University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: Standard Deviation and Standard Pairwise Difference as Measures of Variation

1. Introduction

Variation is a key concern of statistical methods; *our* definition is:

∗ **Variation:** differences in (variate or attribute) values:

- across the individuals in a group, such as;
  - ○ a target population/process, ○ a study population/process, ○ a respondent population,
  - ○ a non-respondent population, ○ a sample;
- arising under repetition [*e.g.*, for a (repeated) measurement or error or a sample average].

We distinguish variation from **variability**, which is the model quantity representing imprecision.

A reminder (where needed) of terminology used in these definitions in given in Appendix 1 starting on page HL100.3.

The discussion which follows in Sections 2 to 4 has the commonly-encountered context of a sample of n elements; however, the essential statistical points would be *un*affected if the context were instead a respondent population of size $\mathbf{N}$ elements or even a study or a target population of respective size $\mathbf{N}_S$ or $\mathbf{N}_T$ elements. Similarly, instead of response variate values y_j, the discussion could equally well involve values x_j or z_{ij} of a focal or non-focal explanatory variate (our $\mathbf{X}$ and $\mathbf{Z}$).

2. Quantifying Data Variation

A familiar measure of variation in statistics is the **standard deviation** (abbreviated 's.d.') – these Statistical Highlights recommend adjectives of 'data' and 'probabilistic' as qualifiers of standard deviation to maintain a useful distinction; *we* also use a symbol (like s) for the former and s.d. [like *s.d.*(*Y*)] for the latter – see Appendix 2 starting on page HL100.5. Our concern in this Highlight #100 is with *data* standard deviation which, for the response variate values of a sample of n elements, is defined as in equation (HL100.1) at the right (as its use of Roman letters y and j indicates); the second form of this equation reminds us of the notation SS_y for the sum of squared deviations from the average.

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2} \equiv \sqrt{\frac{1}{n-1}SS_y} \qquad \text{-----(HL100.1)}$$

Another measure of (data) variation, similar in intent to the standard deviation, is the **standard pairwise difference** (abbreviated 's.p.d.') which, for the response variate values of a sample of n elements, is defined as in equation (HL100.2) at the right. It is noteworthy for the following reasons.

$$s_{pd} = \sqrt{\frac{2}{n(n-1)}\sum_{j<k=1}^{n}(y_j - y_k)^2} \equiv \sqrt{\frac{2}{n(n-1)}SSP_y} \qquad \text{-----(HL100.2)}$$

- It is seldom discussed and so is unfamiliar, even to some statisticians.
- It does not appear to have an agreed-on name – s.p.d. is used in these Statistical Highlights, together with s_{pd} as its symbol.
- The pairwise differences comprise an *un*ordered set and so number $\binom{n}{2} = ½n(n-1)$ differences; the corresponding *ordered* set contains twice as many pairwise differences, although *we* are not concerned with it here (but see Note 3 on page HL100.10).
- The second form of equation (HL100.2) defines the notation SSP_y for the sum of (unordered) squared pairwise differences.
- The adjective 'standard' in *our* name indicates the *same* three processes that are involved in calculating standard deviation:
  - squaring each relevant entity – here, each pairwise difference;
  - calculating the 'average' of these squared values – in contrast to the standard deviation, the denominator for averaging in the s.p.d. *is* the number of entities in the sum, $\binom{n}{2}$, *not* a smaller number (like n – 1 rather than n);
  - taking the square root of this 'average' square.
- The use of a more familiar denominator in the averaging, together with using differences among all pairs of data values (instead of deviations from the average) make the standard pairwise difference a more 'natural' measure of (data) variation.
- Anticipating Section 3 which begins below, as given in equation (HL100.3) at the right, the s.d. is *smaller* than the s.p.d. by a factor of $1/\sqrt{2}$ – that is, the s.d. is about 30% smaller in magnitude; the s.d. squared is *half* the s.p.d. squared.

$$\sqrt{2}s = s_{pd} \ \text{ or: } \ 2s^2 = s_{pd}^2 \qquad \text{-----(HL100.3)}$$

  - It follows from equations (HL100.1), (HL100.2) and (HL100.3) that $SS_y = SSP_y/n$, as given in equation (HL100.4).

$$SS_y = SSP_y/n \qquad \text{-----(HL100.4)}$$

NOTES: 1. The 'd' in the abbreviations s.d. and s.p.d. stands for, respectively, 'deviation' and 'difference'; regrettably, 'deviation from the average' is long-established statistical useage, even though 'difference from the average' is equally appropriate.

2. Although peripheral to this Highlight #100, accurate *numerical* calculation of s.d.s (and s.p.d.s) is potentially of concern; simple examples in the classroom (*e.g.*, based on Σy_j and Σy_j^2) can obscure the serious inaccuracy that can arise when taking differences of many large and nearly equal (floating point) numbers – *e.g.*, see Chan, T.F., Golub, G.H. and R.J. LeVeque, Algorithms for Computing the Sample Variance: Analysis and Recommendations. *The American Statistician* **37**(#3) 242-247 (1983). Stable URL: http://www.jstor.org/stable/2683386

3. The Relationship Between the S.d. and the S.p.d.

We demonstrate the relationship in equation (HL100.3) above by considering successive values of n, starting with n = 2; for convenience, we actually demonstrate the relationship in equation (HL100.4), from which equation (HL100.3) follows.

Case 1: $n=2$. $\quad SS_y = (y_1-\bar{y})^2+(y_2-\bar{y})^2 = (y_1-\frac{y_1+y_2}{2})^2+(y_2-\frac{y_1+y_2}{2})^2 = (\frac{y_1}{2}-\frac{y_2}{2})^2+(\frac{y_2}{2}-\frac{y_1}{2})^2 = \frac{1}{2}(y_1-y_2)^2 = SSP_y/n$ -----(HL100.4a)

We note here how *two* terms of SS_y become *one* term of SSP_y and use of the identity: $(y_j-y_k)^2 \equiv (y_k-y_j)^2$.

We also note in Cases 2 to 4 below that, when $n=3$, three terms of SS_y remain three terms of SSP_y, when $n=4$, four terms of SS_y become six terms of SSP_y and, in general, n terms of SS_y become ½n(n−1) terms of SSP_y.

Case 2: $n=3$.

$$\begin{aligned}
SS_y &= (y_1-\bar{y})^2+(y_2-\bar{y})^2+(y_3-\bar{y})^2\\
&= (y_1-\tfrac{y_1+y_2+y_3}{3})^2+(y_2-\tfrac{y_1+y_2+y_3}{3})^2+(y_3-\tfrac{y_1+y_2+y_3}{3})^2\\
&= (\tfrac{2y_1}{3}-\tfrac{y_2}{3}-\tfrac{y_3}{3})^2+(\tfrac{2y_2}{3}-\tfrac{y_1}{3}-\tfrac{y_3}{3})^2+(\tfrac{2y_3}{3}-\tfrac{y_1}{3}-\tfrac{y_2}{3})^2\\
&= \tfrac{1}{9}[(y_1-y_2)+(y_1-y_3)]^2+\tfrac{1}{9}[(y_2-y_1)+(y_2-y_3)]^2+\tfrac{1}{9}[(y_3-y_1)+(y_3-y_2)]^2\\
&= \tfrac{1}{9}[(y_1-y_2)^2+(y_1-y_3)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_1-y_3)^2+(y_2-y_3)^2]\\
&\quad + \tfrac{2}{9}[y_1^2-y_1y_2-y_1y_3+\cancel{y_2y_3}+y_2^2-\cancel{y_1y_2}-\cancel{y_2y_3}+\cancel{y_1y_3}+y_3^2-\cancel{y_1y_3}-y_2y_3+\cancel{y_1y_2}]\\
&= \tfrac{2}{9}[(y_1-y_2)^2+(y_1-y_3)^2+(y_2-y_3)^2]+\tfrac{1}{9}[(y_1-y_2)^2+(y_1-y_3)^2+(y_2-y_3)^2]\\
&= \tfrac{1}{3}[(y_1-y_2)^2+(y_1-y_3)^2+(y_2-y_3)^2] = SSP_y/n \qquad \text{-----(HL100.4b)}
\end{aligned}$$

Case 3: $n=4$.

$$\begin{aligned}
SS_y &= (y_1-\bar{y})^2+(y_2-\bar{y})^2+(y_3-\bar{y})^2+(y_4-\bar{y})^2\\
&= (y_1-\tfrac{y_1+y_2+y_3+y_4}{4})^2+(y_2-\tfrac{y_1+y_2+y_3+y_4}{4})^2+(y_3-\tfrac{y_1+y_2+y_3+y_4}{4})^2+(y_4-\tfrac{y_1+y_2+y_3+y_4}{4})^2\\
&= (\tfrac{3y_1}{4}-\tfrac{y_2}{4}-\tfrac{y_3}{4}-\tfrac{y_4}{4})^2+(\tfrac{3y_2}{4}-\tfrac{y_1}{4}-\tfrac{y_3}{4}-\tfrac{y_4}{4})^2+(\tfrac{3y_2}{4}-\tfrac{y_1}{4}-\tfrac{y_2}{4}-\tfrac{y_4}{4})^2+(\tfrac{3y_4}{4}-\tfrac{y_1}{4}-\tfrac{y_2}{4}-\tfrac{y_3}{4})^2\\
&= \tfrac{1}{16}[(y_1-y_2)+(y_1-y_3)+(y_1-y_4)]^2+\tfrac{1}{16}[(y_2-y_1)+(y_2-y_3)+(y_2-y_4)]^2\\
&\qquad +\tfrac{1}{16}[(y_3-y_1)+(y_3-y_2)+(y_3-y_4)]^2+\tfrac{1}{16}[(y_4-y_1)+(y_4-y_2)+(y_4-y_3)]^2\\
&= \tfrac{1}{16}[(y_1-y_2)^2+(y_1-y_3)^2+(y_1-y_4)^2+(y_1-y_2)^2+(y_2-y_3)^2+(y_2-y_4)^2\\
&\qquad +(y_1-y_3)^2+(y_2-y_3)^2+(y_3-y_4)^2+(y_1-y_4)^2+(y_2-y_4)^2+(y_3-y_4)^2]\\
&\quad + \tfrac{2}{16}[3y_1^2-y_1y_2-y_1y_3+\cancel{y_2y_3}-y_1y_2-y_1y_4+\cancel{y_2y_4}-y_1y_3-y_1y_4+\cancel{y_3y_4}\\
&\qquad +3y_2^2-\cancel{y_1y_2}-\cancel{y_2y_3}+\cancel{y_1y_3}-\cancel{y_2y_4}-\cancel{y_1y_2}+\cancel{y_1y_4}-\cancel{y_2y_3}-\cancel{y_2y_4}+\cancel{y_3y_4}\\
&\qquad +3y_3^2-\cancel{y_1y_3}-y_2y_3+\cancel{y_1y_2}-\cancel{y_1y_3}-\cancel{y_3y_4}+\cancel{y_1y_4}-y_2y_3-\cancel{y_3y_4}+\cancel{y_2y_4}\\
&\qquad +3y_4^2-\cancel{y_1y_4}-y_2y_4+\cancel{y_1y_2}-y_3y_4-\cancel{y_1y_4}+\cancel{y_1y_3}-y_2y_4-y_3y_4+\cancel{y_2y_3}]\\
&= \tfrac{2}{16}[(y_1-y_2)^2+(y_1-y_3)^2+(y_1-y_4)^2+(y_2-y_3)^2+(y_2-y_4)^2+(y_3-y_4)^2]\\
&\qquad +\tfrac{2}{16}[(y_1-y_2)^2+(y_1-y_3)^2+(y_1-y_4)^2+(y_2-y_3)^2+(y_2-y_4)^2+(y_3-y_4)^2]\\
&= \tfrac{1}{4}[(y_1-y_2)^2+(y_1-y_3)^2+(y_1-y_4)^2+(y_2-y_3)^2+(y_2-y_4)^2+(y_3-y_4)^2] = SSP_y/n \qquad \text{-----(HL100.4c)}
\end{aligned}$$

Case 4: $n=n$.

$$\begin{aligned}
SS_y &= (y_1-\bar{y})^2+(y_2-\bar{y})^2+\ldots\ldots+(y_{n-1}-\bar{y})^2+(y_n-\bar{y})^2\\
&= (y_1-\tfrac{y_1+y_2+\ldots\ldots+y_{n-1}+y_n}{n})^2+(y_2-\tfrac{y_1+y_2+\ldots\ldots+y_{n-1}+y_n}{n})^2+\ldots\ldots+(y_{n-1}-\tfrac{y_1+y_2+\ldots\ldots+y_{n-1}+y_n}{n})^2+(y_n-\tfrac{y_1+y_2+\ldots\ldots+y_{n-1}+y_n}{n})^2\\
&= (\tfrac{(n-1)y_1}{n}-\tfrac{y_2}{n}-\ldots\ldots-\tfrac{y_{n-1}}{n}-\tfrac{y_n}{n})^2+(\tfrac{(n-1)y_2}{n}-\tfrac{y_1}{n}-\tfrac{y_3}{n}-\ldots\ldots-\tfrac{y_{n-1}}{n}-\tfrac{y_n}{n})^2+\ldots\ldots\\
&\qquad +(\tfrac{(n-1)y_{n-1}}{n}-\tfrac{y_1}{n}-\ldots\ldots-\tfrac{y_{n-2}}{n}-\tfrac{y_n}{n})^2+(\tfrac{(n-1)y_n}{n}-\tfrac{y_1}{n}-\ldots\ldots-\tfrac{y_{n-2}}{n}-\tfrac{y_{n-1}}{n})^2\\
&= \tfrac{1}{n^2}[(y_1-y_2)+(y_1-y_3)+\ldots\ldots+(y_1-y_{n-1})+(y_1-y_n)]^2+\tfrac{1}{n^2}[(y_2-y_1)+(y_2-y_3)+\ldots\ldots+(y_2-y_{n-1})+(y_2-y_n)]^2+\ldots\ldots\\
&\quad +\tfrac{1}{n^2}[(y_{n-1}-y_1)+(y_{n-1}-y_2)+\ldots\ldots+(y_{n-1}-y_{n-2})+(y_{n-1}-y_n)]^2+\tfrac{1}{n^2}[(y_n-y_1)+(y_n-y_2)+\ldots\ldots+(y_n-y_{n-2})+(y_n-y_{n-1})]^2\\
&= \tfrac{1}{n^2}[(y_1-y_2)^2+(y_1-y_3)^2+\ldots\ldots+(y_1-y_{n-1})^2+(y_1-y_n)^2+(y_2-y_1)^2+(y_2-y_3)^2+\ldots\ldots+(y_2-y_{n-1})^2+(y_2-y_n)^2+\ldots\ldots\\
&\quad +(y_1-y_{n-1})^2+(y_2-y_{n-1})^2+\ldots\ldots+(y_{n-2}-y_{n-1})^2+(y_{n-1}-y_n)^2+(y_1-y_n)^2+(y_2-y_n)+\ldots\ldots+(y_{n-2}-y_n)^2+(y_{n-1}-y_n)^2]\\
&\quad +\tfrac{n-2}{n^2}(n-1)y_1^2-\tfrac{2}{n^2}[y_1y_2+y_1y_3-y_2y_3+y_1y_2+y_1y_4-y_2y_4+\ldots\ldots+y_1y_{n-2}+y_1y_n-y_{n-2}y_n+y_1y_{n-1}+y_1y_n-y_{n-1}y_n]\\
&\quad +\tfrac{n-2}{n^2}(n-1)y_2^2-\tfrac{2}{n^2}[y_1y_2+y_2y_3-y_1y_3+y_2y_4+y_1y_2-y_1y_4+\ldots\ldots+y_2y_{n-2}+y_2y_n-y_{n-2}y_n+y_2y_{n-1}+y_2y_n-y_{n-1}y_n]\\
&\quad +\ldots\ldots+\tfrac{n-2}{n^2}(n-1)y_j^2-\tfrac{2}{n^2}[y_1y_j+y_2y_j-y_1y_2+y_1y_j+y_3y_j-y_1y_3+\ldots\ldots]+\ldots\ldots\\
&\quad +\tfrac{n-2}{n^2}(n-1)y_{n-1}^2-\tfrac{2}{n^2}[y_1y_{n-1}+y_2y_{n-1}-y_1y_2+y_1y_{n-1}+y_3y_{n-1}-y_1y_3+\ldots\ldots+y_{n-3}y_{n-1}+y_{n-1}y_n-y_{n-3}y_n+y_{n-2}y_{n-1}+y_{n-1}y_n-y_{n-2}y_{n-1}]\\
&\quad +\tfrac{n-2}{n^2}(n-1)y_n^2-\tfrac{2}{n^2}[y_1y_n+y_2y_n-y_1y_2+y_1y_n+y_3y_n-y_1y_3+\ldots\ldots+y_{n-3}y_n+y_{n-1}y_n-y_{n-3}y_{n-1}+y_{n-2}y_n+y_{n-1}y_n-y_{n-2}y_{n-1}]\\
&= \tfrac{2}{n^2}[(y_1-y_2)^2+(y_1-y_3)^2+\ldots\ldots+(y_1-y_{n-1})^2+(y_1-y_n)^2+(y_2-y_3)^2+(y_2-y_4)^2+\ldots\ldots+(y_2-y_{n-1})^2+(y_2-y_n)^2+\ldots\ldots\\
&\qquad +(y_{n-3}-y_{n-2})^2+(y_{n-3}-y_{n-1})^2+(y_{n-3}-y_n)^2+(y_{n-2}-y_{n-1})^2+(y_{n-2}-y_n)^2+(y_{n-1}-y_n)^2]\\
&\quad +\tfrac{n-2}{n^2}[(y_1-y_2)^2+(y_1-y_3)^2+\ldots\ldots+(y_1-y_{n-1})^2+(y_1-y_n)^2+(y_2-y_3)^2+(y_2-y_4)^2+\ldots\ldots+(y_2-y_{n-1})^2+(y_2-y_n)^2+\ldots\ldots\\
&\qquad +(y_{n-3}-y_{n-2})^2+(y_{n-3}-y_{n-1})^2+(y_{n-3}-y_n)^2+(y_{n-2}-y_{n-1})^2+(y_{n-2}-y_n)^2+(y_{n-1}-y_n)^2]\\
&= \tfrac{1}{n}[(y_1-y_2)^2+(y_1-y_3)^2+\ldots\ldots+(y_1-y_{n-1})^2+(y_1-y_n)^2+(y_2-y_3)^2+(y_2-y_4)^2+\ldots\ldots+(y_2-y_{n-1})^2+(y_2-y_n)^2+\ldots\ldots\\
&\qquad +(y_{n-3}-y_{n-2})^2+(y_{n-3}-y_{n-1})^2+(y_{n-3}-y_n)^2+(y_{n-2}-y_{n-1})^2+(y_{n-2}-y_n)^2+(y_{n-1}-y_n)^2] = SSP_y/n
\end{aligned}$$

The three demonstrations above for $n=3$, 4 and n are each set out in seven steps (indicated by '=') with noteworthy points as follows:

- Deviations from the average in SS_y in step 1 have become pairwise differences by step 4, which consists of coefficients of $1/n^2$ of each of n terms squared, each containing the sum of $n-1$ paired differences;

(*continued*)

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: Standard Deviation and ..... Difference (continued 1)

- Multiplying out the n terms squared in going from step 4 to step 5 generates two types of paired difference terms:
  - $n-1$ paired differences *squared*, making $n(n-1)$ such terms in total which are $2\binom{n}{2}$ paired differences squared and comprise *two* copies of SSP_y;
  - $(n-1)(n-2)$ *cross products* of paired differences, making $n(n-1)(n-2)$ such terms in total – multiplied out, these cross products generate $4n(n-1)(n-2)$ individual terms which comprise:
    - ○ $n(n-1)(n-2)$ y_j^2 terms made up of n distinct such terms occurring $n-1$ times in each of $n-2$ copies;
    - ○ $3n(n-1)(n-2)$ $y_j y_k$ terms made up of three sets of $\binom{n}{2}$ distinct such terms occurring twice in each of $n-2$ copies;
      - + $2n(n-1)(n-2)$ of the $y_j y_k$ terms have a '–' sign, $n(n-1)(n-2)$ have a '+' sign and cancel half the terms with a '–' sign, leaving $n(n-1)(n-2)$ such terms with a '–' sign – they comprise two identical sets of $\binom{n}{2}$ terms $y_j y_k$ in $n-2$ copies;
      - + only one of the two identical sets of $\frac{1}{2}n(n-1)(n-2)$ $y_j y_k$ terms remaining after the cancellation is shown in each step 5, with coefficient 2 in the numerator outside the square brackets [ ];
      - + the $n(n-1)(n-2)$ cancellations are shown in step 5 of Cases 2 and 3 but it is not feasible to show them in Case 4.

    Taken together, these terms y_j^2 and $y_j y_k$ form $n-2$ copies of SSP_y.

    The fact that $n-1=2$ in Case 2 and $n-2=2$ in Case 3 allows for simpler grouping of the y_j^2 and $y_j y_k$ terms in step 5 of these two cases compared with the general Case 4.

The numbers of terms y_j^2 and $y_j y_k$ given above are those required to contribute $n-2$ copies of SSP_y to the relationship in equation (HL100.4) on page HL100.1 because:

- the $\binom{n}{2}$ paired differences squared of SSP_y, when multiplied out, generate $4\binom{n}{2} = 2n(n-1)$ individual terms in total, which divide equally into $n(n-1)$ of both y_j^2 and $y_j y_k$; hence, $n-2$ copies of SSP_y involve $n(n-1)(n-2)$ of each of the two types of term;
  - the $n(n-1)(n-2)$ terms y_j^2 are made up of n distinct such terms occurring $n-1$ times in each of the $n-2$ copies of SSP_y;
  - the $n(n-1)(n-2)$ terms $-y_j y_k$ are made up of $\binom{n}{2}$ distinct such terms occurring twice in each of the $n-2$ copies of SSP_y.

4. Summary

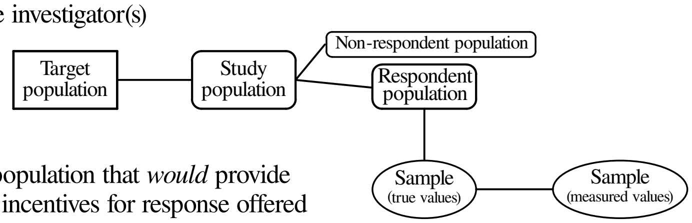
The foregoing discussion involves the following matters of statistical interest.

- Variation is a key concern of statistical methods.
- Two (of several possible) ways of quantifying data variation are the (familiar) s.d. and the (*un*familiar) s.p.d.
  - These two attributes are similar in intent and equivalent algebraically but differ numerically by a factor of $1/\sqrt{2} \simeq 0.7071$.
- For the process of averaging, comparing the s.d. and the s.p.d. provides a rationale for a denominator that reflects the number of constraints on the entities in the numerator (and, conceivably, an antidote for introductory texts' obsession with $n-1$ *vs.* n).
  - For the s.d., the n deviations from the average in the numerator add to zero *regardless* of the particular set of n data values – this *one* constraint is reflected in the $n-1$ (rather than n) denominator, often described (unhelpfully) in statistics as 'losing one degree of freedom'. The idea that constraints determine the denominator appropriate for averaging occurs in other statistical methods; for example, the method of 'least squares' for fitting 'regression' models and 'analysis of variance'.
    - ○ This latter name, despite its enticing acronym, 'ANOVA', is also unhelpful terminology; 'variance' (as normally understood) is *not* involved and 'partitioning sums of squared differences' describes more closely what is actually involved, but there is little hope of replacing the enshrined widespread use of 'ANOVA' by the evocative but unappealing 'PASSDI'.
      - + Squared differences divided by their 'degrees of freedom' are (carelessly) called 'mean (instead of average) squares'.
  - The denominator of a data s.d. is a source of much (unhelpful) 'explanation' (and student anguish) in introductory statistics.
- Suitable algorithms are needed to avoid numerical inaccuracy when calculating s.d. (and s.p.d.), most notably for large data sets.

It is unfortunate that the s.p.d. is rarely even mentioned in introductory statistics teaching.

5. Appendix 1 – Limited Glossary [A more detailed Glossary is given in Statistical Highlight #91.]

The schema at the right below reminds us that data-based investigating is concerned primarily with five groups of elements:



∗ **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 occur for *in*animate elements].

∗ **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.

(*continued overleaf*)

∗ **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 overleaf at the lower right reminds us).

∗ **Attribute**: a quantity defined as a function of response (and, perhaps, explanatory) variates over the five groups of elements defined overleaf at the bottom of page HL100.3 and above.

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 – *e.g.*, see Statistical Highlight #6.

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

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

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

∗ **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. [The population-process distinction is discussed in Statistical Highlight #94.]

∗ **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 *explanatory* variate whose relationship to the *response* variate is involved in the Answer(s) to the Question(s).

∗ **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.

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

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

∗ **Model:** a provisional idealized symbolic description of a real-world phenomenon.

We use Greek letters for **parameters** of statistical and probability models (*e.g.*, μ, σ and π).
– **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 an investigation.
The first response model in the STAT 231 Course Notes is the one in equation (HL100.6) on the facing page HL100.5.

In *any* situation where an Answer is based, in whole or in part, on a mathematical model, we should bear in mind a maxim of the late Dr. George E. P. Box, a respected U.S. statistician: *All model are wrong, some are useful.*

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

Repetition is inherent in our definitions of a confidence interval, an estimator, inaccuracy and imprecision.

∗ **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 the response model in equation (HL100.6) on the facing page HL100.5.

The idea of independence is a (mathematical) *idealization* – the usual state of affairs in the real world is one of ***dependence*** (*i.e.*, *lack* of independence).

∗ **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 Statistical Highlight #6.

Use of the terminology in this Glossary, and of some useful notation, is illustrated in Appendix 2 which starts on the facing page HL100.5; Statistical Highlight #91 provides a more detailed Glossary.

University of Waterloo W. H. Cherry

STATISTICS and STATISTICAL METHODS: Standard Deviation and ..... Difference (continued 2)

6. Appendix 2 – S.d. in a Broader Statistical Context

This Appendix 2, slightly adapted from material in Statistical Highlight #77, provides a broader context for the role of the s.d. in statistical methods than do Sections 1 to 4 of this Highlight #100, which are primarily concerned with using the s.d. (and the s.p.d.) to quantify (data) variation. This broader context is using sample attributes as *estimates* of the corresponding (respondent) population attributes, based on a *model* involving unit (or element) inclusion probabilities arising from the process of *equiprobable* selecting (EPS) of the sample.

✻ In introductory probability, we use *upper* case *italic* letters (usually near the end of the alphabet, like X, Y and Z) for random variables and the corresponding *lower* case letters (*e.g.*, x, y and z) for their values. *We* further distinguish *upper* case **bold** letters for *population* quantities; for example, $\mathbf{Y}_\mathrm{i}$ is the response variate for the i*th* element in the respondent population.

- The line through population symbols is to enable us to distinguish *italic* and **bold** upper-case *handwritten* letters and we *say*, for instance, '$\mathbf{Y}_\mathrm{i}$ cross' for the response variate of the i*th* element in the respondent population.

Table HL100.1:

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

✻ In introductory statistics courses, the number of elements in the population (also called the population *size*) is seldom considered explicitly; in these Statistical Highlights, this attribute is denoted $\mathbf{N}$ ('**N** cross').

- Including the population size in survey sampling theory is sometimes called dealing with a *finite* population, but this is *unhelpful* terminology (perhaps carrying over from mathematics where infinity often arises). A population in statistics is a real-world entity and so, by its nature, has a finite number of elements.

✻ When investigating a Question with a *descriptive* aspect [one whose Answer will involve primarily values for population/process attributes (past, present, future)], a useful way to think of (or to 'model') the response variate of a respondent population element is shown in equation (HL100.5) at the right;

$$\mathbf{Y}_\mathrm{i} = \overline{\mathbf{Y}} + \mathbf{R}_\mathrm{i} \quad \text{-----(HL100.5)}$$

this 'model' is useful because it expresses a quantity we can observe ($\mathbf{Y}_\mathrm{i}$) in terms of an attribute of interest ($\overline{\mathbf{Y}}$, the population *average*) and a quantity ($\mathbf{R}_\mathrm{i}$, the *population residual* for element i) whose behaviour is amenable to probability modelling which, ultimately, enables us to quantify imprecision due to sample error and (in some contexts) measurement error.

- The **response model** for a *non*-comparative Plan – the type of Plan usually appropriate to answer a Question with a descriptive aspect – involving equiprobable (simple random) selecting (EPS) of n units from an *un*stratified population is:

$$Y_j = \mu + R_j, \quad j = 1, 2, ..., \mathrm{n}, \quad R_j \sim N(0, \sigma), \quad \text{independent}, \quad \text{EPS}, \quad \text{-----(HL100.6)}$$

where Y_j is a random variable whose distribution represents the possible values of the measured response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over;

R_j is a random variable (called the *residual*) whose distribution represents the possible *differences*, from the structural component of the model, of the measured value of the response variate for the jth unit in the sample of n units selected equiprobably from the respondent population, if the selecting and measuring processes were to be repeated over and over.

The symbol μ in the model (HL100.6) and two related quantities $\hat{\mu}$ and $\tilde{\mu}$ have the following meanings:

μ: a parameter representing the *average* ($\overline{\mathbf{Y}}$) of the measured response variate of the elements of the respondent population.

$\hat{\mu}$: the least squares *estimate* of μ – a *number* whose value is derived from an appropriate set of *data*;

- ○ for the model (HL100.6), $\hat{\mu} = \overline{y}$, reflecting the intuitive idea that, under EPS, the measured sample average estimates the measured respondent population average;

$\tilde{\mu}$: the least squares *estimator* of μ – a *random variable* whose distribution represents the possible values of the estimate $\hat{\mu}$ if the selecting, measuring and estimating processes were to be repeated over and over;

- ○ for the model (HL100.6), $\tilde{\mu} = \overline{Y}$, the random variable representing the sample average under EPS.

✻ The model parameter σ, and two related quantities $\hat{\sigma}$ and $\tilde{\sigma}$, arise in the context of response models like (HL100.6) above:

σ: the (probabilistic) *standard deviation* of the normal model for the distribution of the residual, is a model parameter representing the (data) *standard deviation* ($\mathbf{S}$) of the measured response variate of the respondent population elements; this (data) standard deviation (and, hence, σ) *quantifies* the *variation* of the measured response variate over the elements of the respondent population – as this variation increases, so does $\mathbf{S}$ (and, hence, so does σ).

- ○ two *other* characteristics of variation are its *location* and its *shape* – these are, respectively, 0 and normal for (HL100.6);

$\hat{\sigma}$: the least squares *estimate* of σ – a *number* whose value is derived from an appropriate set of *data*;

$\tilde{\sigma}$: the least squares *estimator* of σ – a *random variable* whose distribution represents the possible values of the estimate $\hat{\sigma}$ if the selecting, measuring and estimating processes were to be repeated over and over.

How σ is *estimated*, reflected by differing expressions for $\hat{\sigma}$, depends on the sampling protocol in the Plan for the investigation and the response model appropriate for this Plan; for the model (HL100.6), $\hat{\sigma}$ is given by equation (HL100.7).

$$\hat{\sigma} = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^2} \quad \text{-----(HL100.7)}$$

(*continued overleaf*)

∗ In this Appendix 2, we meet below *four* quantities from the theory of survey sampling to which the term *standard deviation* is applied; the first two and an associated quantity S are analogues of the three σs, given overleaf near the bottom of page HL100.5, which arise in a response model context. A difference is that, *un*like $\hat{\sigma}$, s is called a standard deviation.

– $\mathbf{S}$: the respondent *population* (data) standard deviation – it is defined in equation (HL100.8) and is a *number* which quantifies the variation over the respondent population of the response $\mathbf{Y}$ about its average $\overline{\mathbf{Y}}$;

$$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2} \qquad \text{-----(HL100.8)}$$

○ like most population attributes except $\mathbf{N}$, usually the value of $\mathbf{S}$ is *un*known;

Table HL100.2: SUMMARY OF STANDARD DEVIATIONS

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

– s: the *sample* (data) standard deviation – it is defined in equation (HL100.9) and is a *number* which quantifies the variation over the sample of the response y about its average $\overline{y}$; the expression for s in equation (HL100.9) is the *same* as that for $\hat{\sigma}$ for the model (HL100.6) given overleaf in equation (HL100.7) at the bottom of page HL100.5 – it is also like equation (HL100.1) on page HL100.1 *except* that y and j are *italic* letters, indicating that *data* values can reasonably be treated as the values of *random variables* under the probabilities which arise from using EPS as the sample selecting method;

$$s = \sqrt{\frac{1}{\mathrm{n}-1}\sum_{j=1}^{\mathrm{n}}(y_j - \overline{y})^2} \qquad \text{-----(HL100.9)}$$

○ under EPS, s is used to estimate $\mathbf{S}$ – that is, to provide a value we can use for $\mathbf{S}$;

○ this Appendix 2 is concerned with only *one* sampling protocol – EPS from an unstratified population – and so there is only *one* expression for the estimate (s) of $\mathbf{S}$;

– S: the *estimator* corresponding to s – it is a *random variable*, of which s is one (realized) *value*;

+ $s.d.(\overline{Y})$: the standard deviation of the sample average – under EPS, it provides a theoretical basis for quantifying uncertainty due to sample error in estimates of respondent population attributes like an average or total;

+ $\widehat{s.d.}(\overline{Y})$: the *estimated* standard deviation of the sample average which, under EPS, is the basis for calculating *values* for the end points of confidence intervals for respondent population attributes like an average or total.

The expressions from survey sampling theory for $s.d.(\overline{Y})$ and $\widehat{s.d.}(\overline{Y})$ differ in that $\mathbf{S}$ is replaced by its estimate s – see equations (HL100.10) and (HL100.11). Also, as shown in Statistical Highlight #77 in Section 5 near the bottom of page HL77.5 and in Appendix 3 on page HL77.9, S^2 is an *un*biased estimator of $\mathbf{S}^2$ but S is a *biased* estimator of $\mathbf{S}$.

$$s.d.(\overline{Y}) = \mathbf{S}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \qquad \text{-----(HL100.10)}$$

$$\widehat{s.d.}(\overline{Y}) = s\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}} \qquad \text{-----(HL100.11)}$$

∗ We capitalize on having *two* words – average and mean – in English to make a useful distinction for a measure of *location*:

– the *average* is a measure of location for a set of *data*;

– the *mean* is a measure of location for (the distribution of) a *random variable*.

However, for the magnitude of *variation* there is only *one* term – standard deviation – for the commonly-used measure, and this can be a source of confusion. Ideally, we would like:

– the *(new word)* as a measure of variation for a set of *data*,

– the *standard deviation* as a measure of variation for a *random variable*,

but the use of 'standard deviation', regardless of context, is too well-established in statistics for this ideal to be attainable. A compromise, to assist beginning students, is to distinguish a *data* standard deviation from a *probabilistic* standard deviation – see Table HL100.3 at the right. Note that *we* use one symbol (*e.g.*, $\mathbf{S}$, s) for a data standard deviation and the abbreviation *s.d.* for a probabilistic standard deviation.

Table HL100.3

Respondent population standard deviation	$\mathbf{S}$	*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	$\widehat{s.d.}(\overline{Y})$	

Statistical Highlight #76 portrays visually the distinction between the sample standard deviation (s; represented by the 16 'hooked' horizontal lines in each diagram) and the standard deviation of the sample average [$s.d.(\overline{Y})$; as estimated from the 16 sample averages and denoted $s_{\overline{y}}$ near the lower right-hand corner of each diagram].

∗ The standard deviation of $\overline{Y}$ is sometimes referred to as the *standard error* of $\overline{Y}$ (*e.g.*, Barnett, pp. 26, 45) but this term is avoided in these Statistical Highlights because it is used by different authors for *both* $s.d.(\overline{Y})$ and $\widehat{s.d.}(\overline{Y})$ (see also Cochran, pages 24, 25-27 and 53), potentially confusing a quantity and its estimate.

REFERENCES: 1. Barnett, V. *Sample Survey Principles and Methods.* Second edition, Edward Arnold, London, 1991; (*First* edition: *Elements of Sampling Theory.* The English Universities Press Ltd., London, 1974).

2. Cochran, W. G. *Sampling Techniques.* John Wiley & Sons, Inc., New York, 3rd Edition, 1977.

∗ The following suggestions may help avoid confusion arising from (careless use of) terminology discussed above.

● when you encounter the word *mean*, be sure you understand whether it refers to:

– an average of data (and whether the data are from a *sample* or a *census*), **OR**

– a random variable [and whether it is an individual random variable or a (linear) combination (*e.g.*, an average, sum or

difference)], **OR**
- a parameter of a response model or probability model.

- when you encounter the term *standard deviation* or *standard error*, be sure you understand whether it refers to:
  - the variation of data (and whether the data are from a *sample* or a *census*), **OR**
  - a random variable [and whether it is an individual random variable or a (linear) combination (*e.g.*, an average, sum or difference)], **OR**
  - a parameter of a response model or probability model.
- when you encounter the word *inaccuracy*, remember that it is a real-world quantity and is defined only in the context of *repetition* of a *process* – like selecting and/or measuring.
  - *Estimating* bias (a model quantity) *differs* from inaccuracy in that it *de*creases with *in*creasing sample size and so may not be of much practical concern in actual sample surveys – see also the second-last paragraph (–) overleaf on page HL100.8.

[There is further discussion of bias in Statistical Highlight #7 and in Appendices 3 and 4 on pages HL77.9 and HL77.10 in Statistical Highlight #77.]

7. Appendix 3 – The Data S.d. Denominator: the s.d. $n-1$ *vs* n saga

Intrepid readers may enjoy the challenge presented by the following.

> The statistical terminology of **degrees of freedom** is often first encountered in introductory statistics courses when the divisor, $n-1$, in the calculation of s^2 is introduced. The sample $\overline{y}$ 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 $(y-\overline{y})$ 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, $\overline{y} = y$, and the deviation is $y-\overline{y} = 0$. However, when we square this and try to divide by $n-1$, 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 $n-1$ degrees of freedom, or essentially $n-1$ 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.

Curiously, a (sympathetic) reviewer singled out discussion like this for praise for its clarity as an explanation of a troublesome statistical issue; even more curious is its suggestion of calculating the sample standard deviation by replacing $\overline{y}$ by (the known value of?) μ (our $\overline{\mathbf{Y}}$) as part of a process of *estimating* (*un*known) $\overline{\mathbf{Y}}$.

When discussing the sample standard deviation, often in the form of the sample variance with a formula like equation (HL100.12), introductory texts are apt to mention, usually briefly, the 'population' standard deviation with a formula like equation (HL100.13). From the perspective of *our* notation, and the distinctions it enables these Statistical Highlights to maintain, expressions like (HL100.12) and (HL100.13) do notable disservice to the discipline of statistics and its teaching in a variety of ways.

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\overline{x})^2}{n-1} \qquad \text{-----(HL100.12)}$$

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N} \qquad \text{-----(HL100.13)} \quad !$$

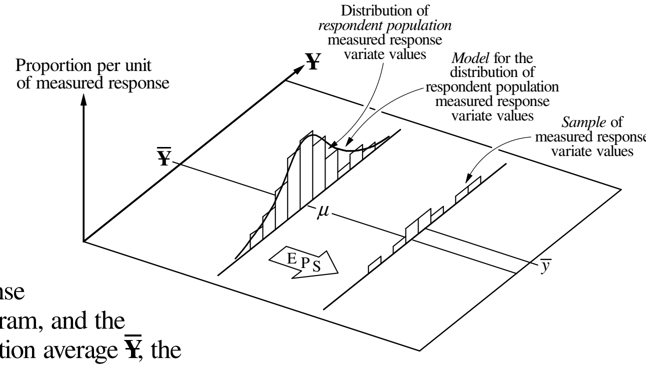
- Emphasis should be on s.d., *not* variance, because:
  - the units of s.d. are those of the variate, but the units of variance are variate units *squared* – for instance, a variate in centimetres has s.d. in cm but variance in cm^2 (a unit of *area*), a variate in \$ has s.d. in \$ but variance in \$^{2};
  - s.d., but *not* variance, can be visualized in suitable pictorial displays;
  - sums of variances in statistical theory can be misinterpreted as implying that variations *add* whereas they actually *add like Pythagoras* (*i.e.*, they add as squares under a square root).

  The traditional emphasis on variance, rather than s.d., in statistical theory may be due, in part, to the inconvenience of typesetting square roots; this was particularly so with manual typesetting but, even now, software packages may produce aesthetically unappealing square roots unless there is human intervention to override default settings. Raising expressions in brackets to the one-half power in place of a square root sign is a visually unevocative alternative to be avoided.

- Letter 'y' would be preferable to 'x' as the variate in expressions like equations (HL100.12) and (HL100.13) for two reasons;
  - in introductory courses, earlier illustrative contexts usually involve *response* (not explanatory) variates;
  - later in such courses when simple linear regression is discussed, 'x' must change to 'y' and the new 'x' looks like the old one but is an *explanatory* variate, which can be confusing for students.

  The response-explanatory variate distinction should be made early in an introductory course and maintained throughout by use of appropriate notation; later discussion of *causation*, and the statistical issues raised by trying to establish it, requires the distinction between focal and *non*-focal explanatory variates (our **X** and **Z**).

(*continued overleaf*)



- Using σ instead of $\mathbf{S}$ and μ instead of $\bar{\mathbf{Y}}$ invites confusing the model and the real world (the respondent population), which is always unwise and occasionally disastrous – for instance, see Statistical Highlight #31.
- Maintaining the population-sample distinction is not helped by use of the *same* variate symbol (lower case italic x) and same summation index (italic i) in equations like (HL100.12) and (HL100.13) – compare Table HL100.1 on page HL100.5.

To emphasize these two distinctions visually, the display at the right shows the respondent population as a histogram of its response variate values, the (normal) model as an idealization of this histogram, and the sample (selected by EPS) as its histogram; the respondent population average $\bar{\mathbf{Y}}$, the model mean μ and the sample average $\bar{y}$ are also shown. [To avoid complicating the diagram unduly, the respondent population (data) s.d. $\mathbf{S}$, the model (probabilistic) s.d. σ, and the sample (data) s.d. s are not labelled.]

- The use of ‘x’ as the generic variate letter, and of *italics* as the default for symbols in equations, may be carry-overs from mathematics; both are *un*suited to statistics – N in equation (HL100.13) overleaf on page HL100.7 looks like a random variable with n in equation (HL100.12) as its value.
  - Unsuitable mathematical *terminology*, which sometimes finds its way into introductory statistics texts, is ‘dependent variable’ and ‘independent variable’ instead of the evocative ‘response variate’ and ‘explanatory variate’.
    - ‘Independent’ in *statistical* contexts is always open to misunderstanding, especially when used as a qualifier of ‘variable’ – recall the Glossary discussion near the bottom of page HL100.4 in Appendix 1.
- The divisor of N (our $\mathbf{N}$) in the ‘population’ s.d. in equation (HL100.13) is not consistent with the $n-1$ in the sample s.d. of equation (HL100.12); when introductory texts try to justify this anomaly, they seem to invoke one of three ‘explanations’.
  - When the sample size is n = 1, equation (HL100.10) in the middle of page HL100.6 can be manipulated [as shown in equation (HL100.14)] to produce an expression for $s.d.(\bar{Y})_{n=1}$ that is the analogue of equation (HL100.13).

$$s.d.(\bar{Y})_{n=1} = \mathbf{S}\sqrt{1-\frac{1}{\mathbf{N}}} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i-\bar{\mathbf{Y}})^2\cdot\frac{\mathbf{N}-1}{\mathbf{N}}} = \sqrt{\frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i-\bar{\mathbf{Y}})^2} \qquad \text{-----(HL100.14)}$$

  Despite appearances, this (probabilistic) s.d. is *not* the population (data) s.d., for reasons given on the lower half of page HL94.9 in Statistical Highlight #94; it also reminds us why *we* distinguish data and probabilistic standard deviations, as in Table HL100.3 at the lower right of page HL100.6.
  - Another ‘explanation’ involves the well-intentioned but dangerous memory aid for setting so-called ‘mean (really, average) square’ divisors: *the number of degrees of freedom lost is the number of parameters estimated* – *e.g.*, in simple linear regression, we estimate two model parameters (slope and intercept) and so use a divisor of n − 2 in the ANOVA table.

  The (mistaken) story goes that, in the sample s.d., we estimate *one* ‘parameter’ μ (really, one population attribute $\bar{\mathbf{Y}}$) by $\bar{y}$ and so use a divisor of n − 1; by contrast, in the ‘population’ s.d., we estimate *no* ‘parameters’ and so use a divisor of $\mathbf{N}$.
    + This memory aid is dangerous because the *real*, but less obvious, idea is the number of *constraints* – recall the third bullet (●) of the Summary on page HL100.3. For example, in the method of least squares, whether we do the minimization by calculus or linear algebra, there are *two* constraints – two partial derivatives set of zero, two vector dot products set to zero for perpendicularity, as shown on the overleaf side under Model 4 in Statistical Highlight #73 – and so the relevant divisor is n − 2. Thus, ‘number of parameters estimated’ is really more a ‘symptom’ than a ‘cause’ of ‘loss of degrees of freedom’ – attentive readers will have noticed a version of the confusion that can arise from the memory aid in the excerpt given at the start of this Appendix 3 overleaf in the middle of page HL100.7.

    In the context of equations (HL100.12) and (HL100.13) [our equations (HL100.9) and (HL100.8)], there is the same *one* constraint on the deviations of (in our notation) the y_js and the $\mathbf{Y}_i$s from their respective averages $\bar{y}$ and $\bar{\mathbf{Y}}$, and so n − 1 and $\mathbf{N}-1$ are the appropriate divisors, making *consistent* the definitions of the population and sample s.d.s, $\mathbf{S}$, s and s.
  - The most confused ‘explanation’ of n − 1 in equation (HL100.12) [even though it is really N (our $\mathbf{N}$) in equation (HL100.13) that needs ‘explaining’] invokes **estimating bias**, an arcane topic for which the enthusiasm of statisticians should be the inverse of how often the matter is mentioned in introductory statistics courses. It is argued that if the divisor in equation (HL100.12) were to be n (as ‘expected’) instead of n − 1, (the random variable) S^2 is *not* an unbiased estimator of σ^2 but a divisor of n − 1 makes it unbiased; this is untrue [σ^2 is being confused with $\mathbf{S}^2$ – see the comments to the left of equation (HL100.11) on page HL100.6] but, regardless, it is largely beside the point because our concern in practice is with $\mathbf{S}$ and s, *not* their squares and, also, values for confidence intervals are *un*affected by whether n − 1, n or even n + 1 is involved.
    + In addition, estimating bias *de*creases with increasing sample size, a dramatic difference in behaviour from **inaccuracy** that should be emphasized – see, for example, Statistical Highlights #15 and #17.

The foregoing discussion shows how a dismal catalogue of misunderstandings and carelessness can be avoided by disciplined use of suitable terminology and notation, to the advantage of statistics and students trying to master its many useful ideas.

∗ It should be of concern to statistics educators when, to ‘learn’ from muddled presentations, we make students abandon essen-

STATISTICS and STATISTICAL METHODS: Standard Deviation and ..... Difference (continued 4)

tial pedagogical precepts like logic, coherence and consistency and substitute uncritical memorization of absurdities.

* In addition, achieving the desirable goal of a statistically literate society is hindered when a subject as vital and ubiquitous as introductory statistics is, through misunderstandings and carelessness, made to appear painful and/or boring and/or useless.

Readers with notable statistical persistence may find the following of interest as an epitome of the foregoing discussion in this Appendix 3; it comes from *The American Statistician* of August, 1983, Vol. 37(#3), page 254. Of particular concern in the context of Appendix 3 are matters raised by the paragraphs which begin: *A few authors refer...* and *Our main reason for...* .

TO BIAS OR NOT TO BIAS

In a recent, brief book review in *The Journal of the American Statistical Association* (see Stephenson 1982) of our text *Elementary Statistics* (Lindgren and Berry, 1981), one-quarter of the space was devoted to commenting on our defining the sample standard deviation with a divisor of n rather than $n-1$. One of our prepublication reviewers had reckoned that a survey of elementary texts would show $n-1$ to be the heavy favorite and, indeed, the convenience sample of such texts on my shelf is 23 to 4 in favor of $n-1$. This may mean, of course, that most people who write new texts have studied from older texts that use the $n-1$. But it is unsettling to hear the report we heard that one professor of statisictics vowed he would resign if forced to teach using n as a divisor. How is it that this point can evoke such passion?

Among those in my survey, some authors state simply and unequivocally that dividing by $n-1$ is better, or best. Some say "statisticians have found" that such is the case, or that there are "subtle theoretical reasons". They are undoubtedly referring to the reason most commonly given by those who give reasons: The version with divisor n for the variance is biased in estimating the variance, and using $n-1$ "corrects" for this bias.

Some authors state that "we want an estimate that is unbiasd", as though this were quite obvious. It is not. I have yet to see a reason why unbiasedness is something we "want". Indeed, it is well known that in the normal case, the divisor $n+1$ is best, if best is meant in the sense of mean squared error. (The notion of variance itself is based on mean squared deviations.)

I'd be interested to know of an instance in which one really wants to estimate a variance. Ordinarily we seek to estimate parameters or quantities that have some direct, intuitive interpretation. The standard deviation of a distribution may well be something to estimate, but of course (as everyone knows) the squre root of an unbiased estimator of variance is biased in estimating the standard deviation.

A few authors refer to the notion of degrees of freedom, pointing out that among the deviations $X_i - \bar{X}$ there are only $n-1$ algebraically independent quantities. True, but this kind of argument would also seem to suggest that when one considers a data set as constituting a population of size N, the *population* variance should be defined with ! divisor $N-1$.

I should say that a few authors do understand the situation, even though ! they may chose to use $n-1$. Snedecor and Cochran and Dixon and Massey say that they use it for practical reasons in more complex problems. Mendenhall and Ott say that sample variance is a misnomer when applied to the unbiased estimate. Breiman says take your pick – and use the unbiased version if it is aesthetically pleasing to you.

Our main reason for choosing to divide by n is that this is most natural and easy to teach, for averaging will already have been introduced as ! a process of summing and dividing by n. And then it doesn't matter whether a given set of data is considered as a sample or a population (or neither!) – the standard deviation is the same numercal measure ! of dispersion in every case. Moreover, formulas for the correlation coefficient are more straightforward with n as divisors. (To be sure, the t statistic would have an $n-1$ in it; but this may be an advantage because it reminds the student of the number of degrees of freedom.)

Having said all this, we may end up switching to $n-1$ just to sell more books. But why should we have to?

Bernard W. Lindgren
School of Statistics
University of Minnesota
Minneapolis, MN 55455

References

BREIMAN, LEO (1973), *Statstics: With a View Toward Applications*, Boston: Houghton Mifflin.

DIXON, W.J. and MASSEY, F.J. (1900), *Introduction to Statistical Analysis* (3rd ed.), New York: McGraw-Hill.

LINDGREN, BERNARD W. and BERRY, DONALD A. (1981), *Elementary Statistics*, New York: Macmillan.

SNEDECOR, G.W. and COCHRAN, W.G. (1967), *Statistical Methods* (6th ed.), Ames: Iowa State University Press.

STEPHENSON, ROBERT W. (1982), Review of *Elementary Statistics* by Bernard W. Lindgren and Donald A. Berry, *Journal of the American Statistical Association*, 77, 215.

8. Appendix 4 – A Broader Perspective on the S.p.d.

An earlier discussion of the s.p.d. is given by Peter M. Heffernan in *The American Statistician* of May 8, 1988, Vol. 42(#2), pages 100-102, in an article entitled *New Measures of Spread and a Simpler Formula for the Normal Distribution*; the author raises the following matters of interest. [Regrettably, this source (or one like it) does not appear to be in the circumscribed statistical lexicon from which most of the plethora of introductory texts seem to draw their content.]

* Near the bottom of page 100, a reference to Gini, C.W. (1912) entitled *Variabilita e Mutabilita* in *Studi Economico-Gluridici Della R. Universita di Cagliari*, 3, part 2, 1-158 appears to be a much earlier use of the s.p.d.
* In the left-hand column of page 101, the curious population s.p.d. discussed in Note 3 overleaf on page HL100.10 is said to be equal to the (probabilistic) quantity δ in equation (HL100.15) [in our notation] at the right – see also equation (HL94.21) near the bottom of page HL94.7 in Statistical Highlight #94.

$$\delta = \sqrt{E[(Y_1 - Y_2)^2]} \qquad \text{-----(HL100.15)}$$

* In the right-hand column of page 101, a justification for the population s.d. 'anomaly' is the idea that a population of size 1 has zero standard deviation if s.d. is defined with a divisor of **N**, whereas the sample s.d. with divisor n − 1 is undefined when

(*continued overleaf*)

n = 1, in keeping with such a sample providing no information about variation in the population from which it was selected. [This distinction may be appealing mathematically but is not useful statistically because it makes the (zero) difference of each population element from itself part of variation; it also perpetuates the population-sample s.d. 'anomaly' – see Note 3 below.]

∗ As introduced in, for example, Figure 5.1 of the STAT 220 Course Materials, the probability density function of the normal distribution is equation (HL100.16) at the right. Using the s.p.d., it takes the simpler form (HL100.17) [with δ as defined in equation (HL100.15) overleaf on page HL100.9].

$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}[\frac{y-\mu}{\sigma}]^2} \quad ; \quad -\infty < y < \infty \qquad \text{-----(HL100.16)}$$

$$g(y) = \frac{1}{\sqrt{\pi}\delta} e^{-[\frac{y-\mu}{\delta}]^2} \quad ; \quad -\infty < y < \infty \qquad \text{-----(HL100.17)}$$

Off-setting this simpler form is the loss of the visual interpretation of σ as indicating the two points of inflection of the p.d.f. Interested readers can consult the Heffernan article for further discussion of the above four (and other) matters.

NOTES: 3. An interesting side-light on the sample s.d. n − 1 *vs* n saga arises in the the following paragraph lower in the left-hand column on page 100 of the article. [Its last sentence appears to refer, in part, to the ideas described in the third point (∗) overleaf at the bottom of page HL100.9 – see also Note 4 below].

> The spread of a sample and the spread of a population are each defined to be the square root of a simple average of squared pairwise differences. The distinctions between the two measures is that, in the case of a sample, we average all $n(n-1)$ ordered pairs of distinct points whereas, in the case of a population, we average over all n^2 ordered pairs of points including pairs consisting of a point and itself (the extra differences are all 0 but cause the divisor to be increased). This can be seen to be sensible. !

If we denote the proposed 'population' s.p.d. by s_{pda} (the 'a' indicates *all* pairwise differences), the relationship to *our* s.p.d. introduced on page HL100.1 is as given in equation (HL100.18) – our s.p.d. would have a denominator of n(n−1) if it were to be based on all distinct *ordered* pairwise differences and the numerators are the same (as the paragraph reprinted above points out).

$$s_{pda} = \sqrt{\frac{n-1}{n}}\, s_{pd} \qquad \text{-----(HL100.18)}$$

Thus, s_{pda} corresponds to s.d. with a divisor of n, just as our s_{pd} corresponds to one with divisor n − 1 (as also pointed out in the article's next paragraph). From a statistical perspective, s_{pda} (with its inclusion of the n zero pairwise 'self'-differences) is a bizarre way of quantifying variation and so s_{pda} further discredits the idea of n (or N) as a divisor for the standard deviation – see also equation (HL94.13) near the bottom of page HL94.6 in Statistical Highlight #94.

- Equation (HL100.18) shows that, for a population of N elements, s_{pda} is the right-most expression in equation (HL100.14) on page HL100.8 and so also is $s.d.(\overline{Y})_{n=1}$ – recall the second point (∗) overleaf on page HL100.9.

If it had been generally accepted early on in statistics that there is (sometimes) more to calculating an average than dividing a sum by its number of components, the oceans of ink (or, more recently, the acres of pixels) devoted (unprofitably) to the n − 1 *vs* n saga (and its appendage, the bogus population-sample s.d. 'anomaly') could have been avoided. Accepting that the process of averaging is not as straight-forward as it appears at first sight simply mirrors other situations we encounter routinely. Sadly, insisting on maintaining the simple view of averaging can set a useful statistical insight about its divisor – unhelpfully named 'degrees of freedom' – on a path to inconsistency and absurdity and, in doing so, tarnish the image of the whole discipline.

4. Further to the foregoing discussion [in particular, the second point (∗) overleaf at the bottom of page HL100.9] is the following paragraph from page 101 of the article; paragraphs which follow it in the article pursue its ideas.

> It seems reasonable to subtract the points from themselves in defining the spread of a population, because we want this spread to equal the spread of the random variable obtained by sampling from the population *with replacement*. Since sampling is with replacement, the same point can be sampled more than once, so the average over all possible pairs that could be obtained by sampling from the population includes pairs consisting of a point and itself; that is, the spread of the random variable reflects the spread of the population with self-differences included. !

One problem here is that mathematical considerations are driving statistics in an unhelpful direction. Sampling (really selecting) *with* replacement is sometimes a convenience (even a necessity) for mathematical tractability in the development of statistical theory but, in the real world, it is selecting with*out* replacement that is relevant, even if there are situations where the difference between selecting with and without replacement has no meaningful effect on Answers.

- A limiting case is n = 1 when there is *no* with/without replacement difference [note equation (HL100.14)].

A second matter is why the paragraph takes sampling as being relevant to how (population) variation is *quantified* (as distinct from *estimated*); it may be that the same (or a related) statistical issue is involved as in the discussion of equation (HL100.14) on page HL100.8 – see the bullet (●) above in Note 3.

The paragraph also seems to encourage ignoring the helpful data/probabilistic standard deviation distinction.

At its heart, the s.d. n − 1 *vs* n saga is about which among competing statistical issues is given primacy:

- how to calculate an average – simple or nuanced?
- how to quantify variation in populations and in samples – the same or a different way? using s.d. or s.p.d.?

Despite much (not always informed) debate, statistics seems to give primacy to nuanced averaging. This Highlight #100 justifies this choice (using the s.p.d.) and describes both how the competing issues can then properly be accommodated and difficulties arising had a different choice been made; it also identifies statistical issues which, in this context, are distractions to be avoided.