University of Waterloo W. H. Cherry

ERROR: Bias and Rms Error

For a random variable Y and some constant c (and where 'E' denotes probabilistic 'expectation'), we have:

$$\begin{aligned} E\{[Y-\mathrm{c}]^2\} &= E\{[E(Y)-\mathrm{c}+Y-E(Y)]^2\} = E\{[E(Y)-\mathrm{c}]^2+[Y-E(Y)]^2+2[E(Y)-\mathrm{c}][Y-E(Y)]\} \\ &= E\{[E(Y)-\mathrm{c}]^2\}+E\{[Y-E(Y)]^2\}+2E\{[E(Y)-\mathrm{c}][Y-E(Y)]\} \\ &= [E(Y-\mathrm{c})]^2+E\{[Y-E(Y)]^2\}+2[E(Y)-\mathrm{c}]E[Y-E(Y)] \\ \textit{i.e.,}\quad E\{[Y-\mathrm{c}]^2\} &= [E(Y-\mathrm{c})]^2+[s.d.(Y)]^2 \qquad \text{because } E[Y-E(Y)]\equiv 0. \qquad \text{-----(HL7.1)} \end{aligned}$$

If we now think of Y as a random variable whose distribution represents the possible values of a *response variate* $\mathbf{Y}$ and c as a *true* value, the left-hand side of equation (HL7.1) is a **mean squared error** and $E(Y-\mathrm{c})$ in the first term on the right-hand side is a **bias**; we can therefore interpret equation (HL7.1) as:

$$\text{mean squared error} = \text{bias}^2 + \text{standard deviation}^2. \qquad \text{-----(HL7.2)}$$

Taking the square root so we are working on the *same* scale as the variate represented by Y, the **root mean squared error** is:

$$\text{rms error} = \sqrt{\text{bias}^2 + \text{standard deviation}^2}. \qquad \text{-----(HL7.3)}$$

Thus, the rms error is *one* concept that *combines* the two model quantities of bias and (probabilistic) standard deviation, or it can be used as a model for the two corresponding real-world entities of inaccuracy and imprecision.

Equation (HL7.3) provides useful insights about bias and variation in the context of survey sampling (to answer a Question with a *descriptive* aspect); different cases depend on how broad our focus is in terms of *which* true value c represents.

∗ The narrowest focus is *measuring* when c is the true value of the response variate $\mathbf{Y}$; equation (HL7.3) is then:

$$\text{measuring rms error} = \sqrt{\text{measuring bias}^2 + \text{measuring standard deviation}^2}. \qquad \text{-----(HL7.4)}$$

∗ For measuring *and* sampling, c is the true value of the *respondent* population attribute of $\mathbf{Y}$ and then:

$$\begin{array}{c}\text{measuring \textit{and} sampling}\\ \text{rms error}\end{array} = \sqrt{\text{measuring + sampling bias}^2 + \text{measuring and sampling standard deviation}^2}; \qquad \text{-----(HL7.5)}$$

NOTE: 1. $\begin{array}{c}\text{Measuring and sampling}\\ \text{standard deviation}\end{array} = \sqrt{\text{measuring standard deviation}^2 + \text{sampling standard deviation}^2}. \qquad$ -----(HL7.6)

∗ For measuring *and* sampling *and* non-responding, c is the true value of the *study* population attribute of $\mathbf{Y}$ and then, under our assumption that non-response is *deterministic* (*not* stochastic – see Note 3 below and overleaf on page HL7.2):

$$\begin{array}{r}\text{measuring and sampling and}\\ \text{non-responding rms error}\end{array} = \sqrt{\begin{array}{l}\text{measuring + sampling}\\ \text{+ non-responding bias}^2\end{array} + \text{measuring and sampling standard deviation}^2}. \qquad \text{-----(HL7.7)}$$

∗ For measuring *and* sampling *and* non-responding *and* specifying, c is the true value of the *target* population attribute of $\mathbf{Y}$ and then, under our assumption that specifying the study population also is *deterministic*:

$$\begin{array}{r}\text{measuring and sampling}\\ \text{and non-responding and}\\ \text{studying rms error}\end{array} = \sqrt{\begin{array}{l}\text{measuring + sampling}\\ \text{+ non-responding}\\ \text{+ studying bias}^2\end{array} + \text{measuring and sampling standard deviation}^2}. \qquad \text{-----(HL7.8)}$$

NOTES: 2. In printed materials other than these Statistical Highlights (*e.g.*, see Cochran, p. 15), equation (HL7.1) [or (HL7.2)] is is usually discussed only with respect to *estimating* bias. Although we have relatively little to say about estimating bias in these Statistical Highlights, it is useful to recognize the following [see also Statistical Highlight #21]:

- **Estimating bias** (a *model* quantity) is the difference between the mean of an estimator and the value of the corresponding population attribute (or model parameter – see the schema in Note 5 overleaf on page HL7.2); for example, under EPS:
  - the random variable $\overline{Y}$ representing the sample *average* $\overline{y}$ is an *un*biased estimator of the respondent population average $\overline{\mathbf{Y}}$ because $E(\overline{Y}) = \overline{\mathbf{Y}}$ or $E(\overline{Y}) - \overline{\mathbf{Y}} = 0$; BUT
  - the sample ratio $r = \overline{y}/\overline{x}$ is a *biased* estimator of the respondent population ratio $\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}}$ because $E(R) \neq \mathbf{R}$ or $E(R) - \mathbf{R} \neq 0$, and likewise for S as an estimator of the respondent population (data) standard deviation $\mathbf{S}$ (although S^2 is an *un*biased estimater of $\mathbf{S}^2$) [*e.g.*, see the top and bottom of page HL77.5 in Section 5 and Appendix 3 on page HL77.9 in Statistical Highlight #77].
- The rms error of an estimator is of interest because, while we prefer an *un*biased estimator of a population attribute, there are times when a *biased* estimator has only *small* bias and appreciably *smaller* standard deviation than an available *un*biased estimator; we *may* then prefer the biased estimator with *smaller* rms error.
- *Un*like (real-world) inaccuracy, estimating bias *de*creases in magnitude with increasing sample size – see, for example, Note 21 starting near the bottom of page HL77.9 in Statistical Highlight #77.

REFERENCE: Cochran, W. G.: *Sampling Techniques*. Third edition, John Wiley & Sons, New York, 1977, ISBN 0-471-16240-X.

3. Which elements of the study population fall in the respondent and non-respondent populations depends on the *incentives* offered for response – different incentives will presumably, in general, result in *different* sets of the study

(*continued overleaf*)

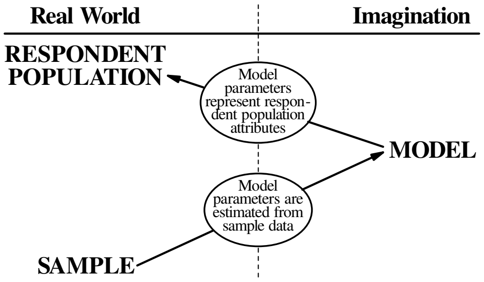
NOTES: (cont.) 3. population elements in the two populations. For *given* incentives for response (as specified in the protocol for selecting elements in a *particular* investigation), statistical theory to manage non-response error can be based on:

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

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

4. Equation (HL7.8) overleaf on page HL7.1 can be thought of as an extension to *variation* of the behaviour of averages (when answering a Question with a *descriptive* aspect) in equation (HL18.1) and its pictorial version at the lower right of page HL18.4 in Section 5 of Statistical Highlight #18.

5. Mathematical models are all idealizations and all are products of the intellect and the imagination; as shown in the schema at the right, *we* think of the model as a *link* between the *sample* and the *respondent population*.



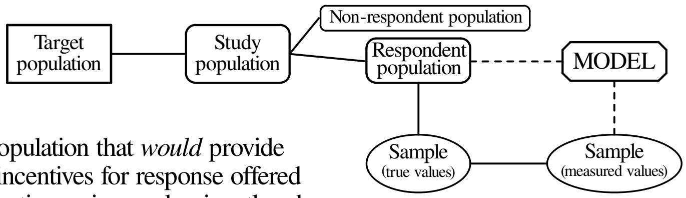
To maintain the distinction between the real world (represented by the data) and the model, we use different words – 'average' and 'mean' – for their measures of location; unfortunately, we do not have this option for the two measures of variation, which are both called 'standard deviation'. In the early 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 HL7.1 at the right.

Table HL7.1

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

Appendix [optional reading]

The schema at the right below shows five groups of elements which *we* distinguish for data-based investigating and which are involved in the definitions of *our* six error categories (*e.g.*, given on the lower half of page HL8.1 in Statistical Highlight #8). Also, the model is included in the schema, as a *link* between the sample and the respondent population (as in Note 5 above).



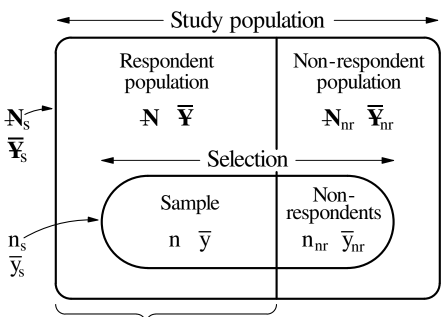
* **Target population**: the group of elements to which the investigator(s) want Answer(s) to the Question(s) to apply.
* **Study population**: a group of elements *available* to an investigation.
* **Respondent population:** those elements of the study population that *would* provide the data requested 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 6 below.
* **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).

NOTE: 6. 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 $\cancel{\bar{\mathbf{Y}}}$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:

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



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