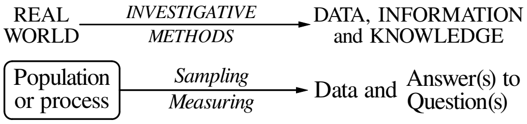
University of Waterloo W. H. Cherry

STATISTICAL TERMINOLOGY, NOTATION and DISTINCTIONS

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.

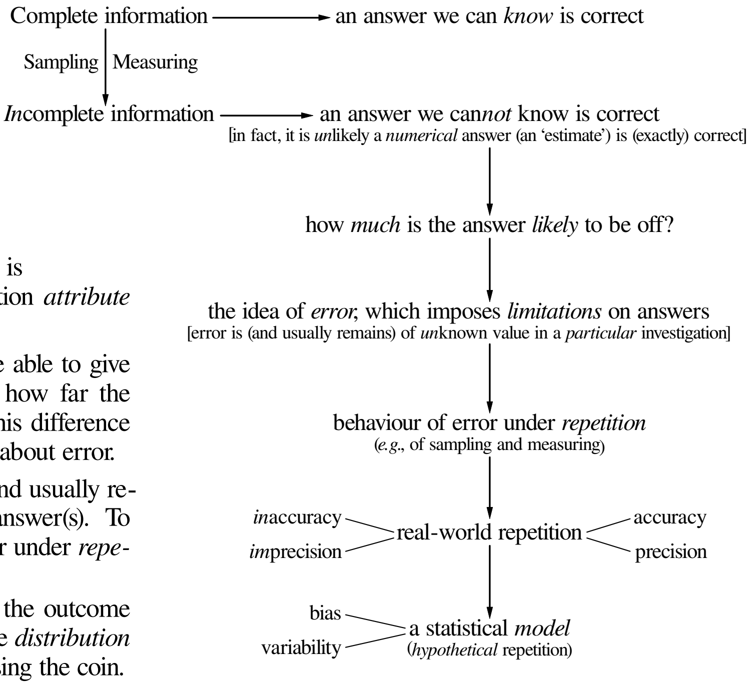
This introduction presupposes initially that data-based investigating is concerned with answer(s) to question(s) about a collection of *elements* which comprise a *population*. For example:

- ○ 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, Repetition, Accuracy and Precision [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.

(*continued overleaf*)

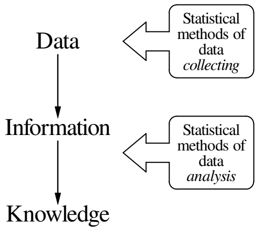
∗ A statistical *model*, including its *assumptions*, allows us to achieve our aim of quantifying (some of) the uncertainty in our estimate(s); part of identifying the limitations on an Answer is to assess model error arising from the difference between ('idealized') *modelling assumptions* and the real-world situation.

∗ We emphasize *sample error* and *measurement error* in this introductory overview; other categories of error, which need more background to understand, are (besides *model error*) *study error, non-response error* and *comparison error.*
- There can also be *non*-statistical error, although extra-statistical knowledge is generally required to assess it.

Error in our six categories is discussed in detail in the seventeen Statistical Highlights #6 to #22; Highlight #6 provides an introductory overview.

NOTES: 1. Another diagrammatic summary of what statistics is about is given at the right; it shows a broad division of the focus of statistical methods into data *collecting* and data *analysis.*

Instances of important questions to be answered (*i.e*., matters needing data-based investigating) are given by P. Calamai: *Why journalists can't add* [Chapter 3 (pages 15-24) in *Statistics, Science, and Public Policy*, A.M. Hertzberg and I. Krupta (editors), Queen's University, Kingston, Ontario K7L 3N6, 1998]:

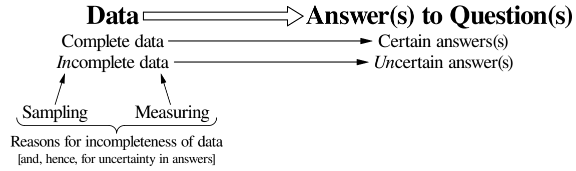

Data
Information
Knowledge
Statistical methods of data *collecting*
Statistical methods of data *analysis*

> What are the big things? They are issues such as silicone breast implants, the downsizing of the middle class, gender pay equity, the consumer price index, violence against women, the poverty line, drinking and pregnant women, race and crime, international debt comparisons, unemployment rates, the disappearance of the Atlantic cod, the tainted blood scandal and, of course, BSE and CJD, two sets of initials known only to a few researchers until Mad Cow Disease hit the headlines in 1996.

2. The following are illustrations of the precept that *certainty* requires *complete information* and, conversely, *in*complete information inescapably yields *un*certainty [ignorance (usually of the size or magnitude) of error]:
   - ● In the proof of a theorem in mathematics, we have a set of axioms and the rules of logic; within this system of complete information, there can be *certainty* the theorem is true.
   - ● Games like chess and bridge involve small populations of elements (32 pieces and 52 cards, respectively) and a set of rules; within these systems of complete information, there can be *certainty* which player has won a game.
   - ● In data-based investigating, information boundaries are seldom as clearly defined as in mathematics, chess and bridge, and practical considerations impose sources of *in*completeness and, hence, of *un*certainty, which include:
     - sampling: although we want Answer(s) to apply to a *population*, we seldom have the resources to investigate more than a *sub*set of the population (a *sample*) – the idea of sample error reminds us a sample seldom agrees *exactly* with the population;
     - measuring: even apparently straight-forward quantities like length and weight have *inherent* uncertainty in the values we obtain (the idea of measurement error), while measuring quantities like the number of cod in the North Atlantic fishing grounds or people's *attitudes* or *opinions* (*e.g*., on mandatory registration of firearms, capital punishment, abortion) may introduce much *greater* measuring uncertainty.

   A diagrammatic summary of these matters is given at the right above.

DATA-BASED INVESTIGATING



3. An illustration of *sample* error is polling to estimate a proportion of interest, typically from a national sample of about 1,000 to 1,500 adults. The *sample* proportion is a point estimate of the *population* proportion but is unlikely to be *exactly* equal to it; however, in a properly designed and executed poll, it is likely close enough (*i.e*., to have small enough sample error) to be an Answer with *acceptable* limitations, provided that *measurement* error (*e.g.*, arising from the questionnaire, the interviewer and the interview process) has also been adequately managed.

3. Why Do Terminology and Notation Matter in Statistics?

Literary discipline is usually needed for clear and succinct presentation of ideas in any respectable subject area, but introductory statistics teaching materials are notable for one or more of their obscurity, carelessness and muddled thinking. This unhappy state of affairs may be because:

- ● statistical ideas tend to be more than usually troublesome to explain,
- ● statistical writers tend to be more than usually undisciplined and careless.

Whatever the reasons for inadequate teaching materials, the consequences are that:

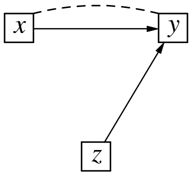
- it is difficult for students trying to learn from such materials to maintain their intellectual integrity,
- students' intellectual challenges arising from inadequacy of presentation compromise their capacity to engage with the (important and useful) *statistical* ideas being presented.

(*continued*)

STATISTICAL TERMINOLOGY, NOTATION and DISTINCTIONS (continued 1)

Three excerpts (from among many like them) illustrate the foregoing bleak assessment – should the reader laugh or cry?

> Two explanatory variables or lurking variables are **confounded** when their effects on a response variable are mixed together. When many variables interact with each other, confounding of several variables can prevent us from drawing conclusions about causation. In the diagram at the right which illustrates confounding, both the explanatory variable *x* and the lurking variable *z* may influence the response variable *y* – the dashed line shows an association, the arrows show a cause-and-effect link. Because *x* is confounded with *z*, we cannot distinguish the influence of *x* from the influence of *z*. We cannot say how strong the direct effect of *x* on *y* is. In fact, it is hard to say if *x* influences *y* at all.



> A Google search in May, 2022, yields a variety of results for **the law of large numbers**; for example:
>
> ○ *As a sample size grows, its mean gets closer to the average of the whole population.* (Investopedia)
>
> ○ *A theorem that describes the result of performing the same experiment a large number of times.* (Wikiedia)
>
> ○ *If you repeat an expriment a large number of times, what you obtain should be close to the expected value.* (Probability Course)
>
> ○ *As the number of identically distributed randomly generated variables increases, their sample mean (average) approaches their theoretical mean.* (Encyclopedia Britannica)
>
> ○ *A theorem that describes the result of repeating the same experiment a large number of times.* (Corporate Finance)
>
> ○ *If we take a sample of n observations of a random variable, the sample mean approaches the expected value of the random varable as n approaches infinity.* (Khan Academy)
>
> ○ *The frequencies of events with the same likelihood of occurrence even out, given enough trials or instances. As the number of experiments increases, the actual ratio of outcomes will converge on the expected, or theoretical, ratio of outcomes.* (Whatis.com)
>
> ○ *As a sample size becomes larger, the sample mean gets closer to the expected value.* (Statology)
>
> ○ *When we do an experiment a large number of times, the average result will be very close to the expected result.*
> *In other words, in the long run, random events tend to average out at the expected value.* (Math is Fun)
>
> ○ *The relative frequency of an event will converge on the probability of the event, as the number of trials increases.* (Statistics Dictionary)

> The statistical terminology of **degrees of freedom** is often first encountered in introductory statistics courses when the divisor, $\mathrm{n}-1$, in the calculation of s^2 is introduced. The sample $\overline{\mathrm{y}}$ is used as the location from which to measure deviations when computing s^2; so it has been arranged arbitrarily to make the sum of the deviations $(\mathrm{y}-\overline{\mathrm{y}})$ add to zero, which it would not ordinarily do if we used the location value μ. This constraint on the deviations is called loss of a degree of freedom. Its effect is most severe when the sample size is 1. Then y is the only observation, $\overline{\mathrm{y}} = \mathrm{y}$, and the deviation is $\mathrm{y}-\overline{\mathrm{y}} = 0$. However, when we square this and try to divide by $\mathrm{n}-1$, we find ourselves illegally trying to divide by zero. With only one observation, we thus have no way of using only sample data to compute s^2 as an estimate of σ^2. When we have only 1 observation and want to estimate σ^2, we lose the only degree of freedom we have by estimating μ by y, and we say we have no degrees of freedom left for estimating σ^2.
>
> When we want to estimate σ^2, we have to 'pay' a degree of freedom for having to estimate μ, and this leaves us with $\mathrm{n}-1$ degrees of freedom, or essentially $\mathrm{n}-1$ observations worth of information, for estimating σ^2. A 2×2 contingency table loses 3 degrees of freedom because of constraints imposed by marginal totals. The idea of degrees of freedom keeps coming up in more complicted statistical methods, and so we need to become comfortable with it. Such statistical methods may lose additional degrees of freedom by estimating more things.

Introductory statistics teaching materials are a vast body but, rather than quantity, we need materials that are correct, clear and useful. Achieving any *one* of these three characteristics is a demanding task; together, the three are a daunting undertaking. If the current dismal situation is to be rectified, teaching materials need to be built around three distinctions between:

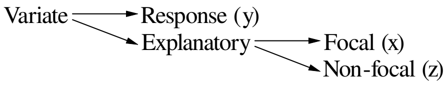
∗ the population and the sample, ∗ the real world and the model, ∗ the individual case and behaviour under repetition.

Such teaching materials can be achieved by appropriate terminology and notation presented and maintained with unremitting discipline; it is also useful to discard some (*un*helpful) current terminology. Rigorously maintaining these distinctions fosters a mind-set which routinely recognizes what statistical methods can, and can*not*, accomplish. [These three distinctions are a seminal insight of the writer's colleague, Prof. R. Jock MacKay.]

4. Terminology, Notation and Distinctions

From the start, we avoid unthinking adoption from mathematics of matters *un*suited to statistics.

∗ Rather than (dependent and independent) 'variables', we have 'variates' which (as summarized in the schema at the right) are designated as response (letter y) or explanatory, with the latter being focal (letter x) or non-focal (letter z).



∗ Instead of only using *italics* as is common for typesetting mathematics, we exploit more broadly letter case and face/style:

- upper case **bold** letters are for **population** quantities,
- lower case Roman letters are for (sample) data,
- upper case *italic* letters are for ***random variables***,
- lower case *italic* letters are for *values* of random variables;
- also, lower case Greek letters [*e.g.*, μ (mu), σ (sigma), π (pi)] are used for (probability and response) *model parameters.*

(*continued overleaf*)

As indicated for a response variate in the first line of Table 1 at the right, these notation conventions enable maintaining the population-sample and real world-model distinctions.

Table 1

Real world	**Model**
¥, y	Y, y
$\overline{\textbf{¥}}$, $\overline{\mathrm{y}}$	$\overline{Y}$, $\overline{y}$, μ
₱, p	P, p, π

- ● The line through the **bold** letter (we *say* ‘y cross’) is to distinguish **¥** from *Y* in handwritten symbols.
- ● From the beginning, *we* use letter y as the *generic* variate, which is a *response* for much of the discussion in introductory statistics; this helps avoid confusion from (the surprisingly common practice in introductory texts of) starting with x as the generic variate (a carry-over from mathematics?) and then having to switch to y when (say) linear regression is discussed.

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

The second line of Table 1 is for an *attribute* [of a *group* of elements (or units) like a population (or a sample)], specifically an *average* in the context of our simplest response model of equation (1). We say (or write): *the* ***population*** *average* $\overline{\textbf{¥}}$ *is represented by the* ***model*** *mean* μ, *for which the* ***estimator*** *is the random* variable $\overline{Y}$ *and the* ***estimate*** *is the sample average* $\overline{\mathrm{y}}$ (= $\overline{y}$).

- ● Implicit (and easily overlooked) in this statement is the *investigator’s* responsibility to ensure that the Plan for the investigation makes it reasonable to treat $\overline{\mathrm{y}}$ (the sample average – a *number* – calculated from data) as $\overline{y}$ (the value of the random variable representing the sample average in the *model*).
- ● The *length* of the statement makes it tempting to try to shorten it but, as words are omitted, the distinctions it makes become obscured or are lost altogether (*e.g.*, ‘μ is estimated by $\overline{\mathrm{y}}$’) – see also Table 2 below.
- ● Symbols are often subscripted; *e.g.*, Y_j and R_j in model (1) and **Ƶ**$_1$, **Ƶ**$_2$, ..., **Ƶ**$_\mathrm{k}$ for multiple non-focal explanatory variates.

The third line of Table 1 shows another (useful) attribute, a *proportion* (letter p); reserving letter ‘p’ for proportions entails *our* using ‘Pr( )’ (Roman letters) for *probability*, in contrast to the (random variable look-alike) ‘*P*’ of most introductory statistics texts.

A feature of these Materials is their emphasis on *standard deviation* rather than *variance* – the latter, which (unhelpfully) is the focus of traditional statistics teaching, is mentioned only with reluctance for reasons given on page HL91.22 in Statistical Highlight #91 [see also the comment (●) above Note 9 on page 2.138 in Figure 2.17 in the STAT 332 Course Materials (1995 curriculum)].

– Standard deviations entail the routine typesetting of square roots, an *onerous* task for achieving aesthetic outcomes.

For standard deviations (*and* averages), maintaining the real-world/model distinction is aided by terminology given in Table 2 at the right below; we exploit the availability of two words in English for measures of location but, with *one* term for *variation* and *variability*, these Materials suggest, particularly for beginning students, (mentally) adding the adjectives ‘data’ and ‘probabilistic’.

- + The *average-mean* distinction is absent from current statistical discourse – there also seems to be almost wilful (unnecessary) use of ‘mean’, perhaps to sound erudite, because we learn about averages in grade school where it would be inappropriate to raise the (nuanced) matter of a divisor *other than* the number of observations (widely denoted n, Roman ‘n’ in these Materials).

Table 2

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

- + The (unprofitable) *n−1 vs. n saga* is about the ‘correct’ divisor for calculating (under constraints) the average deviation from the average for a (data) standard deviation – see Appendix 3 on pages HL100.7 to HL100.9 in Statistical Highlight #100.
- + In contrast to (nuanced) averaging of *data*, at least in introductory statistics there is *no* disagreement about how to calculate the mean and standard deviation of a random variable – see, for example, Figures 5.9 and 7.11 in the STAT 220 Course Materials.

The foregoing matters can be seen *in use* in Appendix 1 on pages TND-7 and TND-8 and, in a regression context, in Figure 16.1 in the STAT 231 Course Materials; the Glossary entries in Statistical Highlight #91 also provide more detail.

We distinguish *four* populations by evocative adjectives:

✻ the **target** population; ✻ the **study** population; ✻ the **respondent** population; ✻ the **non-respondent** population.

These populations are central to developing the first two categories of *error* in Table 3 at the right below; we see them, for instance, as part of the right-hand schema on the lower half of page TND-6 and of the diagram at the upper right of page HL91.11 in Statistical Highlight #91. [Some contexts involve a *process* rather than a population – for example, see Statistical Highlight #94].

These Materials maintain throughout that (real-world) populations in statistics are *finite*; infinite ‘populations’ are a (sometimes useful) *model* – see, for example, Appendix 4 on page HL77.10 in Statistical Highlight #77.

The individual case-repetition distinction is already *implicit* in our use of:

- ● *error*, which arises in an *individual* investigation, AND:
- ● a response *model* like the one in equation (1) above, because such models only describe behaviour under *repetition*.

The distinction is *explicit* in the terminology for our six categories of error; as shown in Table 3 at the right, (numerical) error under repetition can become either inaccuracy or imprecision, depending on whether its sign or magnitude is involved. *Repetition* is evoked by adding ‘-ing’ to the eleven adjectives for inaccuracy and imprecision.

Table 3

Individual case	.... **Repetition**	
Error	**Inaccuracy**	**Imprecision**
Study	Studying	(Studying)
Non-response	Non-responding	(Non-responding)
Sample	Sampling	Sampling
Measurement	Measuring	Measuring
Comparison	Comparing	Comparing
Model	Modelling	-----

- ● Studying and non-responding imprecision are de-emphasized in the right-hand column of Table 3 because these Materials assume a *deterministic* (*not* a *stochastic*) process for specifying the study population and, for people as elements or units,

(*continued*)

STATISTICAL TERMINOLOGY, NOTATION and DISTINCTIONS (continued 2)

deciding whether to respond (under given incentives). [Elsewhere, there *are* stochastic models for the response process.]

- Measurement error arises for an individual measurement but *overall* error involves ***attribute*** *measurement error.*
  - Usually, attribute measurement error error is managed in an investigation by managing measurement error.
  - The *effect* of measurement error may differ between variates and an attribute (like the slope of a regression line).
  - Our usual concern is with *sample* attribute measurement error because a census is rare in practice.
- Model error (in the last line of Table 3) is the only category of error *not* defined in terms of attributes; its nature means it is seldom appropriate to consider modelling imprecision, as indicated by the final dash (-----) in Table 3.

Our six error 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 manage the six categories of error are summarized in Table HL6.1 on pages HL6.4 and HL6.5 in Statistical Highlight #6.

NOTE: 4. It is unfortunate that, for two essential concepts, statistics deals *directly* with ***in***accuracy and ***im***precision, whereas common useage involves the (more familiar and, seemingly, more straight forward) inverses, namely accuracy and precision.

5. The FDEAC Cycle

A characteristic of statistics courses is their concern (some might say obsession) with *data analysis.* For later specialized courses, this concern is usually appropriate but, in introductory courses where students typically encounter the subject at a serious level for the first time, it is *imperative* they learn that there is much *more* to statistics than data analysis. They need to recognize the hard truth that unless many *difficult* tasks of error management are undertaken successfully *before* data analysis, its answer(s) are likely to be fatally compromised by severity of limitations. The tabular summary of the components of the FDEAC cycle in Table 4 below (an overview of a *process* for data-based investigating) provides a basis for students to start to assimilate these vital ideas; Table 4 is elaborated in Statistical Highlights #88 and #89.

Another lesson to learn is that *in*adequate management of any *one* category of error can result in severe limitation on answer(s), despite all *other* error categories being adequately managed.

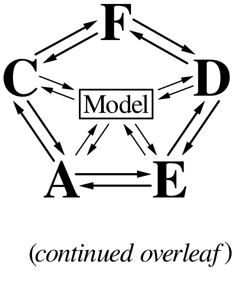
Table 4: The FDEAC cycle: a *structured process* for data-based investigating

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***

The kudos from being party to arcane mathematical and computational techniques, and the ‘excitement’ of reaching more quickly the goal of answers, combine to make it widespread in statistics teaching and practice to devote *in*adequate resources to error management in the relative drudgery of Question formulation, developing a proper Plan and its careful execution; the consequence of answers that likely embody *falsehoods* can easily be (and routinely is) ignored.

The diagram at the right makes the following points about the FDEAC cycle in relation to the *model*:

- ○ 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



(*continued overleaf*)

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 there are clear Question(s), part of the Plan to answer these Question(s) is an appropriate model (*e.g.*, for sampling and measuring processes, including calculating sample size) [discussed in Figure 2.13 of STAT 332];

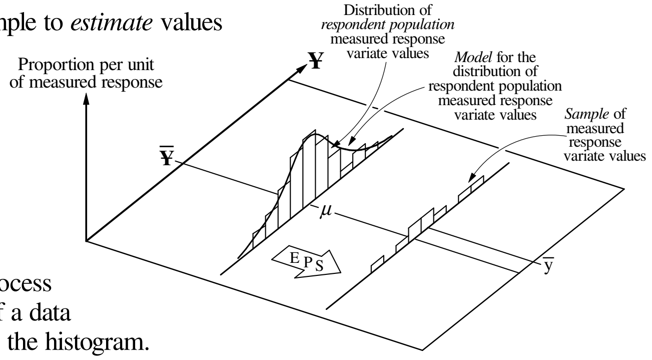
– the Execution stage generates data used to estimate model parameters representing study population attributes, and error in these estimates depends on how the data are generated (*e.g.*, imprecision and inaccuracy of measuring processes);

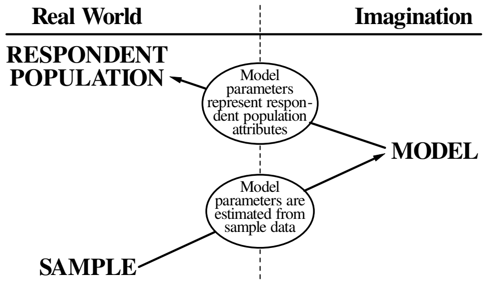
– the Analysis stage may use model-based formal methods of data analysis and then the model must be appropriate for the method(s) of analysis that will obtain from the data the information needed to answer the Question(s);

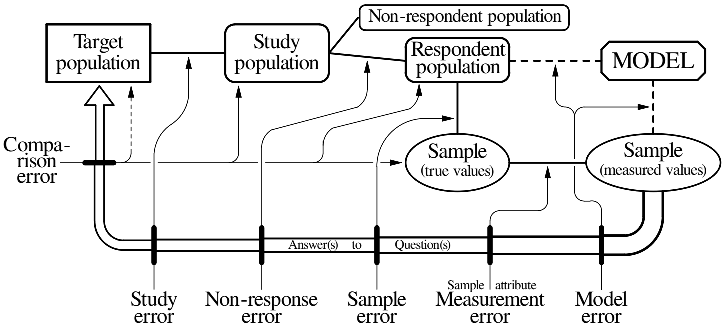
– the Conclusion stage must consider the model and its assumptions in assessing limitations on Answer(s) and the model must be chosen with Answer(s) and their acceptable limitations in the Question context in mind.

The three diagrams below are pictorial reminders of, respectively, the respondent population-model-sample sequence, our view of the model as a link in this sequence, and of how our four populations, the sample and the six categories of error comprise a process for data-based investigating.

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 Questions; we distinguish a *point* estimate from an *interval* estimate (defined on page HL91.11 in Statistical Highlight #91). When the normal model is appropriate for the distribution of the response variate values, the model mean μ is estimated by the sample average $\overline{\mathrm{y}}$ (= $\overline{y}$) and σ is estimated by the sample standard deviation s (= 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.







6. Terminology to Avoid or Use With Care

As well as defining appropriate and evocative terminology, clarity is aided by *not* duplicating existing terms with (equivalent) 'feel-good' words (with possibly ambiguous statistical connotations) – in statistics, we eschew 'elegant variation' in wording.

∗ **Applicability** (of an Answer) refers to study error and/or sample error.

∗ **Generality** (of an Answer) usually refers to sample error; in DOE, **generality** (or a **wider inductive basis**) may refer to whether the Plan involves a factorial treatment structure so that interaction effect(s) can be estimated.

– **Generalizability** refers to *study* error and **generalization** to *sample* error in the social sciences.

∗ **Reliability** [usually] refers to adequate precision (attained by managing *im*precision) [sometimes to adequate accuracy].

∗ **Sensitivity** (ability to detect an effect) refers to adequate precision (attained by managing *im*precision).

∗ **Strength** (of an Answer) means precision so **weakness** means *im*precision.

∗ **Trustworthiness** (of an Answer) means accuracy so **untrustworthiness** means *in*accuracy.

∗ **Validity** (of an Answer) means accuracy so **invalidity** means *in*accuracy.

As well as the clarity that results from using the ideas only of 'accuracy,' 'precision' and 'error,' Statistical Highlight #91, in the Glossary starting on page HL91.7, suggests restricted use or avoidance as statistical terminology for the following words or phrases:

degrees of freedom, error (in regression models), experiment, hypothesis testing/test of hypothesis, independent, pivotal/pivotal quantity, population parameter, probable error, random, randomization, relative frequency, representative sample, sample statistic, sampling bias, significance level, simple random sampling, statistic, survey, test of significance, test statistic, universe, variance.

(*continued*)

STATISTICAL TERMINOLOGY, NOTATION and DISTINCTIONS (continued 3)

7. Appendix 1: The Notation in Use – Illustrations from the STAT 332 Course Materials [optional reading]

The following tables (with STAT 332 page references) illustrate the notation advocated and used in these Materials.

Figure 2.3, page 2.37

Table 2.3.1:

SYMBOL			DESCRIPTION
Random variable	Value	Respondent population	
Y	y	$\cancel{\mathbf{Y}}$	Response variate
–	j	i	Summation index
–	x	$\cancel{\mathbf{X}}$	Focal explanatory variate
–	z	$\cancel{\mathbf{Z}}$	Explanatory variate
–	n	$\cancel{\mathbf{N}}$	Number of units/elements
$\overline{Y}$	$\overline{y}$	$\overline{\cancel{\mathbf{Y}}}$	Average (sum ÷ number)
R	r	$\cancel{\mathbf{R}}$	Residual [or Ratio]
S	s	$\cancel{\mathbf{S}}$	Standard deviation

Figure 2.3, page 2.38

Table 2.3.2: SUMMARY OF STANDARD DEVIATIONS

Response Models	Survey Sampling
σ Model parameter	$\cancel{\mathbf{S}}$ Respondent population standard deviation – an attribute
$\hat{\sigma}$ Estimate of σ	s Sample standard deviation – we take s = s to estimate $\cancel{\mathbf{S}}$
$\tilde{\sigma}$ Estimator of σ	S Estimator corresponding to s – a random variable

Figure 2.3, page 2.38

Table 2.3.3

Respondent population standard deviation	$\cancel{\mathbf{S}}$	*data* standard deviation
Sample standard deviation	s	
Standard deviation of the sample average	$s.d.(\overline{Y})$	*probabilistic* standard deviation
Estimated standard deviation of the sample average	$\widehat{s.d.}(\overline{Y})$	

Figure 2.3, page 2.40

Table 2.3.5: ....QUANTITY......	RESPONDENT POPULATION	.......SAMPLE [MODEL]....................
Size (elements/units)	$\cancel{\mathbf{N}}$	n
Response	$\cancel{\mathbf{Y}}_i$ (i = 1, 2, ..., $\cancel{\mathbf{N}}$)	y_j (j = 1, 2, ..., n) [r.v.s are Y_j with value y_j]
Average	$\overline{\cancel{\mathbf{Y}}} = \frac{1}{\cancel{\mathbf{N}}}\sum_{i=1}^{\cancel{\mathbf{N}}}\cancel{\mathbf{Y}}_i = \frac{1}{\cancel{\mathbf{N}}}\,{}_T\cancel{\mathbf{Y}}$	$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ [r.v. is $\overline{Y}$ with value $\overline{y}$]
Total	${}_T\cancel{\mathbf{Y}} = \cancel{\mathbf{N}}\overline{\cancel{\mathbf{Y}}} = \sum_{i=1}^{\cancel{\mathbf{N}}}\cancel{\mathbf{Y}}_i$	${}_Ty = \cancel{\mathbf{N}}\overline{y}$ [r.v. is ${}_TY$ with value ${}_Ty$]
Standard deviation	$\cancel{\mathbf{S}} = \sqrt{\frac{1}{\cancel{\mathbf{N}}-1}SS_{\cancel{\mathbf{Y}}}} \equiv \sqrt{\frac{1}{\cancel{\mathbf{N}}-1}\sum_{i=1}^{\cancel{\mathbf{N}}}(\cancel{\mathbf{Y}}_i-\overline{\cancel{\mathbf{Y}}})^2}$	$s = \sqrt{\frac{1}{n-1}SS_y} \equiv \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j-\overline{y})^2}$ [r.v. is S with value s]

Figure 2.3, page 2.45

Table 2.3.7:	Elements	Clusters	Relationships
Respondent population	$\cancel{\mathbf{N}}$	$\cancel{\mathbf{M}}$	$\cancel{\mathbf{N}} = \cancel{\mathbf{M}}L$, $L = \cancel{\mathbf{N}}/\cancel{\mathbf{M}}$
Sample	n	m	n = mL, L = n/m

Also: $\overline{\cancel{\mathbf{Y}}}_i = \frac{1}{L}\sum_{k=1}^{L}\cancel{\mathbf{Y}}_{ik}$ is the average response of the ith population cluster,
$\overline{y}_j = \frac{1}{L}\sum_{k=1}^{L} y_{jk}$ is the average response of the jth sampled cluster,
the subscript *ec* in Table 2.3.8 at the right denotes 'equal-sized clusters'.

Figure 2.3, page 2.45

Table 2.3.8

EPS of elements	Page	EPS of clusters	Page
$\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$	2.40	$\overline{y}_{ec} = \frac{1}{m}\sum_{j=1}^{m}\overline{y}_j$	2.105
$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j-\overline{y})^2}$	2.40	$s_{ec} = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(\overline{y}_j-\overline{y}_{ec})^2}$	2.106
$\cancel{\mathbf{S}} = \sqrt{\frac{1}{\cancel{\mathbf{N}}-1}\sum_{i=1}^{\cancel{\mathbf{N}}}(\cancel{\mathbf{Y}}_i-\overline{\cancel{\mathbf{Y}}})^2}$	2.40	$\cancel{\mathbf{S}}_{ec} = \sqrt{\frac{1}{\cancel{\mathbf{M}}-1}\sum_{i=1}^{\cancel{\mathbf{M}}}(\overline{\cancel{\mathbf{Y}}}_i-\overline{\overline{\cancel{\mathbf{Y}}}})^2}$	2.105
$s.d.(\overline{Y}) = \cancel{\mathbf{S}}\sqrt{\frac{1}{n}-\frac{1}{\cancel{\mathbf{N}}}}$	2.41	$s.d.(\overline{Y}_{ec}) = \cancel{\mathbf{S}}_{ec}\sqrt{\frac{1}{m}-\frac{1}{\cancel{\mathbf{M}}}}$	2.105
$\overline{y} \pm {}_{\alpha}t^*_{n-1}s\sqrt{\frac{1}{n}-\frac{1}{\cancel{\mathbf{N}}}}$	2.42	$\overline{y}_{ec} \pm {}_{\alpha}t^*_{m-1}s_{ec}\sqrt{\frac{1}{m}-\frac{1}{\cancel{\mathbf{M}}}}$	2.106

Figure 2.10, page 2.81

Table 2.10.1: ....QUANTITY.......	RESPONDENT POPULATION	...............SAMPLE [MODEL]................
Size (elements/units)	$\cancel{\mathbf{N}}$	n
Indicator variate	$\cancel{\mathbf{V}}_i$ (i = 1, 2, ..., $\cancel{\mathbf{N}}$)	v_j (j = 1, 2, ..., n) [r.v.s are V_j with value v_j]
Proportion	$\cancel{\mathbf{P}} = \frac{1}{\cancel{\mathbf{N}}}\sum_{i=1}^{\cancel{\mathbf{N}}}\cancel{\mathbf{V}}_i = \overline{\cancel{\mathbf{V}}}$	$p = \frac{1}{n}\sum_{j=1}^{n} v_j = \overline{v}$ [r.v. is P with value p (= $\overline{v}$)]
Number (or frequency)	$\cancel{\mathbf{N}}\cancel{\mathbf{P}}$	(number in sample = np)

Figure 2.14, page 2.112

Table 2.14.7:		...... Figure		
		2.3	2.10	2.14
Population	Response	$\cancel{\mathbf{Y}}_i$	C_i or ${}^cC_i$	$\overline{\cancel{\mathbf{Y}}}_i$
	Attribute(s)	$\overline{\cancel{\mathbf{Y}}}$, $\cancel{\mathbf{S}}$	$\cancel{\mathbf{P}}$	$\overline{\overline{\cancel{\mathbf{Y}}}}$, $\cancel{\mathbf{S}}_{ec}$
	Size	$\cancel{\mathbf{N}}$	$\cancel{\mathbf{N}}$	$\cancel{\mathbf{M}}$
Model	Random variable(s)	$\overline{Y}$, S	P	$\overline{Y}_{ec}$, (S_{ec})
	Value(s)	$\overline{y}$, s	p	$\overline{y}_{ec}$, s_{ec}
Sample	Response	y_j	C_j or ${}^cC_j$	$\overline{y}_j$
	Attribute(s)	$\overline{y}$, s	p	$\overline{y}_{ec}$, s_{ec}
	Size	n	n	m

Figure 2.10, page 2.88

Table 2.10.9

QUANTITY	SYMBOL
Population proportion	$\cancel{\mathbf{P}}$
Sample proportion	p
Random variable	P
A value of P	p
Model parameter	π
Probability	Pr

(*continued overleaf*)

Figure 2.15, page 2.113 and **Figure 2.17, page 2.131**

Table 2.15.1: **Quantity**	**Respondent** **Population**		**Sample** (Estimate) (= model *value*)		**Model** (*Estimator*)	
Average	$\bar{\mathbf{Ɏ}}$	$\bar{\mathbf{X̵}}$	$\bar{\mathrm{y}}\ (= \bar{y})$	$\bar{\mathrm{x}}\ (= \bar{x})$	$\bar{Y}$	$\bar{X}$
Total	${}_{\mathrm{T}}\mathbf{Ɏ}$	${}_{\mathrm{T}}\mathbf{X̵}$	${}_{\mathrm{T}}\mathrm{y} = \mathbf{N̵}\bar{\mathrm{y}}\ (= {}_{\mathrm{T}}y)$	${}_{\mathrm{T}}\mathrm{x} = \mathbf{N̵}\bar{\mathrm{x}}\ (= {}_{\mathrm{T}}x)$	${}_{\mathrm{T}}Y$	${}_{\mathrm{T}}X$
Ratio	$\mathbf{R̵} = \bar{\mathbf{Ɏ}}/\bar{\mathbf{X̵}} \equiv {}_{\mathrm{T}}\mathbf{Ɏ}/{}_{\mathrm{T}}\mathbf{X̵}$		$\mathrm{r} = \bar{\mathrm{y}}/\bar{\mathrm{x}} \equiv {}_{\mathrm{T}}\mathrm{y}/{}_{\mathrm{T}}\mathrm{x}\ (= r)$		R	

Figure 2.16, page 2.125

Table 2.16.2: ......QUANTITY......	**RESPONDENT POPULATION**	................**SAMPLE [MODEL]**................
Size: units ≡ clusters	$\mathbf{M̵}$	m
elements	$\mathbf{N̵} = \sum_{\mathrm{i}=1}^{\mathbf{M̵}} \mathbf{Ł}_{\mathrm{i}}$	$\mathrm{n} = \sum_{j=1}^{\mathrm{n}} \mathrm{l}_j$
Cluster size	$\mathbf{Ł}_{\mathrm{i}}\ (\mathrm{i} = 1, 2, \ldots.., \mathbf{M̵})$	$\mathrm{l}_j\ (j = 1, 2, \ldots.., \mathrm{m})$ [r.v. is L_j with value l_j]
Average cluster size	$\bar{\mathbf{Ł}} = \mathbf{N̵}/\mathbf{M̵} = \frac{1}{\mathbf{M̵}}\sum_{\mathrm{i}=1}^{\mathbf{M̵}} \mathbf{Ł}_{\mathrm{i}}$	$\bar{\mathrm{l}} = \mathrm{n}/\mathrm{m} = \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}} \mathrm{l}_j$ [r.v. is $\bar{L}$ with value $\bar{l}$]
Response	$\mathbf{Ɏ}_{\mathrm{ik}}\ (\mathrm{i} = 1, 2, \ldots.., \mathbf{M̵};\ \mathrm{k} = 1, 2, \ldots.., \mathbf{Ł}_{\mathrm{i}})$	$\mathrm{y}_{jk}\ (j = 1, 2, \ldots.., \mathrm{m};\ k = 1, 2, \ldots.., l_j)$ [r.v. is Y_{jk} with value y_{jk}]
Cluster average	$\bar{\mathbf{Ɏ}}_{\mathrm{i}} = \frac{1}{\mathbf{Ł}_{\mathrm{i}}}\sum_{\mathrm{k}=1}^{\mathbf{Ł}_{\mathrm{i}}} \mathbf{Ɏ}_{\mathrm{ik}} = \frac{1}{\mathbf{Ł}_{\mathrm{i}}}\,{}_{\mathrm{T}}\mathbf{Ɏ}_{\mathrm{i}}$	$\bar{\mathrm{y}}_j = \frac{1}{\mathrm{l}_j}\sum_{k=1}^{l_j} \mathrm{y}_{jk} = \frac{1}{\mathrm{l}_j}\,{}_{\mathrm{T}}\mathrm{y}_j$ [r.v. is $\bar{Y}_j$ with value $\bar{y}_j$]
Cluster total	${}_{\mathrm{T}}\mathbf{Ɏ}_{\mathrm{i}} = \sum_{\mathrm{k}=1}^{\mathbf{Ł}_{\mathrm{i}}} \mathbf{Ɏ}_{\mathrm{ik}} = \mathbf{Ł}_{\mathrm{i}}\bar{\mathbf{Ɏ}}_{\mathrm{i}}$	${}_{\mathrm{T}}\mathrm{y}_j = \sum_{k=1}^{l_j} \mathrm{y}_{jk} = \mathrm{l}_j\bar{\mathrm{y}}_j$ [r.v. is ${}_{\mathrm{T}}Y_j$ with value ${}_{\mathrm{T}}y_j$]
Average cluster total	${}_{\mathrm{T}}\bar{\mathbf{Ɏ}} = \frac{1}{\mathbf{M̵}}\sum_{\mathrm{i}=1}^{\mathbf{M̵}} {}_{\mathrm{T}}\mathbf{Ɏ}_{\mathrm{i}}$	${}_{\mathrm{T}}\bar{\mathrm{y}} = \frac{1}{\mathrm{m}}\sum_{j=1}^{\mathrm{m}} {}_{\mathrm{T}}\mathrm{y}_j$ [r.v. is ${}_{\mathrm{T}}\bar{Y}$ with value ${}_{\mathrm{T}}\bar{y}$]
Average	$\bar{\mathbf{Ɏ}} = \sum_{\mathrm{i}=1}^{\mathbf{M̵}} {}_{\mathrm{T}}\mathbf{Ɏ}_{\mathrm{i}} \Big/ \sum_{\mathrm{i}=1}^{\mathbf{M̵}} \mathbf{Ł}_{\mathrm{i}}$	$\bar{\mathrm{y}}_{uc} = \sum_{j=1}^{\mathrm{m}} {}_{\mathrm{T}}\mathrm{y}_j \Big/ \sum_{j=1}^{\mathrm{m}} \mathrm{l}_j$ [r.v. is $\bar{Y}_{uc}$ with value $\bar{y}_{uc}$]
Total	${}_{\mathrm{T}}\mathbf{Ɏ} = \sum_{\mathrm{i}=1}^{\mathbf{M̵}} {}_{\mathrm{T}}\mathbf{Ɏ}_{\mathrm{i}} = \mathbf{N̵}\bar{\mathbf{Ɏ}} = \mathbf{M̵}\,{}_{\mathrm{T}}\bar{\mathbf{Ɏ}}$	${}_{\mathrm{T}}\mathrm{y} = \mathbf{N̵}\bar{\mathrm{y}}_{uc}$ (if $\mathbf{N̵}$ is known) ${}_{\mathrm{T}}\mathrm{y} = \mathbf{M̵}\,{}_{\mathrm{T}}\bar{\mathrm{y}}$ (if $\mathbf{N̵}$ is *un*known) [r.v. is ${}_{\mathrm{T}}Y$ with value ${}_{\mathrm{T}}y$]

8. Appendix 2: Notation – Managing Falsehood

The foregoing eleven tables from the STAT 332 Course Materials occur near the start of the relevant Figure to summarize the notation that will be used in that Figure. The tables also have a broader purpose – they provide continuing emphasis on maintaining the three distinctions, listed on the lower half of page TND-3, which are vital for effective statistics teaching. The tables have the following features:

- the terms ‘respondent population’, ‘sample’ and ‘model’ are prominent in the column and row headings,
- the sample is taken as the source of real-world data, as is the case in most data-based investigating,
- the respondent population columns notation is upper-case **bold** letters with high visual impact,
- the sample columns have low-visual-impact lower-case Roman letters, emphasizing the population-sample distinction
- the model columns are distinguished by their *italic* letters and (usually) ‘random variable’ (r.v.) label(s),
- as illustrated in equations (2) and (3) at the right [taken from the upper half of page 2.125 in Figure 2.16], a quantity and its estimate are notationally *different* in appearance, a useful reminder (*not* dependent on knowing the *meanings* of the symbols) of the distinction between behaviour under repetition and the individual case.

$$s.d.(R) \simeq \sqrt{\frac{1}{(\mathbf{N̵}-1)\bar{\mathbf{X̵}}^2}\sum_{\mathrm{i}=1}^{\mathbf{N̵}}(\mathbf{Ɏ}_{\mathrm{i}} - \mathbf{R̵}\mathbf{X̵}_{\mathrm{i}})^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N̵}})} \qquad \text{-----(2)}$$

$$\hat{s.d.}(R) \simeq \sqrt{\frac{1}{(\mathrm{n}-1)\bar{x}^2}\sum_{j=1}^{\mathrm{n}}(y_j - r x_j)^2(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N̵}})} \qquad \text{-----(3)}$$

This illustration is for standard deviation, but the same is true of estimating bias, with multiple examples in Figures 2.15, 2.16 and 2.17; the difference in visual impact is higher in the more complicated expressions in these three later figures than in the first instance encountered as equations (2.3.9) and (2.3.17) on page 2.41 in Figure 2.3.

Distiguishing *observed* (Roman) y from a *model* (italic) *y* to distinguish the observed sample average and standard deviation, $\bar{\mathrm{y}}$ and s, from the model quantities $\bar{y}$ and *s*, reminds us of the importance in the Design stage of the FDEAC cycle of developing a Plan that makes it *reasonable* to treat $\bar{\mathrm{y}}$ as $\bar{y}$ and s as *s* (as in, for example, Table 2.3.5 on the upper half of page TND-7).

Failure to observe the foregoing distinctions makes it easier to overlook the question to be asked about *all* data: *How were they generated?* If this question is *not* asked, it becomes more likely that statistical methods will be applied in situations where their underlying assumptions (like probability selecting, accurate measuring processes, independent measuring, distributional assumptions) are violated, leading to answers with such severe limitations that they are likely to embody falsehoods.

- Our error categories (for example, study error, non-response error, sample error, measurement error and model error in the case of sample surveys) provide a framework for assessing the severity of the limitations imposed by error on answers.