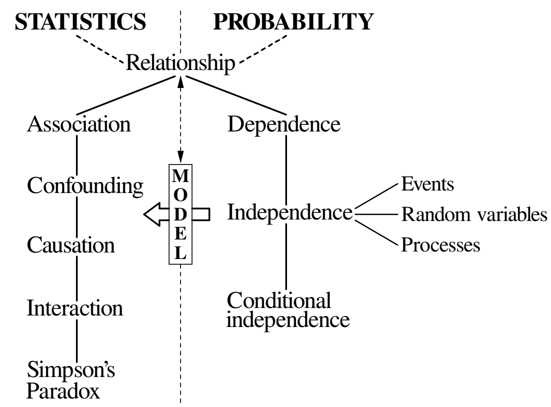
We distinguish variation from **variability**, which is defined lower on this page HL38.4 immediately above Note 1.

∗ **Independence, Independent:** a dictionary definition is: *not subject to the control, influence or determination of another or others.*
- **Independent measurements:** measurements are independent when the operator's knowledge of the value arising from one execution of the measuring process does *not influence* the value from any other execution.
  The Plan for an investigation needs to address the issue of measurement independence.
- **Independent events (Probabilistic independence):** events A and B are independent when the probabilities of *events A* and B are such that $\Pr(A|B) = \Pr(A)$ and $\Pr(B|A) = \Pr(B)$.
- **Independent random variables:** two random variables are independent when their *joint* probability (density) function is the *product* of their *marginal* probability (density) functions.
  This is the sense of 'independent' in response models like those summarized in Statistical Highlight #71.

As summarized on the left of the schema at the right, a relationship in *statistics* is often considered in terms of one or more of association, confounding, causation, interaction and Simpson's Paradox (see also the schema on page HL57.1 in Statistical Highlight #57).

In *probability* (on the right of the schema), a relationship is considered in terms of *dependence*, which comes in great variety and is often difficult to mathematize; as a consequence, introductory courses emphasize ***in**dependence*, as it applies to events, random variables and processes. Even the first two of these three involve an appreciable set of ideas and may be all a course has time to discuss.

Connection between statistical and probabilistic considerations of a relationship arises in the probability *models* statistics uses in the Analysis stage of the FDEAC cycle.



  ○ Emphasis on *in*dependence in introductory courses can obscure the fact that independence is a mathematical *idealization*. In the real world, ***de**pendence* is the norm – it may be that the behaviour of *every* particle in the universe depends on (*i.e.*, is affected by) *every other* particle, no matter how minute the degree of dependence.
    ○ This may be why lurking variates are usually so *numerous* when answering Questions with a causative aspect.

∗ **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.

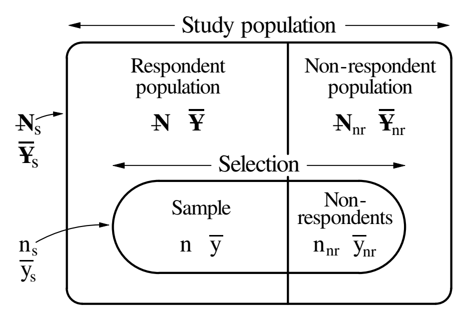
∗ **Imprecision:** *standard deviation* of error (*i.e.*, its *haphazard* component exhibited as *variation*) under *repetition*.
- **Sampling imprecision:** standard deviation of sample error under repetition of selecting and estimating.
- **Measuring imprecision:** standard deviation of measurement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

∗ **Accuracy:** the inverse of *in*accuracy.
∗ **Precision:** the inverse of *im*precision.
∗ **Bias:** the *model* quantity representing *inaccuracy.*
∗ **Variability:** the *model* quantity representing *imprecision*.

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 $\cancel{\mathbf{N}}$s and ns refer to *numbers of elements or units*;
- the $\overline{\cancel{\mathbf{Y}}}$s and $\bar{y}$s are *averages* of a response variate $\cancel{\mathbf{Y}}$ of the elements or units.

The relationships among the numbers of elements or units are:

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



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

(*continued*)

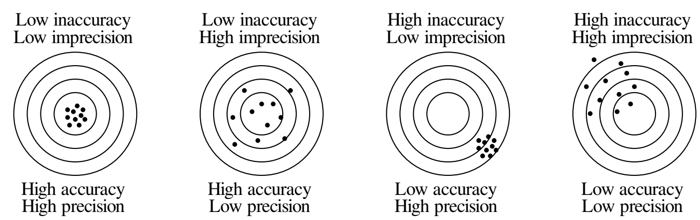
University of Waterloo W. H. Cherry

MEASURING: Measuring Processes (continued 2)

3. Measuring Processes [The title matter of this Highlight #38.]

When measuring, on an element or unit, a variate whose value does not change, we recognize that:

- making *one* measurement provides a value for the variate but no information about **measurement error** – the difference between a measured value and the true (or long-term average) value of the variate;
- making *more than one* measurement of the *same* variate on the *same* element or unit [the process of **repeated measuring** (or **repetition**)] and calculating their (data) standard deviation and average allows us to:
  - *see* that repeated measurements of the same quantity usually do *not* agree (exactly) with each other;
  - *quantify* **measuring imprecision** – the (data) *standard deviation* of the repeated measurements;
  - *quantify* **measuring inaccuracy** – the *average* of the repeated measurements minus the *true* value being measured.
    - + Measuring a *known* value (*i.e.*, a **standard**) to quantify measuring inaccuracy is called **calibrating** the measuring process.
    - + A classic discussion of measuring inaccuracy by W.J. Youden is summarized in Statistical Highlight #15.
  - The *average* of repeated measurements is likely to be *closer* to the *long-term* average than an *individual* measurement – that is, the average has lower *im*precision (or higher precision) than individual measurements. Alternatively, we can say the *average* of repeated measurements is likely to have measurement error of *smaller magnitude* than an *individual* measurement.
    - + These (equivalent) statements are the meaning of the (familiar) idea that the *average* of repeated measurements is a 'better' value than just *one* measurement (provided, of course, that the measurements are independent – see below).
  - We recognize that the sign and magnitude of error (which applies to a *particular* case) under *repetition* lead to the ideas of inaccuracy and imprecision, whose images are provided by patterns of shots on a target, as shown below at the right.



  - Implicit in the idea of *repetition* is that the measurements are **independent**, meaning the operator's knowledge of the value arising from one execution of the measuring process does *not influence* the value from any other execution.
    - + This (more informal) meaning of 'independence' should not be confused with **probabilistic independence**, defined on the upper half of the facing page HL38.4.
  - When elements or units are people and the measuring instrument is a questionnaire, repeated measuring is usually not feasible because respondents will likely recall their previous answers and so compromise the (measuring) independence that is required statistically.
    - + When we distinguish measuring 'physical' variates (involving one or more of length, mass and time) from measuring that requires a questionnaire, the key *statistical* difference is compromised ability for repeated measuring of the same element or unit. There appear to be few (or no) *statistical* issues with the seemingly different nature of the *sources* of measurement error in the two situations [*e.g.*, imperfections in the components (discussed below) of the measuring process for 'physical' variates, ignorant and/or careless and/or untruthful responses to a questionnaire].
      - ⊙ In the same vein, for a questionnaire to measure self-esteem, for instance, it is a *subject-area* (extra-statistical) matter to define what is *actually* being measured [*e.g.*, see pages HL42.2 and HL42.3 in Statistical Highlight #42].

  It is *un*clear how well the statistical issue of compromised repeated measuring is addressed by a questionnaire with *many* (perhaps hundreds of) questions, some of which are well-separated versions (or *in*versions) of the *same* question.

The foregoing discussion involves measuring variate values of *elements* (or *units*) but the effect of a measuring process on an *attribute* value is more important statistically – see equation (HL38.2) and the schema at the lower right of page HL38.7 in Appendix 1. For example, under a model for measuring inaccuracy where bias is *constant* (*i.e.*, *not* dependent on the value measured – for example, a ruler for which the first one centimetre is missing):

- inaccuracy will contaminate individual measured values and (often not recognized) is *un*affected by *averaging* – this is the situation with the estimate of the *intercept* of a least-squares regression line;
  - the average of measured values has lower *imprecision* than its individual measurements (*e.g.*, see Statistical Highlight #76);
- inaccuracy is *zero* for a *difference*, which is typically involved in comparisons, in estimates of standard deviations, and in the *slope* of a least-squares regression line [given the right in equation (HL38.1)].

$$\hat{\beta}_1 = \frac{\sum_{j=1}^{n} y_j(z_j - \overline{z})}{\sum_{j=1}^{n}(z_j - \overline{z})^2} \qquad \text{-----(HL38.1)}$$

We distinguish four *components* of a measuring process:

- the **measuring instrument** or **gauge**; ∗ the **operator(s)**; ∗ the **measuring protocol**; ∗ the **element** or **unit measured**.

Distinguishing these components makes it easier to identify *explanatory variates* which affect measured values and which may therefore be a source of measurement error; we consider in turn *statistical* matters associated with the four components.

Matters about the **measuring instrument** or **gauge** are:

(*continued overleaf*)

- Decreasing the *imprecision* of a measuring instrument usually involves increasing *cost*; for example, when measuring length:
  - a *ruler* costs about $5 and can be read to 0.1 mm;
  - a *pair of calipers* costs about $50 and can be read to 0.01 mm;
  - a *micrometer* costs about $500 and can be read to 0.001 mm (1 micron);

  in this instance, each decrease in imprecision by a factor of ten increases cost by about a factor of ten.
- Higher *cost* of a measuring instrument does *not* necessarily mean higher *accuracy* (lower *in*accuracy).
- In the context of a sample survey, the measuring *instrument* is the *questionnaire*; it is curious that investigators, who would *not* undertake assembly of the types of instrumentation used in a laboratory (*e.g.*, balances, spectrophotometers), often approach the task of developing the questionnaire with little recognition of the difficulty or importance of doing so successfully.
  - *Stability* of a measuring instrument – its ability to yield the *same* measured value in the same circumstances at points separated in time – is important but is not relevant to *all* measuring instruments; we usually distinguish:
    - *short*-term stability; – *long*-term stability.

    What constitutes a 'short' or 'long' time scale for stability is context dependent.

Matters about the **operator** are:

- In a clinical trial (used in medical research to assess, for example, the efficacy of a drug or surgical procedure), as indicated in Table HL38.1, blinding treatment *assessors* manages operator effect on *measuring inaccuracy* by trying to make assessment *independent* of the participant's treatment.

Table HL38.1:

Blinding of ...	**Short name**	**Statistical purpose**
Participants	Single blind	Manage *comparison error*
Treatment administrators	Double blind	Manage *comparison error*
Treatment assessors	Triple blind	Manage *measuring inaccuracy*

  - To be **blind** means not to know, for any element or unit, whether it is in the *treatment* group or the *control* group (which usually receives a dummy treatment known as a **placebo**).
  - The short names in the second column of the table for the blinding are *not* recommended because they do not distinguish adequately among the eight possible combinations of which group(s) are blind.

  Blinding of operator(s) as to the nature of the element or unit being measured may also be used in a medical diagnostic laboratory, where measuring *inaccuracy* is managed by analyzing **standards** at regular intervals concurrently with the primary task of analyzing biological materials.
- In a self-administered questionnaire (received in the mail or on line, for example), the *operator* is also the *respondent.*

The **measuring protocol** is the instructions for how to use the measuring instrument; one of its purposes is to promote *uniformity* in how different operators make measurements and so to try to make negligible, in the context of the investigation, any operator effect on the measured value obtained from the measuring instrument.

- Clear measuring protocol(s) and adherence to them by operators on different shifts are vital in a multi-shift manufacturing operation if a consistent product is to come from the different shifts (see also Notes 2 and 3 below and on page HL38.7).

Matters about the **element or unit measured** are concerned with the act of measuring *changing* the element or unit being measured or the value it yields. For example:

- Maclean's ranking of Canadian universities might make universities change their operations in ways that would improve their ranking but make no substantive change to the quality of the educational experience they offer students.
- Households selected for a panel used to obtain Nielsen ratings of TV programs might change their TV viewing habits as a consequence of *knowing* their viewing habits are being monitored.
- The interviewer administering a questionnaire (the 'operator') might (unintentionally) influence the person responding.
- A *slanted* question on a questionnaire may have a different effect on different (types of) respondents.

An extreme case is when measuring *destroys* the element or unit (*e.g.*, in quality assurance, firing shotgun cartridges or measuring bursting pressures of plastic bags or condoms); destructive measuring precludes the statistical benefits of repeated measuring on the same element or unit. (This is the same *statistical* issue as attempting repeated measuring when a questionnaire is involved).

NOTES: 2. When suppliers and assembly operations disagree about whether manufactured parts meet specifications, a common reason is measuring *inaccuracy* – the measuring processes used to check the parts at the supplier and assembly plants *dis*agree because they have not been calibrated to standards that *agree* with each other.

3. Assessing inaccuracy and imprecision of measuring process(es) to be used in an investigation, and the factors which affect them, is a common reason for one or more *sub*-FDEAC cycle(s) within the 'main' FDEAC cycle. An example is an industrial **gauge R&R** investigation.
   - * **Repeatability** of a gauge is the variation [expressed as an appropriate (data) standard deviation] of repeated measurements on each of a sample of (10, say) parts by *one* operator using the gauge;
   - * **Reproducibility** of a gauge is the between-operator variation [expressed as an appropriate (data) standard deviation] of two measurements, one by each operator using the gauge, on each of a sample of (10, say) parts.
     - The two operators are usually assumed to have *equal* repeatability.

   Repeatability quantifies the imprecision of a gauge under the most *favourable* conditions for operator effect.

University of Waterloo W. H. Cherry

MEASURING: Measuring Processes (continued 3)

NOTES: (cont.) 3. Reproducibility quantifies how this (lowest) imprecision is affected (increased) by having *two* operators.

- Investigators undertaking a measuring process assessment should take to heart the comments, made in 1966, by the U.S. National Bureau of Standards (now the National Institute of Standards and Technology), one of the world's premier measuring organizations:

> A major difficulty in the application of statistical methods to the analysis of measurement data is that of obtaining suitable collections of data. The problem is more often associated with conscious, or perhaps unconscious, attempts to make a particular process perform as one would like it to perform rather than accepting the actual performance ..... Rejection of data on the basis of arbitrary performance limits severely distorts the estimate of real process variation. Such procedures defeat the purpose of the ..... program. Realistic performance parame- ters require the acceptance of all data that cannot be rejected for cause.

SOURCE: Freedman, D., Pisani, R. and R. Purves: *Statistics.* First Edition, W. W. Norton & Company, New York, 1980, page 95.

4. A measuring process of statistical interest is so-called **randomized response**, whose simplest version involves an iterviewer asking a 'Yes/No' question about past 'sensitive' behaviour of the person being interviewed (the 'interviewee'), usually with the goal of estimating the population proportion of people who will admit they have engaged in the behaviour (*e.g.*, abortion, illicit drug use, viewing child pornography, money laundering, terrorism); randomized response was developed to manage two difficulties such investigations encounter:

∗ the interviewee may find the question too sensitive to give a truthful answer, AND/OR:

∗ the interviewer may be under a legal obligation to report the behaviour of a respondent who answers 'Yes'.

Randomized response manages both matters by having a box containing a number of (say, 100) cards, each with one of two questions in known proportions (say, 20% of cards have the first question, 80% have the second):

– *Is your birthday in July?* – *Have you ever engaged in ... ?*,

where the first question has a *known* distribution of *answers*, the second question names the behaviour of interest.

The interviewee selects a card 'at random *with* replacement' from the (well-mixed) box and answers it.

○ If the person *has* engaged in the behaviour *and* has selected a card containing the question about it, (s)he is not necessarily divulging sensitive information by answering 'Yes' and so may be more likely to answer truthfully [provided the interviewer *has* convinced the interviewee of their protection under randomized response];

○ the interviewer is legally protected because (s)he does not know to which question a 'Yes' answer applies.

An introductory version of the probabilistic basis of estimating the population proportion under randomized response, from a sample selected from the population, is the topic of Question A4-12 of the STAT 220 assignments.

- Question A4-16 (the 'three convicts' problem) raises an (*un*expected) probabilistic issue with the warder's *response* to convict A (see Statistical Highlight #49).

4. Appendix 1: – Overall Error for Questions with a Descriptive Aspect [The title matter of Statistical Highlight #18.]

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. -----(HL38.2)

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 (HL38.2); symbols are defined in Table HL38.2 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 (▪);

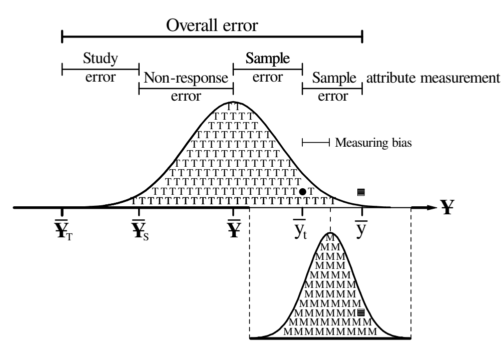
⊙ there is *no* sampling *bias* – the mean of the sampling distribution is $\bar{Y}$;

⊙ there *is* measuring *bias* – the horizontal distance between $\bar{y}_t$ 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 HL38.2: SYMBOL DEFINITIONS

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



(*continued overleaf*)

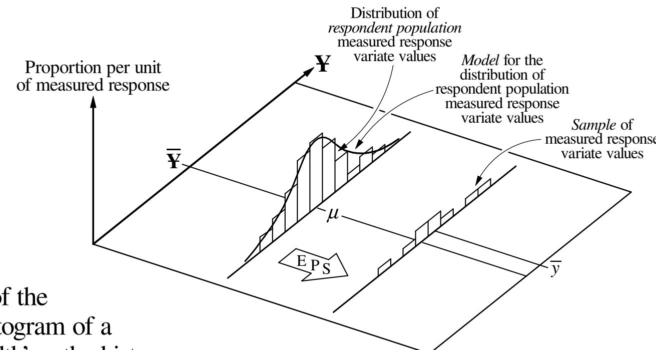
5. Appendix 2 – Statistical Modelling and Model Error [optional reading]

∗ **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.

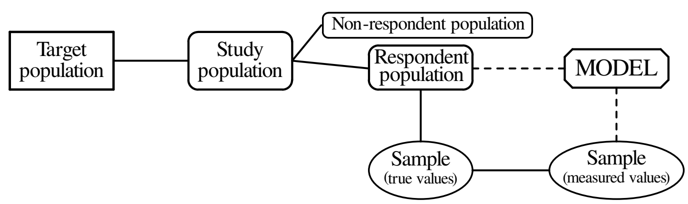
∗ **Model parameter:** a constant (which *we* denote by a *Greek* letter) in a response model that *represents* a respondent population *attribute*; for example, μ represents $\overline{\mathbf{Y}}$ in the response model (HL77.2) near the middle of page HL77.1 in Statistical Highlight #77 – see also the following diagram at the right and the discussion to its left.

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

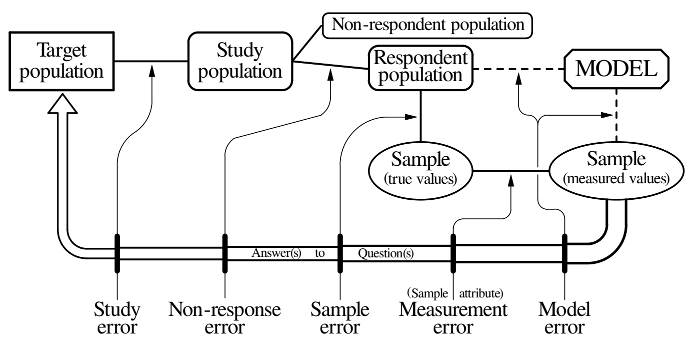
Model-based methods of analysis in statistics use data from a sample to *estimate* values of model parameters which then represent plausible values (in light of the data) for respondent population attributes and, hence, for Answer(s) to Question(s); we distinguish a *point* estimate from an *interval* estimate (defined near the middle of page HL38.2). 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.



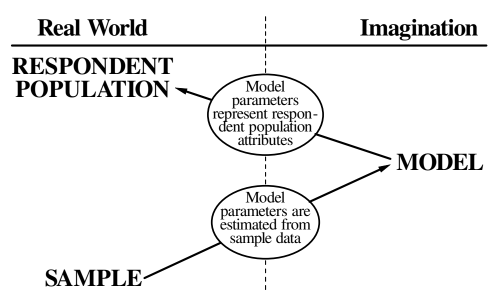
The schema at the lower right of page HL38.2 can be extended to include the model, as shown at the right; *our* view of the model as a link between the respondent population and the sample is elaborated in the schema at the bottom right of this page HL38.8 above Table HL38.3.



Model error, with its mathematical and probabilistic focus, is a broad and complex topic and differs in its nature from the other five categories of error involving attributes; the discussion of Plan components to manage model error (*e.g.*, in Statistical Highligh #18) is restricted to five assumptions (as given on page HL38.3) underlying our (four) response models (summarized in Statistical Highlight #71). The schema at the right above is shown again at the right, as a simplification of the one at the lower right of page HL38.3 by omitting comparison error.



All mathematical models are idealizations and are products of the intellect and the imagination. As indicated in the two schemas above, *we* think of the model as a *link* between the *sample* and the *respondent population*; a more detailed pictorial representation of this idea is shown at the right.



NOTE: 5. 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 HL38.3 at the right.

Table HL38.3

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