University of Waterloo W. H. Cherry

CONFIDENCE INTERVALS: Quantifying Sampling (and Measuring) Imprecision
(for some types of Questions)

When a *measurement* is made, its *accuracy* (even its *meaning*) depends on the existence of a *standard* unit; by 'standard', we mean that all parties concerned agree that this standard represents the *true* value of a unit of the type of measurement. For mass, for example, it is agreed by international treaty that the International Prototype Kilogram, a cylindrical platinum-iridium ingot held at the International Bureau of Weights and Measures near Paris, has a mass that defines the unit *one kilogram*. Ultimately, *all* other measurements of mass depend for their accuracy on a comparison with this standard in Paris; in practice, *direct* comparison with the international standard is extremely rare. Instead, there is a chain of comparisons involving intermediate 'standards' of progressively increasing authenticity; the further a comparison is along the chain, the less frequently it is made. People typically encounter the start of such a chain in the supermarket, where they generally accept that a one kilogram bag of sugar (say) contains this mass (or 'weight') of sugar, to within a certain *tolerance*. (See also Appendix 5 on pages HL2.17 and HL2.18).

1. An Illustration of a Measuring Process – the Mass of NB10

Each country has one or more arms of government or other organizations concerned with standards for the many types of quantities and methods of measurement. The activities of these organizations include the maintenance of honest measures in the marketplace and the preservation and calibration of national standards; the latter are needed because it is not feasible to consult the *inter*national prototype standards on a routine basis. The following data set comes from calibration work in the U.S. on the reproducibility of measurements of mass ('weight'); the data are 100 replicate measurements, made at what was then the U.S. National Bureau of Standards (NBS) in Washington, D.C., of the weight of the U.S. national ten-gram standard, NB10. For convenience, the data are coded as the number of *micrograms* (μg) *below 10 grams*; for example, the first reading ('**Wt:**) of 409 corresponds to a measured weight of 9.999 591 grams. [For ease of reference, the readings are numbered ('#') 1 to 100 in the order in which they were obtained on a roughly weekly basis over the period 1962-63.]

Table HL2.1:

#	Wt.	#	Wt.	#	Wt.	#	Wt.	#	Wt.	#	Wt.	#	Wt.	#	Wt.	#	Wt.	#	Wt.
1	409	11	398	21	408	31	403	41	405	51	404	61	404	71	412	81	408	91	401
2	400	12	403	22	399	32	400	42	409	52	406	62	405	72	406	82	406	92	407
3	406	13	407	23	399	33	410	43	399	53	407	63	392	73	409	83	401	93	412
4	399	14	402	24	402	34	401	44	402	54	405	64	407	74	400	84	412	94	375
5	402	15	401	25	399	35	407	45	407	55	411	65	406	75	408	85	393	95	409
6	406	16	399	26	397	36	423	46	406	56	410	66	404	76	404	86	437	96	406
7	401	17	400	27	407	37	406	47	413	57	410	67	403	77	401	87	418	97	398
8	403	18	401	28	401	38	406	48	409	58	410	68	408	78	404	88	415	98	406
9	401	19	405	29	399	39	402	49	404	59	401	69	404	79	408	89	404	99	403
10	403	20	402	30	401	40	405	50	402	60	402	70	407	80	406	90	401	100	404

For these data:

sum = 40,459,

sum of squares = 16,373,447,

so the average and (data) standard deviation are:

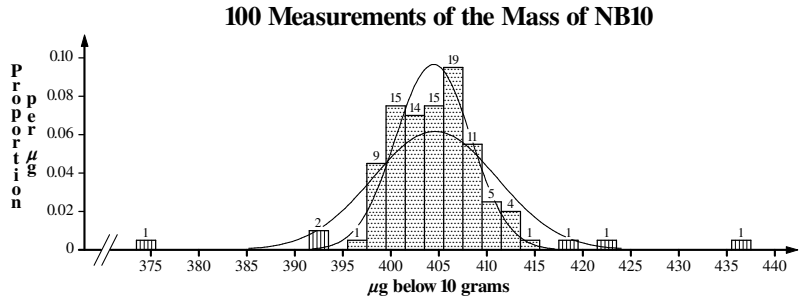
$\bar{y} = 404.59\ \mu g$,

$s = 6.47\ \mu g$.
(6.466 846 μg)

DATA SOURCE: Freedman, D., Pisani, R. & Purves, R.: *Statistics*. W. W. Norton & Company, Inc., 1978, page 91.
The quotation from NBS in Note 1 below is taken from the same source (page 95).

A histogram of the data is given at the right; the bars are 2 μg wide with their edges at half-integer values – for example, the left-most bar covers the interval 373.5 to 375.5 μg below 10 grams. The number at the top of each bar is the number of observations for the bar. Superimposed on the histogram are two normal probability density functions (p.d.f.s); the wider shorter one has mean 404.59 μg and (probabilistic) standard deviation 6.47 μg and the narrower taller one has mean 404.48 μg and (probabilistic) standard deviation 3.88 μg.

100 Measurements of the Mass of NB10



Before we pursue an analysis of the data in Table HL2.1 above, we note from them matters of statistical interest.

NOTES: 1. The data set shows several **outliers** (*i.e.*, observations far from the majority of the data) and is a rare illustration of the *real* performance of a measuring process; it illuminates the following comment by NBS:

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

The invaluable statistical lesson for us is that, if *this* is the performance of a measuring process for which:

- the operator(s) are highly-trained professionals,

(*continued overleaf*)

NOTES: (cont.)

1. ● the measuring instrument is of the highest quality,
   ● the quantity being measured is relatively straight forward,

   we must avoid uncritical acceptance of the stated (or implied) inaccuracy and/or imprecision of the plethora of currently-accessible data whose generation may have met only some (or none) of these three conditions.

   Assessment of the inaccuracy and imprecision of measuring processes should be *integral* to data analysis but, regretably, is seldom so in practice – see also the discussion in the second paragraph on page HL38.1 of Statistical Highlight #38.

2. The diagram overleaf on page HL2.1 illustrates the idea of *modelling* the shape of a distribution arising from *variation* in data – in this instance, we are using:
   ● a normal distribution with mean 404.59 and standard deviation 6.47 as a model for all 100 observations;
   ● a normal distribution with mean 404.48 and standard deviation 3.88 as a model for the central 94 observations, omitting the six most deviant measurements.

   For these models, relevant information calculated from the data set and three subsets of it are given in Table HL2.2 at the right

Table HL2.2

Observations omitted (#)	Sample size	Sum	Sum of squares	Average (μg)	(Data) S.d. (μg)
None	100	40,459	16,373,447	404.59	6.47
375 (94), 437 (86)	98	39,647	16,041,853	404.56	4.78
393 (85), 423 (36)	96	38,831	15,708,475	404.49	4.28
392 (63), 418 (87)	94	38,021	15,380,087	404.48	3.88

   Looking at the histogram overleaf on page HL2.1 with its two superimposed normal p.d.f.s, we see that:
   ○ the $N(404.59, 6.47)$ model does *not* fit the distribution of all 100 observations – most notably, it misrepresents the extent of outliers;
   ○ the $N(404.48, 3.88)$ model is a fair (but not a good) fit to the central 94 observations.

   With data sets like the NB10 observations overleaf in Table HL2.1 on page HL2.1, there is a temptation for investigators not to mention discarding the six outliers and to report *only* the central 94 measurements, because:
   ∗ it is easier to model the shape of their distribution,
   ∗ they conform to the *mis*conception that data from *most* measuring processes can be fitted by a normal model,
   ∗ it is widely believed that the presence of outliers in a data set *necessarily* reflects unfavourably on the investigator,
   ∗ they provide a *less imprecise* estimate of the mass of NB10 at the cost of ignoring the risk of *greater inaccuracy*.

3. In the source cited overleaf on page HL2.1 for the NB10 data and the NBS quotation, a (surprising) interval width of 2½ μg was chosen for the histogram; only three (not six) observations (numbers 94, 86 and 36) were then excluded for the 'central' data set of 97 (not 94) observations, which has average 404.37 μg and (data) standard deviation 4.41 μg.

2. A Confidence Interval for μ Representing the (true) Mass of NB10

A *confidence interval* (abbreviated CI) is a widely-used statistical tool for quantifying the imprecision (*e.g.*, arising from sample and/or measurement error) of an estimate from sample data of (the value of) a population attribute; the simple numerical attribute commonly of interest is the (study or respondent) population average, $\bar{Y}$. The (true) mass of NB10, here denoted μ**M**, can be thought of as the average of a very large (in principle, an infinite) number of measurements made by a measuring process with *no* inaccuracy. The following derivation of a CI for the mass of NB10 from the data in Table HL2.1 uses probabilistic ideas from Figures 5.1, 5.4, 5.5 and 5.6 and statistical ideas from Figures 5.7 and 5.8 of the STAT 231 Course Materials.

We model in words an outcome of the process of measuring the mass of NB10 as:

measured value = true value + haphazard measurement error; -----(HL2.1)

measurement error usually changes in value from measurement to measurement and, for a particular measurement, may be positive or negative (or, conceivably, zero) so the measured value may be above or below (or, conceivably, equal to) the true value. As indicated by the adjective 'haphazard' in equation (HL2.1), we assume initially *no* inaccuracy in the NBS measuring process, meaning that, over a large set of measured values of the mass of NB10, the *average* measurement error is (close to) zero.

Expressing the model in equation (HL2.1) symbolically for the jth measurement of a set of $n = 100$ measurements, using the notation of equation (5.6.7) on page 5.18 in Figure 5.6 of the STAT 231 Course Materials, we have:

$$Y_j = \mu + R_j, \quad j = 1, 2, \ldots, n, \quad E(R_j) = 0, \quad s.d.(R_j) = \sigma, \quad \text{independent}, \quad \text{EPS}, \qquad \text{-----(HL2.2)}$$

where: μ is the model *mean* – it represents the (true) mass of NB10, μ**M**, AND:
σ is the model (probabilistic) *standard deviation* – it quantifies variation in the measuring process, arising from sources like the measuring instrument, the operator(s) and the environmental conditions under which a measurement is made.

Equation (HL2.2) has no bias term (and R_j has zero mean) because of our initial assumption of no inaccuracy in the measuring process; two additional modelling assumptions (whose discussion is pursued in point 6 on pages HL2.7 and HL2.8) are:
∗ the set of measured values can be treated as though it is a sample of size 100 obtained by *equiprobable selecting* (EPS) from the (hypothetical) population of all such values,
∗ the random variables we use to represent the outcomes obtained under repetition of the measuring process are *probabilistically independent* for any two different values of j – see the discussion on page HL38.5 in Statistical Highlight #38.

University of Waterloo W. H. Cherry

CONFIDENCE INTERVALS: Quantifying Sampling and Measuring Imprecision (continued 1)

(for some types of Questions)

Adding the n equations (HL2.2) and dividing each term by n, we have:

$$\frac{1}{n}\sum_{j=1}^{n} Y_j = \frac{1}{n}\sum_{j=1}^{n}\mu + \frac{1}{n}\sum_{j=1}^{n} R_j \qquad \text{OR:} \qquad \overline{Y} = \mu + \overline{R}. \qquad \text{-----(HL2.3)}$$

Using the results in Figure 5.4 on page 5.12 for the average of n probabilistically independent random variables, together with the Central Limit Theorem approximation in equation (5.5.5) on page 5.13 of Figure 5.5, we have:

$$\overline{R} \dot{\sim} N(0, \sigma\sqrt{\tfrac{1}{n}}) \qquad \text{so that:} \qquad \overline{Y} \dot{\sim} N(\mu, \sigma\sqrt{\tfrac{1}{n}}) \qquad \text{and so:} \qquad \frac{\overline{Y}-\mu}{\sigma\sqrt{\frac{1}{n}}} \equiv \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1); \qquad \text{-----(HL2.4)}$$

the denominator of the rightmost expression is written in two forms to emphasize their equivalence.

For a random variable Z with a $N(0,1)$ distribution, we know from STAT 231 page Ap.1 of Table 1 in Appendix B that:

$$\Pr(-1.96 < Z \le 1.96) = 0.95; \qquad \text{-----(HL2.5)}$$

see also the middle diagram of the three at the bottom of the first side (page 5.1) of STAT 231 Figure 5.1.
Using equation (HL2.4) in equation (HL2.5), we have:

$$\Pr\left(-1.96 < \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \le 1.96\right) \simeq 0.95;$$

the probability of 0.95 is now only *approximate* because of the approximate normality in equation (HL2.4).

$$\therefore \quad \Pr(-1.96\sigma/\sqrt{n} < \overline{Y}-\mu \le 1.96\sigma/\sqrt{n}) \simeq 0.95;$$

$$\therefore \quad \Pr(\overline{Y}-1.96\sigma/\sqrt{n} < \mu \le \overline{Y}+1.96\sigma/\sqrt{n}) \simeq 0.95 \qquad \text{(in two steps).} \qquad \text{-----(HL2.6)}$$

The 'probability' expression of equation (HL2.6) differs from a more usual expression like equation (HL2.5) in that:
- its end points are stochastic (not 'fixed') quantities, AND:
- its centre is a 'fixed' (not a stochastic) quantity;

to reflect these differences, we change the wording and the presentation of equation (HL2.6) to:

an approximate 95% confidence interval for μ representing **M** *is* $(\overline{Y}-1.96\sigma/\sqrt{n},\ \overline{Y}+1.96\sigma/\sqrt{n})$. -----(HL2.7)

The foregoing lengthy derivation of the CI expression of equation (HL2.7) is given to show the *reasoning* leading to this widely-used statistical method for estimating from sample data (the value of) a population attribute, in this instance an average. To find a *realized* CI in practice, we work with equation (HL2.7) *as given*, using appropriate values for $\overline{Y}$, σ and n;

an approximate ***realized*** *95% confidence interval for μ representing* **M** *is* $(\overline{y}-1.96\sigma/\sqrt{n},\ \overline{y}+1.96\sigma/\sqrt{n})$. -----(HL2.8)

We will usually omit the adjective 'realized' because, elsewhere, the distinction is *rarely* made between:
- equation (HL2.7), the *general* CI expression involving a random variable, here $\overline{Y}$, AND:
- equation (HL2.8), the expression for a *numerical* CI based on values arising from a particular data set;

rather, equation (HL2.7) may be omitted and equation (HL2.8) presented as the (unqualified) CI for a mean representing an average.

For the NB10 mass measurements in Table HL2.1 on page HL2.1: $\overline{y} = 404.59\,\mu g$, $s = 6.47\,\mu g$ and $n = 100$, so that:

an approximate 95% confidence interval for μ representing **M** *is* (403.32, 405.86) μg below 10 grams, -----(HL2.9)

using, at this point in the discussion, the value of s for that of σ.

3. Understanding Confidence Intervals for a Model Mean Representing a Population Average

To use confidence intervals properly, there is much to understand about them; discussion follows under six headings ('points').

1. **Interpretation of a CI.** A CI is an interval of *plausible* values, in light of the data, for a population attribute (like $\overline{\mathbf{Y}}$) represented by a model parameter (like μ).

2. **The value used for the confidence level.** The CIs given in equations (HL2.7), (HL2.8) and (HL2.9) have a confidence level of 95%; this is the most *common* value in practice, although 90% and 99% are also encountered. To *change* the confidence level, we change the numerical coefficient, as illustrated in Table HL2.3 below – the number in the *first* column of this table, and its designation 'α', is only for completeness after (much) later discussion of another statistical estimating method called a **test of statistical significance.**

Table HL2.3

α	$1-\alpha$	Level	Coeff.	Equation (HL.7)	Equation (HL2.8)	Equation (HL2.9)	CI width
0.10	0.90	90%	1.6449	$(\overline{Y}-1.6449\sigma/\sqrt{n},\ \overline{Y}+1.6449\sigma/\sqrt{n})$	$(\overline{y}-1.6449\sigma/\sqrt{n},\ \overline{y}+1.6449\sigma/\sqrt{n})$	(403.53, 405.65)	2.12
0.05	0.95	95%	1.9600	$(\overline{Y}-1.9600\sigma/\sqrt{n},\ \overline{Y}+1.9600\sigma/\sqrt{n})$	$(\overline{y}-1.9600\sigma/\sqrt{n},\ \overline{y}+1.9600\sigma/\sqrt{n})$	(403.32, 405.86)	2.54
0.01	0.99	99%	2.5758	$(\overline{Y}-2.5758\sigma/\sqrt{n},\ \overline{Y}+2.5758\sigma/\sqrt{n})$	$(\overline{y}-2.5758\sigma/\sqrt{n},\ \overline{y}+2.5758\sigma/\sqrt{n})$	(402.92, 406.26)	3.34

- The numerical coefficient (now to four decimal places) in the fourth column of the table is the ordinate of the negative and positive boundaries of the *central* 90%, 95% and 99% of the area under the $N(0,1)$ p.d.f. – recall the three diagrams at the bottom of the first side (page 5.1) of Figure 5.1 in the STAT 231 Course Materials.
  - Using these values in equation (HL2.8) gives the respective 90%, 95% and 99% CIs in the second-last column of the table.

(*continued overleaf*)

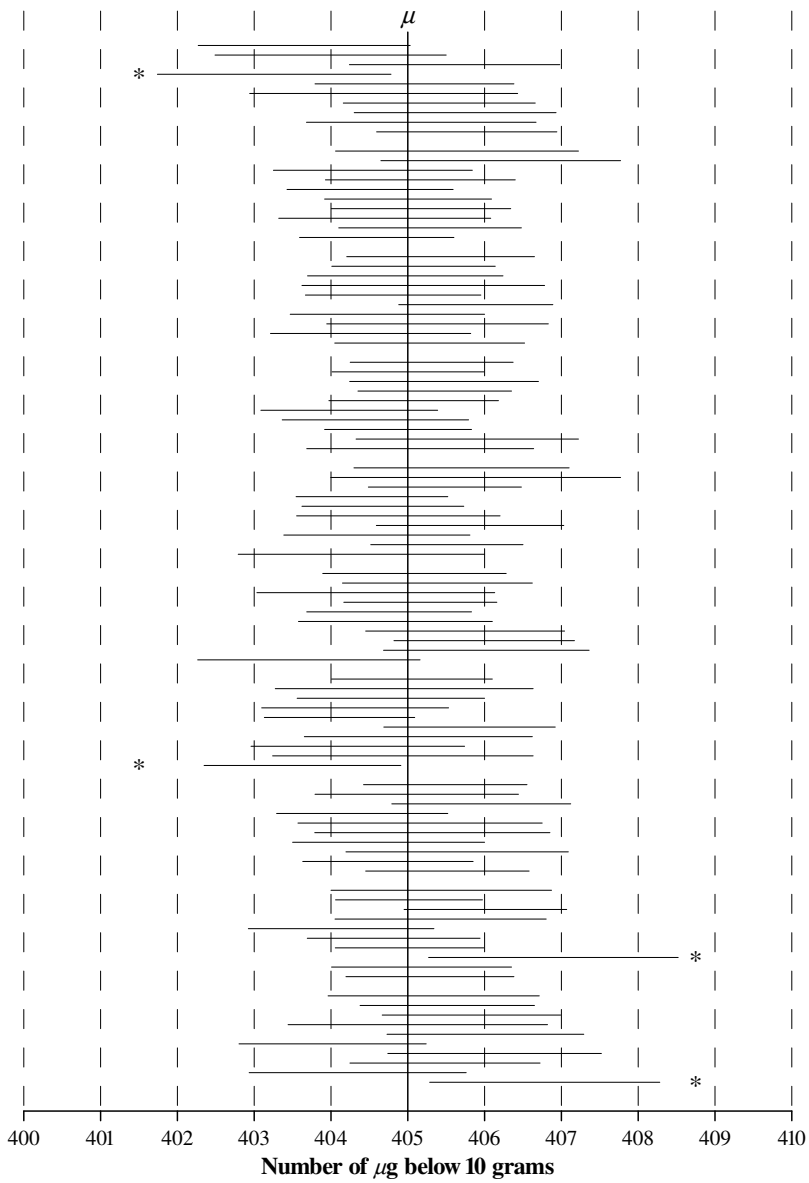
- In principle, *any* value for a confidence level can be chosen; for instance, we could use as the coefficient any (except the first) of the 64 values in the lower table for the $N(0, 1)$ [*i.e.*, the $N(0, 1)$] distribution on page Ap.1 in Appendix B of the STAT 231 Course Materials; in *practice*, levels other than (the round numbers) 95%, 90% and 99% are rarely (if ever) used.
- The lower table on page Ap.1 gives selected *percentiles* for the right-hand half of the $N(0, 1)$ p.d.f. – that is, an $F(z)$ value is the area under the p.d.f. from $-\infty$ to z. However, a CI involves the *central* area from $-z$ to z, so that
  - the coefficient of 1.6449 for a 90% CI appears for an $F(z)$ value of .95,
  - the coefficient of 1.9600 for a 95% CI appears for an $F(z)$ value of .975,
  - the coefficient of 2.5758 for a 99% CI appears for an $F(z)$ value of .995.

  It is a common mistake to forget this minor complication when using a so-called *one-tailed* table (like Table 1 in Appendix B) to look up coefficient values for calculating a CI. It is possible to avoid this complication by reformatting to a *two-tailed* table, but this only trades a minor complication in one situation for one in another.
- The last column of Table HL2.3 overleaf near the bottom of page HL2..3 is the *width* of each CI – its upper end minus its lower end – and it is discussed at the end of the first bullet of point 4, with reference to Table HL2.4, on the facing page HL2.5.

3. **Interpretation of the confidence level.** We use here the idea of repeating over and over, implemented as a computer simulation of a (mere) 100 repetitions of calculating a CI and shown graphically below at the right.

- The simulation requires a value for μ, here taken as 405 μg below 10 grams; also, it uses $N(405, 5)$ and $N(405, 20)$ models for the 'central' and 'outlying' measurements of the mass of NB10, with respective weights of 96% and 4%.
- Each CI in the diagram is shown as a horizontal line running from the value of its lower end to that of its upper end; for example, the first interval is (402.27, 405.03).
  - For visual convenience, there is extra horizontal white space in the diagram after each ten CIs.
- The centre and the width of these hundred CIs change as different simulated samples of 100 observations yield different (realized) values for the random variables $\overline{Y}$ and S representing the sample average and sample (data) standard deviation under repetition of the (sampling and) measuring processes.
- The level of a CI is its *coverage*, meaning that, for a 95% CI, about 95 of every hundred intervals that would be obtained under repetition *would* contain the value of the model parameter representing the population attribute and about 5% would *not*. In the simulation, we see that 96 CIs *do* cross the heavy vertical line at the value of μ and *four* CIs do not – they are marked with an asterisk (*).
  - If the simulation with 100 repetitions were to be done for a 90% or a 99% CI instead of a 95% CI, we would expect to see (respectively) about ten CIs and about one CI that did not cover the value of μ.

Computer Simulation of One Hundred Approximate 95% Confidence Intervals for the True Mass of the U.S. 10-gram standard, NB10



  - When a CI is calculated from data obtained in an investigation, coverage is interpreted in terms of the CIs that would

CONFIDENCE INTERVALS: Quantifying Sampling and Measuring Imprecision (continued 2)
(for some types of Questions)

(hypothetically) be obtained if the investigation were to be repeated over and over with the same Plan.

– The real-world analogue of the simulation is the (daunting) task of repeating one hundred times the process of measuring the mass of NB10 one hundred times – that is, making 10,000 mass measurements.

The inherent uncertainty (due to sample error and measurement error, for example) is reflected in not knowing, for the data from one *actual* investigation, whether the 95% (say) CI calculated from these data is:

○ one of the 95 or so percent of intervals that *do* cover μ, OR:

○ one of the 5 or so percent of intervals that do *not* cover μ.

We are reminded that statistical methods cannot *remove* uncertainty arising from one or more of our six categories of error (although their proper use can *reduce* it) but rather statistical methods *quantify* the uncertainty remaining under the Plan.

- Coverage is a formalization of the change in perspective on page HL2.3 in writing equation (HL2.6) as equation (HL2.7).
- If we want to discuss a CI more generally with*out* specifying the value of its confidence level, we can refer to a $100(1-\alpha)\%$ CI – for example, if $\alpha = 0.05$, this is a 95% CI; this is one way 'α' in the first column of Table HL2.3 on page HL2.3 is useful.
  – In a similar vein, a *symbol* for the CI numerical coefficient taken from the $N(0,1)$ table is $Z^*_{1-\alpha}$ (*e.g.*, $Z^*_{0.95} = 1.9600$)

4. **Factors affecting the width of a CI for the model mean μ representing the population average $\bar{Y}$.** We need to understand factors that affect the width of a CI because:

∗ a *narrower* CI determines the interval of plausible values for the population attribute with *smaller* imprecision;

∗ the Plan for an investigation may want to manage these factors in order to obtain a CI of *specified* width.

For this discussion, we write the CI of equation (HL2.8) as: $\bar{y} \pm Z^*_{1-\alpha}\sigma/\sqrt{n}$; -----(HL2.10)

we see from equation (HL2.10) that *three* factors affect the width of a CI for a mean representing a population average.

- The *confidence level*, which is determined by the coefficient $Z^*_{1-\alpha}$ – the *higher* the confidence level, the *wider* the CI and, conversely, the *lower* the confidence level, the *narrower* the CI.
  – The freedom of an investigator to manipulate the confidence level to alter the width of (usually, to make narrower) a CI is *constrained* because, once data have been collected from the units of the sample, the data contain a *fixed* amount of information relevant to estimating a model parameter representing a population attribute. If the CI calculated from the data is too wide to meet the requirements of the investigation, it can be made narrower only by using a *lower* confidence level. While the *width* of the CI may then be acceptable, its *confidence level* may not be.
    ○ This is the same *idea* as the *Heisenberg Uncertainty Principle* in physics – *more* precise knowledge of one variable (like position) can only be obtained by *less* precise knowledge of another variable (like momentum). The 'fixed' amount of information that must be partitioned between the measurement of two such canonical variables involves Planck's constant, which is so small that effects of the Uncertainty Principle are usually only important at the atomic level.
  – A numerical illustration of the effect of confidence level on CI width is given in the last column of Table HL2.3 on page HL2.3; these widths, supplemented by information for four *higher* confidence levels, are given in Table HL2.4 below. We see, for instance, that:
    ○ the 95% CI is about twenty percent wider than the 90% CI;
    ○ the 99% CI is over thirty percent wider than the 95% CI;
    ○ the 99.9999% CI is nearly three times as wide as the 90% CI.

Table HL2.4: Confidence Levels and CI Widths for the NB10 Data

α	0.10	0.05	0.01	0.001	0.0001	0.000 01	0.000 001
Level (%)	**90**	**95**	**99**	**99.9**	**99.99**	**99.999**	**99.9999**
Coefficient	1.6449	1.9600	2.5758	3.2905	3.8906	4.4172	4.8916
Width (μg)	**2.12**	**2.54**	**3.34**	**4.26**	**5.04**	**5.72**	**6.33**
Ratio	---	1.20	1.31	1.28	1.18	1.14	1.11
Ratio to 90%	1.00	1.20	1.58	2.01	2.37	2.70	2.99

    ○ The six ratios of *successive* CI widths in the second-last line of the table (and, in the last line, the ratio of each CI width to that of the 90% CI) show there is no *simple* quantitative relationship between level and width for a CI.
- The *standard deviation* σ – the *larger* σ, the *wider* the CI and, conversely, the *smaller* σ, the *narrower* the CI.
  – This result is the basis of the intuitive idea that uncertainty due to sample error is likely to *decrease* with *decreasing* variation in the response variate over the elements of the (study or respondent) population.
    ○ In the extreme (and extremely rare) case of a population with the *same* value of the response variate for *all* its elements, there is *no* sampling uncertainty (*zero* sample error) in estimating the average from *any* sample of size $n = 1$.
  – In the context of our introduction to confidence intervals – estimating the mass of NB10 from repeated measurement data – the model standard deviation σ of equation (HL2.2) on the lower half of page HL2.2 quantifies variation in the measuring process; in *this* situation, the magnitude of σ can, to some degree, be managed by the investigator(s) by means of that part of the Plan which deals with the measuring process.

    More generally, σ represents the (data) standard deviation $\mathbf{S}$ of the (study or respondent) population and so its value can*not* be *altered* by the investigator(s).

    Thus, σ affects the width of a CI for μ representing the population average $\bar{Y}$ but the value of $\mathbf{S}$ usually can*not* be altered.
  – Although an investigator usually cannot *alter* the value of $\mathbf{S}$, its *effect* can be reduced, and a narrower CI obtained, by appropriate **stratifying** of the population from which the sample is selected – see Appendix 5 on page HL74.10 of

(*continued overleaf*)

Statistical Highlight #74. The 'costs' of such management of variation are:

- ○ the need to use a more complicated sampling protocol, AND:
- ○ knowledge for all the population elements of the values of a suitable [response or explanatory] variate(s) on which to stratify.

A possible (undesirable) influence of the investigator(s) on the value *used* for **S** is pursued in Note 5 on page HL2.9.

- ● The *sample size* n – the *larger* n, the *narrower* the CI and, conversely, the *smaller* n, the *wider* the CI.
  - – This result is the basis of the intuitive idea that uncertainty due to sample error is likely to *decrease* with *increasing* sample size – informally:
    - ○ a *large* sample is 'better' than a *small* sample because of the likely *smaller* magnitude of sample error, OR:
    - ○ the estimate from a *large* sample is likely to be *closer* to the actual value of the population attribute than the estimate from a *small* sample, PROVIDED THAT: the samples are otherwise equivalent.

    We need to recognize that this proviso is met only in *exceptional* investigations – experience shows that the larger the sample size, the more difficult it becomes to manage inaccuracy and imprecision of estimates derived from sample data.
    - ○ An extreme case is a population *census*, like that carried out quinquenially in Canada; the 2006 census cost about $500 million, reflecting the resources needed to try to manage inaccuracy and imprecision when generating a large data set.
  - – Equation (HL2.10) overleaf near the middle of page HL2.5 shows that the width of a CI for a mean representing a population average depends on the *square root* of n so that, to halve (say) the width of such a CI, the sample size must be increased by a factor of *four* – see also Appendix 3 on pages HL74.6 to HL74.8 of Statistical Highlight #74.

A summary of the three factors which affect the width of a CI for μ representing the population average $\overline{\mathbf{Y}}$ is:

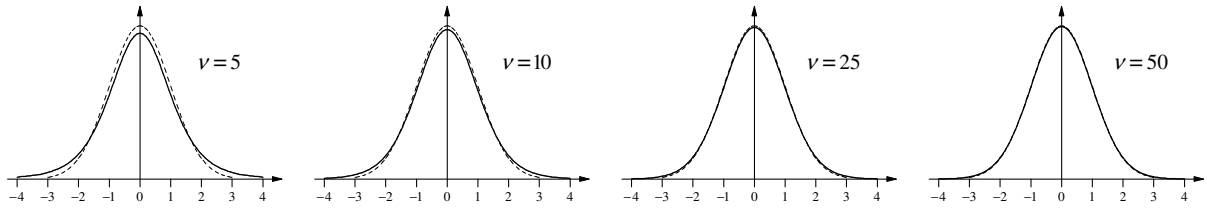
- ∗ once the data have been collected, there is a trade-off between confidence level and interval width;
- ∗ the population standard deviation **S** cannot usually be altered but its effect can be *managed* by a stratified sampling protocol;
- ∗ CI width decreases with increasing sample size but only as the square root of n.

5. **CI numerical coefficients from the t_ν instead of the $N(0,1)$ distribution.** One matter in our introduction to confidence intervals needs to be *changed* to give the usual result for a CI for a mean representing a population average; this matter is the use of the sample standard deviation (s) for the value of σ. In practice, the change involves merely looking up the CI numerical coefficient in a *different* table from that of the $N(0, 1)$ distribution, although the theory underlying this change, which is given in Appendix 3 on pages HL2.14 to HL2.16, is more extensive; this is reminiscent of the lengthy *derivation* on pages HL2.2 and HL2.3 of equation (HL2.7) but its straight-forward *implementation* in equation (HL2.8).

   Under this change, equation (HL2.10) becomes: $\bar{y} \pm t^*_{\nu,1-\alpha}\, s/\sqrt{n}$. -----(HL2.11)

   - ● The t distribution, from which the numerical coefficent $t^*_{\nu,1-\alpha}$ in equation (HL2.11) is obtained, is tabulated in Table 3 on pages Ap. 3 and Ap. 4 in Appendix B of the STAT 231 Course Materials.
     - – The t distribution has one parameter (here denoted ν, the Greek letter *nu*) called its *degrees of freedom* for historical reasons.
       - ○ In a CI for a mean representing a population average, $\nu = n-1$; that is, the degrees of freedom are *one less than the sample size* – see equation (HL2.12) at the right. $\bar{y} \pm t^*_{n-1,1-\alpha}\, s/\sqrt{n}$. -----(HL2.12)
       - ○ The t_ν distribution becomes more like the $N(0, 1)$ distribution as ν increases; the entries in the last line on the overleaf side (page Ap. 4) of Table 3 for an *infinite* number of degrees of freedom are those for the $N(0, 1)$ distribution.
       - ○ The six significant figures for entries in Table 3 for the t_ν distribution abscissas are given for two reasons:
         - + to improve the accuracy of *interpolating* the table for degrees of freedom (table rows) or probabilities (columns) which are not given – interpolating is more accurate for degrees of freedom than for probabilities;
         - + to avoid the practice of carrying *many* figures in CI calculations but, at a late stage, using a numerical coefficient from the t table with only three or four significant figures, although this (careless) practice seldom affects the CI in such a way as to have a practically important effect on an Answer.

         The gap after three significant figures in each entry in Table 3 is to make the entries easier to read.
     - – Like the $N(0, 1)$ distribution, the t_ν distribution has a bell-shaped p.d.f. that is symmetrical about zero and with domain $(-\infty, \infty)$; the two distributions *differ* in that the t_ν distribution has (slightly) less central area and more tail area (sometimes referred to as having *heavier tails*). A visual comparison of the p.d.f.s for the (dashed) $N(0, 1)$ and (solid) t_ν distributions for $\nu = 5$, 10, 25 and 50 degrees of freedom are shown below; at the scale of these diagrams, the two distributions differ almost imperceptibly when ν reaches 50.



     - – The notation $Y \sim t_\nu$ means that the randon variable Y has a t distribution with ν degrees of freedom; in Table 3:
       - ○ column headings are *left-tail areas* [as they are for the $N(0, 1)$ distribution in Table 1 'on page Ap. 1 in Appendix B];

University of Waterloo W. H. Cherry

CONFIDENCE INTERVALS: Quantifying Sampling and Measuring Imprecision (continued 3)

(for some types of Questions)

- the *degrees of freedom* are shown in the left-hand column (and are repeated at intervals at the right of the table to assist with reading across a particular line);
- the two *italicized* columns are those needed (most commonly) to find 95% and 99% confidence intervals;
- an illustration of table look-up is that if the random variable $Y \sim t_{15}$, then: $\Pr(Y \le 2.13\,145) = 0.975$ so that: $\Pr(-2.13\,145 < Y \le 2.13\,145) = 0.95$ and: $\Pr(Y \le -2.13\,145) = \Pr(Y \ge 2.13\,145) = 0.025$.

- We *use* equation (HL2.12) for calculating a CI but we recognize that, for sample sizes that are generally large enough (*e.g.*, $n \ge 50$) to be useful, the similarity of the $N(0, 1)$ and t_{n-1} distributions means that the difference in width of the two CIs is not enough to have a practically important effect on the Answer in most investigations. This is illustrated (when $n = 100$) in Table HL2.5 at the right, for six confidence levels, of $N(0, 1)$ and t_{99} CIs for μ representing $\bar{\mathbf{M}}$ based on the NB10 data in Table HL2.1 on page HL2.1; although the t_{99} CIs *are* wider, the ratios of the two widths in the last column of the table differ from 1 by only a few percent. [These ratios are calculated from the numerical coefficients in Table HL2.5, *not* from the tabulated widths which come from the CIs rounded to 2 decimal places.]

Table HL2..5: Comparison of $N(0, 1)$ and t_ν CIs for the NB10 Data

		..... $N(0,1)$			..... t_{99}			
$1-\alpha$	Level	Coeff.	CI	Width	Coeff.	CI	Width	Width ratio
0.90	90%	1.6449	(403.53, 405.65)	2.12	1.66039	(403.52, 405.66)	2.14	1.009
0.95	95	1.9600	(403.32, 405.86)	2.54	1.98422	(403.31, 405.87)	2.56	1.012
0.99	99	2.5758	(402.92, 406.26)	3.34	2.62641	(402.89, 406.29)	3.40	1.020
0.999	99.9	3.2905	(402.46, 406.72)	4.26	3.39153	(402.40, 406.78)	4.38	1.031
0.999 9	99.99	3.8906	(402.07, 407.11)	5.04	4.05504	(401.97, 407.21)	5.24	1.042
0.999 99	99.999	4.4172	(401.73, 407.45)	5.72	4.65675	(401.58, 407.60)	6.02	1.054

6. **Modelling assumptions underlying the CI derivation.** Now we have equation (HL2.12) on the facing page HL2.6 for finding a CI for μ representing a population average, equations (HL2.7) and (HL2.8) on page HL2.3 are mainly of theoretical interest to us; *their* assumptions for reasonable approximate normality are those for the Central Limit Theorem given in Section 1 of Figure 5.5 on page 5.13 of the STAT 231 Course Materials.

- The response model which is the basis of equation (HL2.12) for the CI is the model of equation (HL2.2) modified to:

$$Y_j = \mu + R_j, \quad j = 1, 2, ..., n, \quad R_j \sim N(0, \sigma), \quad \text{independent}, \quad \text{EPS}; \qquad \text{-----(HL2.13)}$$

thus, for a t_{n-1} distribution CI for μ representing the population average $\bar{\mathbf{Y}}$, there are three modelling assumptions.

– **Assumption 1:** The response variate $\mathbf{Y}_i$, $i = 1, 2, ..., \mathbf{N}$, in the (study or respondent) population has a *normal* distribution – see the model shown superimposed on the histogram of the population response variate values in the diagram at the bottom right of the first side (page HL74.1) of Statistical Highlight #74.

Under repetition over and over of the EPS of Assumption 2 below, the model response variate Y_j then has a normal distribution; we denote its mean by μ and its (probabilistic) standard deviation by σ. These two model parameters represent the respective (study or respondent) population attributes $\bar{\mathbf{Y}}$ and $\mathbf{S}$.

- The (assumed) *exact* normality underlying equation (HL2.12) [and in the model (HL2.13)] is to be contrasted with the *approximate* normality resulting from the Central Limit Theorem approximation underlying equation (HL2.7).

Under the probabilistic independence of Assumption 3 discussed overleaf on page HL2.8, the normality of Y_j yields normality of $\bar{Y}$, the random variable representing the sample average.

To assess the normality assumption for the population $\mathbf{Y}_i$, $i = 1, 2, ..., \mathbf{N}$, we could use a **normal quantile** plot of the sample response variate values y_j, $j = 1, 2, ..., n$; in practice, to allow for when there is a more *general* structural component than μ in the model, we use a normal quantile plot of the **estimated residuals** $\hat{r}_j = y_j - \hat{\mu}$ from the response model, where $\hat{\mu} = \bar{y}$ is the *estimate* of the model parameter μ – see also Appendix 2 on page HL2.13.

- $\mathbf{Y}_i$ and, hence, Y_j are not exactly normal in practice, of course, so the CI from equation (HL2.12) has a confidence level that will *differ* from its stated (or *nominal*) level (*e.g.*, 95%). How the size of this difference is *related* to the degree of departure from normality of the distribution of Y_j is called the **robustness** of the CI estimating process.
  - + Such a process is said to be robust if its outcome – here, the difference between the actual and stated confidence levels of the CI – is relatively *insensitive* to departures from the modelling assumption – here, the normality of Y_j.
  - + The process becomes *less* robust as the outcome becomes *more* sensitive to departures from the assumption.

  Under a departure from normality that leaves the distribution of Y_j 'mound-shaped', the CI estimating process *is* robust; that is, for a CI calculated from equation (HL2.12), the difference between the actual and stated confidence levels should have no practically important effect on an Answer if the distribution of the Y_j is 'mound-shaped'.

  However, under a departure from normality characterized by **outliers** or strong **skewness**, *lack* of robustness is likely – the difference between the actual and stated confidence levels of the CI may have a practically important effect on an Answer – the actual level will be (appreciably) *lower* than the stated level.

  Quantitative treatment of lack of robustness – a source of *model* error – is beyond the scope of this Highlight #2.

Discussion of the next assumption differs depending on whether the context is sampling or measuring.

– **Assumption 2 – sampling:** The sample of n units yielding the response variate values y_j, $j = 1, 2, ..., n$, which are taken as values of the random variable Y_j in the model (HL2.13), is selected by *EPS* (or, more generally, by a *probability* selecting process) from the (study or respondent) population.

(*continued overleaf*)

Statistical Highlight #74 (on pages HL74.2 and HL74.3) shows that Assumption 2, in conjunction with Assumption 1, establishes the normality of the distribution of $\overline{Y}$; in addition, the mean of this distribution is $\overline{\mathbf{Y}}$ [equation (HL74.8)] and its standard deviation is as shown at the right [equation (HL74.11)]. The second term under the square root reflects *lack* of probabilistic independence arising from selecting with*out* replacement; this term is absent for selecting *with* replacement, a process almost *never* used in practice.

$$\mathbf{S}\sqrt{\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}}}$$

- In the *model*, the population attributes $\overline{\mathbf{Y}}$ and $\mathbf{S}$ are replaced by the respective parameters μ and σ which represent them; also, the $1/\mathbf{N}$ under the square root is seldom included.
- The degree to which Assumption 2 is met in an investigation can be assessed by examining the sampling protocol, with the caveats that the investigator(s) provide an adequate description of the protocol and that they adhered to it.

– **Assumption 2 – measuring:** The sample of n measurements yielding the response variate values y_j, j = 1, 2, ..., n, which are treated as values of the random variable Y_j in the model (HL2.13), is selected by *EPS* from the (hypothetical) population of all such measured values.

- The degree to which Assumption 2 is met when investigating a measuring process is *difficult* to assess because:
  + the hypothetical nature of the population makes it unclear which possible measured values it includes,
  + there is no *explicit* sampling protocol which determines which values from this 'population' form the sample.

  It is **wrong** to think that the *haphazard* nature of the process that typically leads to the sample of measured values makes this process *equivalent* to EPS from the population of values that *might* have been obtained.
  - Haphazard selecting is *un*controlled whereas EPS is (tightly) *controlled*; thinking these are equivalent processes is reminiscent of another **wrong** idea about probabilities, that *lack of knowledge* of a probability distribution can be modelled by a *uniform* distribution (see discussion of the Ali Baba 'paradox' in Statistical Highlight #46.

  Investigating a measuring process by repeated measuring – like the NB10 investigation – *may* yield a useful Answer from a CI calculated from equation (HL2.12) but its basis in statistical theory lacks justification for Assumption 2.

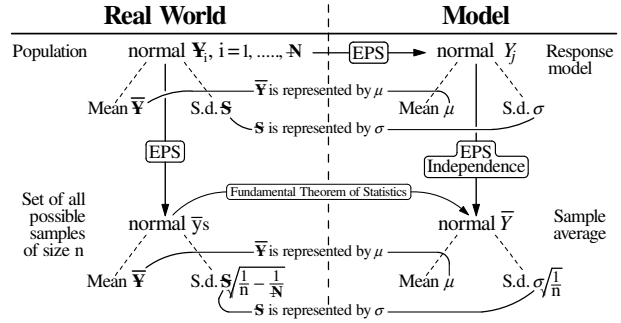
– **Assumption 3:** The residuals R_j are *probabilistically independent* random variables.

- This assumption is met if the sample is selected equiprobably *with* replacement; under the EPS (with*out* replacement) used in practice, Assumption 3 becomes closer to being met as the sample size becomes a progressively smaller proportion of the population size. Thus, violation of Assumption 3 under EPS should have no practically important effect on an Answer from a CI calculated from equation (HL2.12) on page HL2.6 provided $\mathrm{n} \ll \mathbf{N}$.
- A graphical check, if needed, of Assumption 3 is to make a scatter diagram of the *estimated residuals* $\hat{r}_j = y_j - \hat{\mu}$ in the *order* in which the data were collected;
  + a *pattern* in the diagram (*e.g.*, one or more groups in which the points are close to each other) suggests Assumption 3 may *not* be met;
  + *lack* of a pattern (*e.g.*, the diagram shows a haphazard scatter of points) suggests Assumption 3 *is* met.
- In a measuring process investigation (like the NB10 investigation), meeting Assumption 3 can be helped by a Plan which tries to remove from a measured value any *influence* of a previously measured value – see the discussion on page HL38.5 in Statistical Highlight #38.
  + An illustration of such influence, when making measurements in duplicate, is if the operator records the *same* value for the second measurement as for the first, *regardless* of the second value obtained.

A summary of the discussion of the three assumptions is given in the schema at the right; the ideas start with the population response, $\mathbf{Y}_i$, i = 1, ..., $\mathbf{N}$, at the upper left and flow down and to the right across the schema.

Summary of Modelling Assumptions

● Normality ● EPS ● Probabilistic independence



$\mathbf{Y}_i$, i = 1, ..., 10, has a uniform (*not* a normal) distribution in the illustrative diagrams on pages HL74.6 and HL74.7 in Statistical Highlight #74.

NOTES: 4. There are differences of emphasis among introductory statistics courses in presenting a CI for a model mean (like μ) representing a population average (like $\overline{\mathbf{Y}}$).

- In the STAT 231 Course Materials with a more detailed Table 3 in Appendix B of the *t* distribution up to 5,000 degrees of freedom, equation (HL2.12) on page HL2.6 is the *only* CI expression to be used for μ representing $\overline{\mathbf{Y}}$.
- Elsewhere, a distinction is sometimes made between two cases:
  – $\mathbf{S}$ 'known' and the use of equation (HL2.8) on page HL2.3 involving the *normal* distribution, AND:
  – $\mathbf{S}$ '*un*known' but estimated by s and the use of equation (HL2.12) on page HL2.6 and the $t_{\mathrm{n}-1}$ distribution.

  These Materials involve *only* the latter and Example HL2.1 on page HL2.9 is included *solely* to illustrate the former. Some presentations of the distinction confuse a model parameter with a population attribute and refer to 'σ known'.
- Another unnecessary source of confusion is to change from the $t_{\mathrm{n}-1}$ distribution [equation (HL2.12)] to the $N(0, 1)$

CONFIDENCE INTERVALS: Quantifying Sampling and Measuring Imprecision (continued 4)

(for some types of Questions)

NOTES: (cont.)

4. ● distribution [equation (HL2.8)] for calculating a (realized) CI when the sample size exceeds some number like 30.
   - It is true, however, that, for reasonable sample sizes, it seldom has a practically important effect on an Answer about $\overline{\mathbf{Y}}$ whether the CI numerical coefficent is taken from the t_{n-1} or the $N(0,1)$ distribution.
   - The useful **rule of thumb** is that a 95% CI for a model mean representing a population average is *roughly* the sample average plus and minus around twice the sample standard deviation divided by the square root of the sample size.
     - Elsewhere, $\sigma/\sqrt{n}$ may be called the *standard error* and $s/\sqrt{n}$ the *estimated* standard error, although the adjective 'estimated' (which distinguishes the model quantity from its estimate) is, unfortunately, often omitted.

5. The histogram on the first side (page HL2.1) of this Highlight #2 shows there are several possible *outliers* among the 100 measurements of the mass of NB10; the most obvious are the values of 375 (#94) and 437 (#86) μg below 10 grams, but up to four other measurements *may* also be outliers. As summarized in Table HL2.2 on page HL2.2, if these six deviant observations are progressively omitted in pairs from the data set, the average is relatively *un-* affected but the (data) standard deviation steadily decreases – for the 94 observations, it is 3.88 μg, about 60% of its value of 6.47 μg for all 100 measurements.

   Outliers in data sets are common and, as described in Note 2 on the upper half of page HL2.2, dealing with them poses a dilemma for investigators; for example, investigators may:

   ∗ use the *entire* data set to calculate a *wider* CI, leading to a *more* imprecise Answer;
   ∗ document the outliers but omit them when calculating the CI, leading to a *less* imprecise Answer;
   ∗ suppress the outliers with*out* documenting them and again calculate the CI leading to a *less* imprecise Answer.

   The last of these possibilities is unethical but tempting for the reasons given on page HL2.2 at the end of Note 2.

   As indicated in the quotation from NBS given at the bottom of page HL2.1 in Note 1, omitting outliers may *under*estimate the real variation in a measuring process and so the *actual* confidence level of the CI based on a truncated data set may be far enough below its stated level to have a practically important effect on an Answer.

   Treatment of outliers is an important area of statistical methods but there can be disagreement about which of the available strategies should be used in the context of a particular investigation.

   ● An illustration, from the NB10 data set on page HL2.1, of the effect on CI width of omitting outliers is given in Table 2.6 at the right; the decrease in width of the 95% CI calculated from equation (HL2.12), as possible outliers are omitted, could be large enough to have a practically important effect on an Answer.

Table HL2.6

Observations omitted (#)	Sample size (n)	s (μg)	Coeff. $t^*_{n-1,0.95}$	95% CI (μg below 10 g)	Width (μg)	Ratio
None	100	6.47	1.98422	(403.31, 405.87)	2.56	1
375 (94), 437 (86)	98	4.78	1.98472	(403.60, 405.52)	1.92	0.75
393 (85), 423 (36)	96	4.28	1.98525	(403.62, 405.36)	1.74	0.68
392 (63), 418 (87)	94	3.88	1.98580	(403.69, 405.27)	1.58	0.62

   - This illustration also reminds us of the *limitations* on the Answer about $\mathbf{M}$, the (true) mass of NB10, imposed by *model* error from two sources, which may compromise point 6 Assumption 1 and Assumption 2:
     - the lack of *normality* and the effect of the outliers on the *robustness* of the CI;
     - the *haphazard* nature of the sample selecting process for repeated measurements.

Example HL2.1: An egg farmer knows from long experience that her hens produce eggs whose weights vary with a standard deviation of 3.0 grams. She is required by an Egg Marketing Board inspector to produce statistical evidence of the average weight of the eggs she sells. If a sample of 100 eggs selected equiprobably from those produced by her hens over a suitable period of time shows an average weight of 56.2 grams, find a 90%, a 95%, and a 99% confidence interval for the average weight of all eggs from the farm.

Answers: We are told that the sample average, based on $n = 100$ eggs, is $\bar{y} = 56.2$ grams; also, $\mathbf{S}$ represented by σ is *known* to be 3.0 grams in this (unusual) situation. Then, if $\overline{\mathbf{Y}}$ is the average weight of all eggs from the farm and using equation (HL2.8) near the middle of page HL2.3, we have (approximately):

a 90% CI for μ representing $\overline{\mathbf{Y}}$ is: $\bar{y} \pm 1.6449\frac{\sigma}{\sqrt{n}} = 56.2 \pm 1.6449 \times \frac{3.0}{\sqrt{100}} = 56.2 \pm 0.493 = (55.7,\ 56.7)$ grams;

a 95% CI for μ representing $\overline{\mathbf{Y}}$ is: $\bar{y} \pm 1.9600\frac{\sigma}{\sqrt{n}} = 56.2 \pm 1.9600 \times \frac{3.0}{\sqrt{100}} = 56.2 \pm 0.588 = (55.6,\ 56.8)$ grams;

a 99% CI for μ representing $\overline{\mathbf{Y}}$ is: $\bar{y} \pm 2.5758\frac{\sigma}{\sqrt{n}} = 56.2 \pm 2.5758 \times \frac{3.0}{\sqrt{100}} = 56.2 \pm 0.773 = (55.4,\ 57.0)$ grams.

Example HL2.2: Five thousand people register to compete in a community marathon run. A physical education researcher wishes to investigate an index of vital capacity among these runners. He selects by EPS 50 names from the registration list and is able to measure the vital capacity for all 50 of these people; the average is 68.0 and

(*continued overleaf*)

Example HL2.2: (continued) the standard deviation is 15.0 in suitable units. Find a $100(1-\alpha)\%$ confidence interval for the average vital capacity of this population of 5,000 runners, for $\alpha = 0.10$, 0.05, 0.02 and 0.01.

Answers: We are told that the sample average, based on $n = 50$ runners, is $\bar{y} = 68.0$ units; also, the population (data) standard deviation **S** is *un*known (as is usual) and so we use its estimate, the sample standard deviation $s = 15.0$ units, as the value of the model (probabilistic) standard deviation σ. Then, if $\bar{\mathbf{Y}}$ is the average vital capacity of all 5,000 runners in the marathon and using equation (HL2.12) on page HL2.6, we have:

a 90% CI for μ representing $\bar{\mathbf{Y}}$ is: $\bar{y} \pm 1.67655\frac{s}{\sqrt{n}} = 68.0 \pm 1.67655 \times \frac{15.0}{\sqrt{50}} = 68.0 \pm 3.556 = (64.4, 71.6)$ units;

a 95% CI for μ representing $\bar{\mathbf{Y}}$ is: $\bar{y} \pm 2.00958\frac{s}{\sqrt{n}} = 68.0 \pm 2.00958 \times \frac{15.0}{\sqrt{50}} = 68.0 \pm 4.263 = (63.7, 72.3)$ units;

a 98% CI for μ representing $\bar{\mathbf{Y}}$ is: $\bar{y} \pm 2.40489\frac{s}{\sqrt{n}} = 68.0 \pm 2.40489 \times \frac{15.0}{\sqrt{50}} = 68.0 \pm 5.102 = (62.9, 73.1)$ units;

a 99% CI for μ representing $\bar{\mathbf{Y}}$ is: $\bar{y} \pm 2.67995\frac{s}{\sqrt{n}} = 68.0 \pm 2.67995 \times \frac{15.0}{\sqrt{50}} = 68.0 \pm 5.685 = (62.3, 73.7)$ units.

4. Modelling a Measuring Process which *has* Inaccuracy

The derivation of a CI for μ in Section 3 on pages HL2.2 and HL2.3 explicitly *ex*cluded measuring inaccuracy; we now extend the discussion to include it, in part because measuring processes are *rarely* without inaccuracy. We limit discussion to the (simpler) case where inaccuracy is *constant* – its value does not *change* with the magnitude of the measured value. Based on Example HL74.2 on page HL74.5 in Statistical Highlight #74, we distinguish two types of investigation.

1. **Calibrating a measuring process to quantify measuring inaccuracy.** The data needed to calibrate a measuring process are obtained by making m (say) independent measurements of a *standard* for the quantity under investigation – recall the discussion on page HL2.1 at the start of this Highlight #2. To analyze such data, we write the model (HL2.13) as equation (HL2.14):

$$_{M}Y_j = \tau + \delta + R_j, \quad j = 1, 2, \ldots, m, \quad R_j \sim N(0, \sigma_M), \quad \text{independent}, \quad \text{EPS}, \qquad \text{-----(HL2.14)}$$

where: the suffix M on the random variable $_{M}Y_j$ emphasizes that its values are as *measured*,
the subscript M on the parameter σ_M reminds us it represents *measuring* process variation,
τ is the value of the standard – a 'true' value,
δ is the (measuring) *bias* – bias in the model represents (measuring) inaccuracy in the real world.

If we take the response variate as $Y_j = {}_{M}Y_j - \tau$, the *difference* between the response as measured and the true value:

$$Y_j = \delta + R_j, \quad j = 1, 2, \ldots, m, \quad R_j \sim N(0, \sigma_M); \quad \text{independent}, \quad \text{EPS}; \qquad \text{-----(HL2.15)}$$

this is now the model of equation (HL2.13), except that the structural component is δ instead of μ. Hence, a CI for the bias δ representing the measuring inaccuracy is calculated from equation (HL2.12) rewritten as equation (HL2.16):

$$\bar{y}_c \pm t^*_{m-1,1-\alpha} s_M/\sqrt{m} \quad \text{[The subscript c reminds us } \bar{y}_c \text{ is the sample average of } \textit{calibration} \text{ data.]} \qquad \text{-----(HL2.16)}$$

2. **A CI for the mean μ representing the population average $\bar{\mathbf{Y}}$ calculated from inaccurate measurements.** Our verbal and symbolic models in Sections 2 and 3 on pages HL2.2 and HL2.7 for a measuring process with *no* inaccuracy are:

measured value = true value + haphazard measurement error; -----(HL2.1)

$$Y_j = \mu + R_j, \quad j = 1, 2, \ldots, n, \quad R_j \sim N(0, \sigma), \quad \text{independent}, \quad \text{EPS}. \qquad \text{-----(HL2.13)}$$

When measuring inaccuracy *is* included, these models become:

measured value = true value + systematic measurement error + haphazard measurement error;-----(HL2.17)

$$Y_j = \mu + \delta + R_j, \quad j = 1, 2, \ldots, n, \quad R_j \sim N(0, \sigma), \quad \text{independent}, \quad \text{EPS}. \qquad \text{-----(HL2.18)}$$

If the estimating process that leads to the CI expression (HL2.12) from the model (HL2.13) is now applied to the model (HL2.18), it is only possible to estimate the *combination* $\mu + \delta$ of the *two* parameters that make up the structural component of this model; we can estimate μ – the parameter of interest – *only* if we have an estimate of δ from calibrating the measuring process, usually in a separate investigation (a FDEAC *sub*cycle) as described in Point 1 above.

∗ It is seldom recognized that using *in*accurate measurement data – which are the rule, *not* the exception – to calculate a CI for μ from equation (HL2.13) *includes* the effect of measuring inaccuracy *unless* δ is estimated by calibration and removed.

A classic discussion of unrecognized measuring inaccuracy is that of Youden in *Enduring Values*, summarized in Statistical Highlight #15.

NOTES: 6. Different notation is used for representing the true value of a measured quantity – μ in Section 2 on page HL2.1 and τ in Point 1 above – for two reasons.

- The purpose of Section 2, starting from the response model (HL2.2), is to reach equation (HL2.12) on page HL2.6 for calculating a CI for the model parameter that is usually denoted μ; this constraint, together with avoiding changing notation during Section 2, dictated choosing μ, rather than the more evocative τ, to represent **M**.
- In Section 4 above, use of τ in Point 1, representing the (true) value of a *standard*, helps us avoid confusing this τ with μ, which represents the (study or respondent) population average $\bar{\mathbf{Y}}$, in the model (HL2.18).

University of Waterloo W. H. Cherry

CONFIDENCE INTERVALS: Quantifying Sampling and Measuring Imprecision (continued 5)

(for some types of Questions)

NOTES: 7. For calculating a CI for μ representing $\overline{\text{¥}}$ from data generated by an *in*accurate measuring process, we distinguish two approaches:

* Leave the inaccurate measuring process *un*adjusted and use the estimate of the bias from the calibration investigation, together with equation (HL2.19) below, to calculate a CI for μ with*out* (estimated) measuring bias – this bias, representing measuring inaccuracy, is thus (mostly) *removed* at the Analysis stage of the (main) FDEAC cycle.

$$\bar{y} - \bar{y}_c \pm t^*_{\nu,1-\alpha}\sqrt{\frac{S^2}{n} + \frac{S_M^2}{m}}, \text{ where } \nu \text{ is the integer closest to } \left(\frac{S^2}{n} + \frac{S_M^2}{m}\right)^2 \Big/ \left(\frac{1}{n-1}\left(\frac{S^2}{n}\right)^2 + \frac{1}{m-1}\left(\frac{S_M^2}{m}\right)^2\right); \quad \text{-----(HL2.19)}$$

the symbols here are from equations (HL2.12) on page HL2.6 and below and (HL2.16) on the facing page HL2.10.

* Use the estimate of the bias from the calibration investigation to make the effect of measuring inaccuracy *negligible* in the investigation context by appropriate *adjustment* of the measuring process. Data from such a *calibrated* measuring process would usually be used in equation (HL2.12) to calculate a CI for μ with*out* measuring bias – the measuring inaccuracy has, in principle, been (mostly) removed by adjusting the measuring process.

$$\bar{y} \pm t^*_{n-1,1-\alpha}\, s/\sqrt{n}, \text{ which can also be written as: } \bar{y} \pm t^*_{n-1,1-\alpha}\sqrt{\frac{S^2}{n}}. \quad \text{-----(HL2.12)}$$

We see from equation (HL2.19) that managing measuring inaccuracy incurs two 'costs':

- the resources needed for the calibration investigation, AND:
- *in*creased imprecision in the estimating process – there is an *additional* term under the square root in equation (HL2.19) compared with equation (HL2.12).
  - Strictly, equation (HL2.12), used with data from a calibrated measuring process, ignores the *uncertainty* in estimating the bias in the calibration investigation; it therefore *under*estimates the width of the CI for μ without bias. The CI from equation (HL2.19) avoids this difficulty and an investigator who uses equation (HL2.12) has a responsibility to check whether the difference in width of the two CIs would be large enough to make a practically important difference in an Answer about $\overline{\text{¥}}$.
    - The numerical coefficient $t^*_{\nu,1-\alpha}$ in equation (HL2.19) will be slightly *smaller* than $t^*_{n-1,1-\alpha}$ in equation (HL2.12) because $\nu > n-1$; this will compensate slightly for the larger estimated standard error in equation (HL2.19) compared with equation (HL2.12).

Equation (HL2.19) illustrates how, often, a 'cost' of managing inaccuracy is increased *imprecision*.

8. Elements which involve *humans* – individually or in a group like a family, business or institution – can introduce difficult measurement issues; an illustration is measuring a response variate like self-esteem, for which a questionnaire like the one at the right below might be used. Measuring difficulties with human units include:

- there may no *standard* for calibrating a measuring process, so *inaccuracy* is difficult to quantify;
- repeated measuring of the *same* element is equivocal – a later measurement may be influenced by an element *remembering* a previous response – so, like inaccuracy, *imprecision* can be difficult to quantify;
- different measuring processes – for instance, different questionnaires for measuring self-esteem – are difficult to compare because of uncertainties in quantifying inaccuracy and imprecision;
- an element's reaction to components of the measuring process may influence the (value of the) response; such components include:
  - the person administering the questionnaire,
  - the questionnaire – an element's background may affect how (s)he interprets a question;
- differences in environmental conditions may have greater influence on the (values of) responses from a person than from an *in*animate element;
- it may be unclear to what degree the measuring process *is*

Gauging Self-Esteem

The Rosenberg Self-Esteem Scale is based on 10 questions. Respondents are asked to **strongly agree, agree, disagree** or **strongly disagree** with these items.

1. On the whole, I am satisfied with myself.
2. At times I think I am no good at all.
3. I feel that I have a number of good qualities.
4. I am able to do things as well as most other people.
5. I feel I do not have much to be proud of.
6. I certainly feel useless at times.
7. I feel that I am a person of worth, at least the equal of others.
8. I wish I could have more respect for myself.
9. All in all, I am inclined to feel that I am a failure.
10. I take a positive attitude toward myself.

Half the questions are phrased positively and half negatively.
For the **positively** phrased questions – numbers **1, 3, 4, 7 and 10** – score as follows: Strongly agree, 4 points; agree, 3; disagree, 2; strongly disagree, 1.
For **negative** questions – numbers **2, 5, 6, 8 and 9** – reverse the scoring so that strongly agree is worth 1 point, etc. The maximum is 40 points, the minimum 10.

Most people in the general population score in the 30-to-40 range. A much smaller number are in the 20s. A score of 10 to 20 is often associated with clinical depression, according to Timothy Owens of Indiana University.

Sources: Conceiving the Self *by Morris Rosenberg (Basic Books, 1979); Timothy Owens.*

(*continued overleaf*)

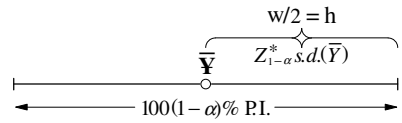
NOTES: (cont.) 8. ● measuring the (intended) response variate – this is, in part, what is called **construct validity** in the social sciences. Our discussion only draws attention to such matters – dealing with them is beyond the scope of these Materials. [The questionnaire overleaf at the lower right of page HL2.11 is taken from Figure 8.8e of the STAT 220 Course Materials.]

5. Calculating a Sample Size in the Plan for an Investigation to Estimate $\overline{\mathbf{Y}}$

The Plan for an investigation may wish to include calculating a sample size that will (under certain assumptions) achieve:

- ∗ a specified total *cost* for the sampling – for example, see equation (HL2.20) at the right; OR:
- ∗ a specified level of *imprecision* for estimating a population attribute – an average in *this* discussion.

$$n = \frac{\text{total budget} - \text{fixed costs}}{\text{cost per unit sampled}} \qquad \text{-----(HL2.20)}$$



Discussion below of the second approach allows for imprecision to be specified in one of three ways – interval *width* (w), interval *half-width* (h) or the *standard deviation of* $\overline{Y}$ [$s.d.(\overline{Y})$] – see the diagram at the right. The interval in this diagram has the population average, $\overline{\mathbf{Y}}$, as its centre – it is therefore a *probability* (*not* a confidence) interval, a PI not a CI; the appropriate numerical coefficent is thus $Z^*_{1-\alpha}$, not $t^*_{n-1,1-\alpha}$.

Specifying imprecision by *interval width*: $w = 2h$ so: $h = w/2$. -----(HL2.21)

Specifying imprecision by *interval half-width*: $h = Z^*_{1-\alpha}s.d.(\overline{Y})$ so: $s.d.(\overline{Y}) = h/Z^*_{1-\alpha}$. -----(HL2.22)

Specifying imprecision by the *standard deviation of* $\overline{Y}$: $s.d.(\overline{Y}) = \mathbf{S}\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}}$ so: $\frac{1}{n} = \left(\frac{s.d.(\overline{Y})}{\mathbf{S}}\right)^2 + \frac{1}{\mathbf{N}}$. -----(HL2.23)

If $n \ll \mathbf{N}$, an *approximate* sample size is: $n_0 = \left(\frac{\mathbf{S}}{s.d.(\overline{Y})}\right)^2 = \left(\frac{\mathbf{S}Z^*_{1-\alpha}}{h}\right)^2 = \left(\frac{2\mathbf{S}Z^*_{1-\alpha}}{w}\right)^2$. -----(HL2.24)

With*out* approximation, using equation (HL2.23): $\frac{1}{n} = \frac{1}{n_0} + \frac{1}{\mathbf{N}} = \frac{1}{n_0}\left(1 + \frac{n_0}{\mathbf{N}}\right)$ so: $n = n_0 \Big/ \left(1 + \frac{n_0}{\mathbf{N}}\right)$. -----(HL2.25)

For any of the three expressions on the right-hand side of equation (HL2.24), we need a value to use for $\mathbf{S}$, the (study or respondent) population (data) standard deviation, which is usually *un*known at the Plan stage of the FDEAC cycle. Possibilities include:

- ∗ using a value obtained in a *previous* investigation of the same (or a similar) population;
- ∗ using a value from data obtained in the *pilot survey* of the current investigation;
- ∗ if there is information on the *range* of response variate values in the population and if approximate normality can be assumed, this range could be taken as $k\mathbf{S}$, where k can be taken as 4, 5 or 6 – see the first side (page 5.1) of STAT 231 Figure 5.1.

A sample size calculated in the Plan should be used with due recognition of its limitations imposed by:

- ⊙ whether it is n_0 or n that has been calculated,
- ⊙ uncertainty in the value used for $\mathbf{S}$,
- ⊙ the likely *response* rate – for example, a 25% response rate (which is currently *optimistic* for a political poll, for instance) would entail an actual sample size four times larger than the calculated value,

and, when imprecision is specified by w or h:

- ⊙ use of $Z^*_{1-\alpha}$ in the sample size calculation compared with the (slightly larger) $t^*_{\nu,1-\alpha}$ or $t^*_{n-1,1-\alpha}$ in the CI calculation.

NOTES: 9. The approximate sample size n_0 is the square of the ratio of the population (data) standard deviation $\mathbf{S}$ to the standard deviation of $\overline{Y}$ – see the first expression on the right-hand side of equation (HL2.24) above.

10. Non-response – the usual terminology when the elements involve humans – or, more generally, missing data are the source of *non-response error* which imposes limitations on an Answer. This category of error is more easily overlooked when, as in the discussion of Section 5 above, conventional terminology of *sample size* is used. The discussion would be better framed using the terminology of equation (HL2.26) below, which is specific to these Materials and distinguishes the elements (or units)) *selected* from the elements (or units) which *respond*:

selection = sample + non-respondents AND: selection size (n_s) = sample size (n) + non-response (n_{nr}). -----(HL2.26)

Using from the start this terminology, which is evocative of a pervasive source of error, would be better than first using conventional terminology and then making good its deficiency by introducing non-response later in the discussion – see also Section 5 on pages 5.24 to 5.26 in Figure 5.7 of the STAT 231 Course Materials.

6. Appendix 1: t_ν Distribution History

The t distribution was developed early this century by William Sealy Gosset (1876-1937), who was born in Canterbury, England, and educated at Winchester and New College, Oxford, where he obtained first-class degrees in mathematics and natural science. On leaving Oxford in the fall of 1899, he joined the famous brewing firm of Guinness in Dublin; he remained with Guinness all his life, becoming in 1935 chief brewer at Park Royal, the firm's newly-established brewery in London.

University of Waterloo W. H. Cherry

CONFIDENCE INTERVALS: Quantifying Sampling and Measuring Imprecision (continued 6)

(for some types of Questions)

In the early 1900s, scientific methods and laboratory determinations were beginning to be seriously applied to brewing, and this led Gosset to recognize the need for adequate methods to deal with small samples when examining relationships among the quality of raw materials for beer such as barley and hops, the conditions of production, and the finished product. The importance of controlling the quality of barley also led Gosset to study the design of agricultural field trials.

Gosset was once described by Sir Ronald Fisher (considered by many to be the father of modern statistical methods) as *the Faraday of statistics*, because he was not a profound mathematician but had a superb intuitive faculty that enabled him to grasp general principles and see their relevance to practical ends.

In his work at the brewery, Gosset had been struck by the importance of assessing the precision of the estimator represented by the average of a small sample. The usual procedure at the time was to calculate the sample average ($\bar{y}$) and sample standard deviation (s) and to proceed as though the corresponding random variable $\bar{Y}$ had a Gaussian distribution with the same average as the population from which the sample was selected and with standard deviation $s/\sqrt{n}$. The difficulty is that s is only an *estimate* of **S**, the population (data) standard deviation, and Gosset's intuition told him that, for small samples, this procedure *over* estimates the precision of $\bar{Y}$ as an estimator of the population average $\bar{\mathbf{Y}}$. This finally led him to develop the *t* distribution for use in place of the Gaussian distribution when **S**, represented in the model by σ, is estimated by s.

Gosset is known mainly by the pseudonym *student*, under which he published most of his statistical work because his employer regarded his findings as proprietary information and did not wish to bring it to their competitors' attention by the use of Gosset's real name and his affiliation with Guinness.

More information about Gosset's contributions to science and to statistics can be obtained from pages 409-413 of the *International Encyclopedia of Statistics*, Volume 1, edited by W. H. Kruskal and J.M. Tanur (The Free Press, New York, 1978) [DC Library call number HA17.I63 1978]; most of the description above is taken from this source.

7. Appendix 2: Least Squares Estimating

Our discussion of CIs in Section 2 on pages HL2.2 and HL2.3 uses the sample average $\bar{y}$ as an estimate of the model mean μ representing the population average $\bar{\mathbf{Y}}$ and, on page HL2.6, the sample (data) standard deviation s as an estimate of the model (probabilistic) standard deviation σ, representing the population (data) standard deviation **S**. In this Appendix, we discuss a *process* – the *method of least squares* – for obtaining expressions for estimates and estimators of parameters in the structural component of the model; for the response model (HL2.13) on page HL2.7, the structural component takes the (simplest) form μ. As summarized in Table HL2.7 at the right below, we introduce additional notation for a measure of location and of variation:

- ∗ a hat on a model parameter denotes an *estimate* (here, $\hat{\mu}$ or $\hat{\sigma}$) – a *value*, like $\bar{y}$ and s;
- ∗ a tilde on a model parameter denotes an *estimator* (here, $\tilde{\mu}$ or $\tilde{\sigma}$) – a *random variable*, like $\bar{Y}$ and S.

Table HL2.7

	Location	Variation
Population attribute	Average $\bar{\mathbf{Y}}$	(Data) S.d. **S**
Model parameter	Mean μ	(Probabilistic) S.d. σ
Estimate – a value	$\bar{y}$ or $\hat{\mu}$	s or $\hat{\sigma}$
Estimator – a random variable	$\bar{Y}$ or $\tilde{\mu}$	S or $\tilde{\sigma}$

We set out the estimating process in numbered steps for two *equivalent* derivations using calculus and linear algebra. The name *least squares* is meant to evoke the method's central idea: *minimizing the sum of the squared residuals* (steps 4 and 5 below).

$Y_j = \mu + R_j, \quad j = 1, 2, \ldots, n; \quad R_j \sim N(0, \sigma); \quad$ independent, EPS.

Step 1: $Y_j = \mu + R_j, \quad j = 1, 2, \ldots, n.$

Step 2: $y_j = \mu + r_j, \quad j = 1, 2, \ldots, n.$

Step 3: $r_j = y_j - \mu, \quad j = 1, 2, \ldots, n.$

Step 4: $g(\mu) = \sum_{j=1}^{n} r_j^2 = \sum_{j=1}^{n} [y_j - \mu]^2.$

Step 5: $\frac{dg}{d\mu} = -2\sum_{j=1}^{n}[y_j - \mu],$

which is zero when: $\sum_{j=1}^{n}[y_j - \hat{\mu}] = 0;$

i.e., $\sum_{j=1}^{n} y_j - n\hat{\mu} = 0,$

or: $\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} y_j = \bar{y},$ the **estimate** of μ.

Step 6: $\frac{d^2g}{d\mu^2} = -2\sum_{j=1}^{n}[-1] = 2n,$

which is **positive**, as required for a **minimum**.

Also: $\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{n}\hat{r}_j^2}{n-q}} = \sqrt{\frac{\sum_{j=1}^{n}[y_j - \hat{\mu}]^2}{n-1}} = \sqrt{\frac{\sum_{j=1}^{n}[y_j - \bar{y}]^2}{n-1}}.$

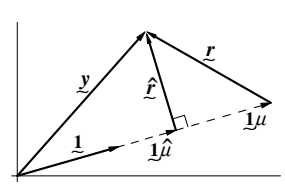
NOTES: 11. These derivations involve random variables

Step 1: $Y_j = \mu + R_j, \quad j = 1, 2, \ldots, n.$

Step 2: $y_j = \mu + r_j, \quad j = 1, 2, \ldots, n.$

Step 3: $\underset{\sim}{y} = \underset{\sim}{1}\mu + \underset{\sim}{r}$, where: $\underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underset{\sim}{1} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \quad \underset{\sim}{r} = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{pmatrix}.$

Step 4: The least squares estimate of μ is determined by the vector $\hat{\underset{\sim}{r}}$, the vector $\underset{\sim}{r}$ of **minimum** length; this occurs when $\underset{\sim}{r}$ is **perpendicular** to the vector $\underset{\sim}{1}$, so we set the dot (or inner) product $\underset{\sim}{1}\cdot\hat{\underset{\sim}{r}}$ to zero.



Step 5: $\underset{\sim}{1}^T\hat{\underset{\sim}{r}} = 0, \quad$ *i.e.*, $\underset{\sim}{1}^T(\underset{\sim}{y} - \underset{\sim}{1}\hat{\mu}) = 0,$

i.e., $\underset{\sim}{1}^T\underset{\sim}{y} - \underset{\sim}{1}^T\underset{\sim}{1}\hat{\mu} = 0,$

or: $\hat{\mu} = (\underset{\sim}{1}^T\underset{\sim}{1})^{-1}(\underset{\sim}{1}^T\underset{\sim}{y}) = n^{-1}\sum_{j=1}^{n} y_j = \bar{y},$ the **estimate** of μ.

Also: $\hat{\sigma} = \sqrt{\frac{\hat{\underset{\sim}{r}}^T\hat{\underset{\sim}{r}}}{n-q}} = \sqrt{\frac{(\underset{\sim}{y} - \underset{\sim}{1}\hat{\mu})^T(\underset{\sim}{y} - \underset{\sim}{1}\hat{\mu})}{n-1}} = \sqrt{\frac{\sum_{j=1}^{n}[y_j - \bar{y}]^2}{n-1}}.$

(*continued overleaf*)

NOTES: (cont.)

11. (Step 1) and their values (Step 2 onwards) in the context of a response *model*; values are therefore represented by *italic* letters like $\bar{y}$. The three modelling *assumptions* (discussed in point 6 of Section 3 on pages HL2.7 and HL2.8) provide justification for regarding the *measured* response variate values y_j, $j = 1, 2, \ldots, n$, of the n sampled elements as values, y_j, $j = 1, 2, \ldots, n$, of random variables, Y_j, $j = 1, 2, \ldots, n$ – see also the top of page HL74.2 in Statistical Highlight #74.

12. The estimate $\hat{\sigma}$ of the model parameter σ is *not* a least squares estimate. The estimate $\hat{\sigma}^2$ is an *un*biased estimate of $\mathbf{S}^2$ but $\hat{\sigma}$ is *not* (quite) unbiased for $\mathbf{S}$ [*e.g.*, see the bottom of page HL77.5 in Section 5 and Appendix 3 on page HL77.9 of Statistical Highlight #77.

13. In the diagram overleaf at the lower right of page HL2.13 in Step 4 of the the linear algebra least squares derivation, the estimated residual vector $\hat{\underset{\sim}{r}}$ lies in the subspace perpendicular to *span* <1>; that is, $\hat{\underset{\sim}{r}}$ lies in the orthogonal complement of a one-dimensional subspace, which is a subspace with $n-1$ dimensions – this is one explanation of the $n-1$ denominator in four expresssions for $\hat{\sigma}$ overleaf at the bottom of the page HL2.13.

14. A **general** form for a response model – a response variate expressed as a **structural component** plus **noise** – is shown in two versions in equation (HL2.27):

$$\left.\begin{array}{l} Y_j = \mu_j + R_j, \\ \underset{\sim}{Y} = \mathbf{X}\underset{\sim}{\Theta} + \underset{\sim}{R} \end{array}\right\} \begin{array}{l} j = 1, 2, \ldots, n, \quad R_j \sim N(0, \sigma), \quad \text{independent}, \quad \text{EPS}, \\ \mu_j = \theta_1 x_{j1} + \ldots + \theta_k x_{jk} + \ldots + \theta_q x_{jq} \quad [\text{explanatory variates } X_1, \ldots, X_k, \ldots, X_q]; \end{array} \qquad \text{-----(HL2.27)}$$

- in the upper version, the structural component μ_j is a linear combination of q (unknown) parameters θ_k and (known) explanatory variate values x_{jk};
- in the lower version, the explanatory variate values form the $n \times q$ matrix $\mathbf{X}$ and the parameters are the $q \times 1$ column vector $\underset{\sim}{\Theta}$.
  - $\underset{\sim}{Y}$ and $\underset{\sim}{R}$ are $n \times 1$ column vectors with $n \gg q$; *i.e.*, there are *many* more observations than there are parameters to be estimated.
  - The *transpose* of $\mathbf{X}$, $\mathbf{X}^T$, in the derivation below is a $q \times n$ matrix.

The response model (HL2.13) on page HL2.7 is the simplest case of the general model (HL2.27) above.

The method of least squares involves minimizing $\|\underset{\sim}{R}\|$, which means that $\hat{\underset{\sim}{R}}$ must be perpendicular to every column of $\mathbf{X}$, so we set the dot product $\mathbf{X} \cdot \hat{\underset{\sim}{R}}$ to zero; that is, we set: $\mathbf{X}^T\hat{\underset{\sim}{R}} = \underset{\sim}{0}$.

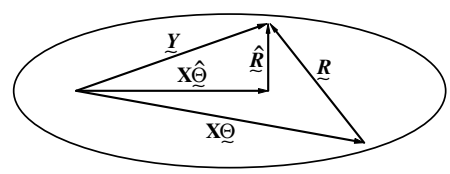
Thus: $\mathbf{X}^T(\underset{\sim}{Y} - \mathbf{X}\hat{\underset{\sim}{\Theta}}) = \underset{\sim}{0}$ or: $\mathbf{X}^T\underset{\sim}{Y} - \mathbf{X}^T\mathbf{X}\hat{\underset{\sim}{\Theta}} = \underset{\sim}{0}$ or: $\mathbf{X}^T\mathbf{X}\hat{\underset{\sim}{\Theta}} = \mathbf{X}^T\underset{\sim}{Y}$ so: $\hat{\underset{\sim}{\Theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underset{\sim}{Y}$. -----(HL2.28)

In equation (HL2.28), we replace $\underset{\sim}{Y}$ by $\mathbf{X}\underset{\sim}{\Theta} + \underset{\sim}{R}$ from the model (HL2.27) to obtain:

$$\hat{\underset{\sim}{\Theta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{X}\underset{\sim}{\Theta} + \underset{\sim}{R}) = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}\Theta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underset{\sim}{R} = \Theta + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\underset{\sim}{R}, \qquad \text{-----(HL2.29)}$$

which shows that $\hat{\underset{\sim}{\Theta}}$ is Θ plus a linear combination of residuals.

In the diagram for $\mathbb{R}^n$ at the right, $\hat{\underset{\sim}{R}}$ must lie in the orthogonal complement of the span of the columns of $\mathbf{X}$, a q-dimensional subspace represented by the ellipse; its orthogonal complement (the subspace in which $\hat{\underset{\sim}{R}}$ lies) has dimension $n-q$, which is the divisor in the leftmost two expressions for $\hat{\sigma}$ overleaf at the bottom of page HL2.13.



8. Appendix 3: t_ν Distribution Theory

Our derivation of the t_ν distribution in this Appendix 3 on pages HL2.14 to HL2.16 involves three other distributions:

- the *normal* distribution which we are already familiar with, AND:
- two distributions we have not yet encountered: – the χ^2 distribution,
  – the K distribution;

as well as leading here to the t_ν distribution, these two distributions have other uses in these Materials.

The χ^2_ν distribution: If $Z_1, Z_2, \ldots, Z_\nu$ are probabilistically independent $N(0, 1)$ random variables, the *sum of their squares*, the random variable X of equation (HL2.30), has a χ^2 distribution with ν degrees of freedom; symbolically, we *write* equation (HL2.31) and we *say* X is 'chi squared with ν degrees of freedom' ('chi' rhymes with 'hi' – it is pronounced 'ki').

$$X = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2 \qquad \text{-----(HL2.30)}$$

$$X \sim \chi^2_\nu \qquad \text{-----(HL2.31)}$$

Because it involves *squares* of $N(0, 1)$ random variables, the χ^2 distribution takes only non-negative values; its p.d.f., for 5, 10, 25 and 50 degrees of freedom, is shown at the upper right of the facing page HL2.15. In these diagrams:

- the four vertical axis scales have the *same* unit size, as do the four horizontal axis scales;
- we see that the χ^2 distribution in *un*symmetrical about its highest point (*un*like the symmetrical normal and t_ν distributions) but it becomes more symmetrical as ν increases.

Two other properties of the χ^2 distribution are:

- its *mean* is ν; if $Z \sim N(0, 1)$, $[sd.(Z)]^2 = E(Z^2) - E[(Z)]^2$ and so: $E(Z^2) = 1$;
  hence, using equation (13.1.30): $E(X) = E(Z_1^2) + E(Z_2^2) + \ldots + E(Z_\nu^2) = \nu$. -----(HL2.32)

We see in the four diagrams that the mean (indicated by↓) lies a little to the *right* of the highest point (↑) of each p.d.f.

University of Waterloo W. H. Cherry

CONFIDENCE INTERVALS: Quantifying Sampling and Measuring Imprecision (continued 7)

(for some types of Questions)

- its *standard deviation* (and, correspondingly, its *spread*) *in*creases as ν increases;
- two probabilistically independent random variables $X \sim \chi^2_\nu$ and $W \sim \chi^2_\tau$ have sum: $X + W \sim \chi^2_{\nu+\tau}$; -----(HL2.33)

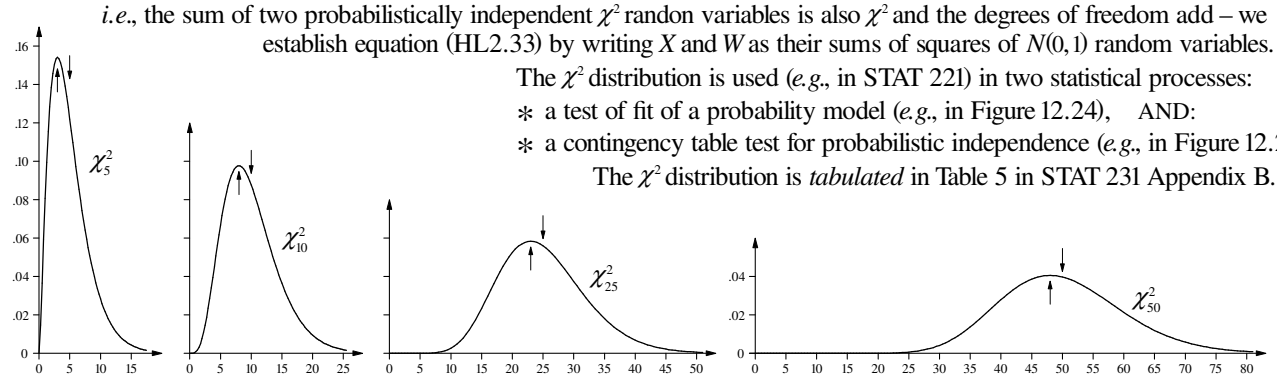
i.e., the sum of two probabilistically independent χ^2 randon variables is also χ^2 and the degrees of freedom add – we establish equation (HL2.33) by writing X and W as their sums of squares of $N(0,1)$ random variables.

The χ^2 distribution is used (*e.g.*, in STAT 221) in two statistical processes:

∗ a test of fit of a probability model (*e.g.*, in Figure 12.24), AND:

∗ a contingency table test for probabilistic independence (*e.g.*, in Figure 12.26).

The χ^2 distribution is *tabulated* in Table 5 in STAT 231 Appendix B.



The K_ν distribution: If the random variable X is χ^2 with ν degrees of freedom, and if the random variable U is the square root of X divided by ν, then U is K with ν degrees of freedom – see equation (HL2.34).

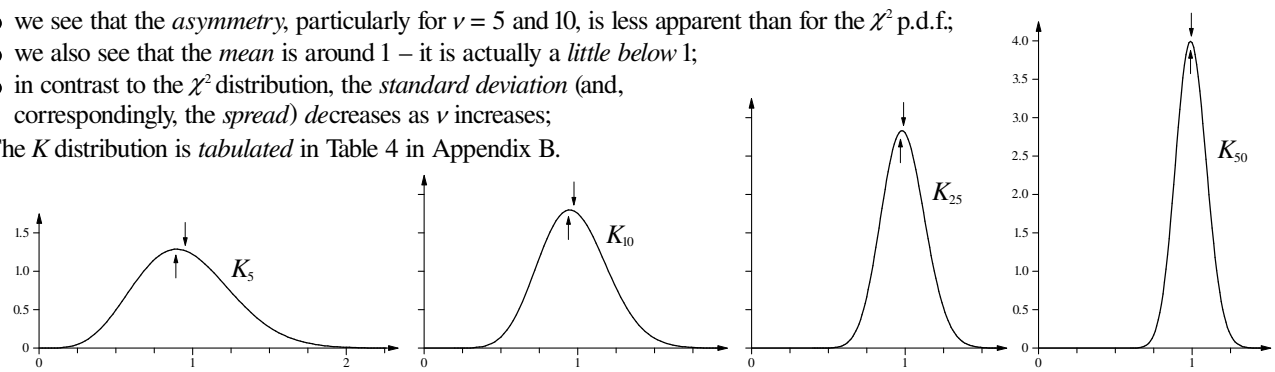
$$\text{If } X \sim \chi^2_\nu \text{ and } U = \sqrt{\frac{X}{\nu}}, \quad \text{then: } U \sim K_\nu. \qquad \text{-----(HL2.34)}$$

i.e., a K_ν random variable is the square root of a χ^2_ν random variable divided by its degrees of freedom.

Like the χ^2 distribution, the K distribution takes only non-negative values; its p.d.f., for 5, 10, 25 and 50 degrees of freedom, is shown below. In these diagrams:

- the four vertical axis scales have the *same* unit size, as do the four horizontal axis scales;
- we see that the *asymmetry*, particularly for $\nu = 5$ and 10, is less apparent than for the χ^2 p.d.f.;
- we also see that the *mean* is around 1 – it is actually a *little below* 1;
- in contrast to the χ^2 distribution, the *standard deviation* (and, correspondingly, the *spread*) *de*creases as ν increases;

The K distribution is *tabulated* in Table 4 in Appendix B.



The distribution of $\tilde{\sigma}$: the random variable $\tilde{\sigma}$ (which can also be denoted S) is the *estimator* of the model parameter σ, the (probabilistic) standard deviation of the normal model for the residuals in equation (HL2.13) on page HL2.7. It is given as two expressions in equation (HL2.35); the first is based on the third expression for $\hat{\sigma}$ at the lower left of page HL2.13, except that the lower-case *y*s representing *values* have been replaced by upper-case letters representing the corresponding *random variables.* The second expression in equation (HL2.35) follows from the first because $Y_j = \mu + R_j$ and $\overline{Y} = \mu + \overline{R}$ – recall equations (HL2.13) and (HL2.3).

$$\tilde{\sigma} = \sqrt{\frac{\sum_{j=1}^{n}[Y_j - \overline{Y}]^2}{n-1}} = \sqrt{\frac{\sum_{j=1}^{n}[R_j - \overline{R}]^2}{n-1}} \qquad \text{-----(HL2.35)}$$

Because $\sum_{j=1}^{n} R_j = n\overline{R}$, the numerator of the second expression in equation (HL2.35) expands as in equation (HL2.36).

$$\sum_{j=1}^{n}[R_j - \overline{R}]^2 = \sum_{j=1}^{n} R_j^2 - n\overline{R}^2 \qquad \text{-----(HL2.36)}$$

Dividing equation (HL2.36) through by σ^2, we obtain equation (HL2.37).

$$\frac{\sum_{j=1}^{n}[R_j - \overline{R}]^2}{\sigma^2} = \sum_{j=1}^{n}\left(\frac{R_j}{\sigma}\right)^2 - \left(\frac{\overline{R}}{\sigma/\sqrt{n}}\right)^2 \qquad \text{-----(HL2.37)}$$

Because $R_j \sim N(0, \sigma)$, the random variables which are the two terms on the RHS of equation (HL2.37) have, respectively, χ^2_n and χ^2_1 distributions, so that from equation (HL2.33) above, the random variable on its LHS has a χ^2_{n-1} distribution, assuming (as is actually the case) that the two RHS random variables are probabilistically independent. Dividing the LHS by $n-1$ (the degrees of freedom of its χ^2 distribution) and taking the square root, we have from equation (HL2.34) above:

$$\sqrt{\frac{\sum_{j=1}^{n}[R_j - \overline{R}]^2}{(n-1)\sigma^2}} = \frac{\tilde{\sigma}}{\sigma} \sim K_{n-1} \quad \text{or:} \quad \tilde{\sigma} \sim \sigma K_{n-1}. \qquad \text{-----(HL2.38)}$$

The t_ν distribution: If the random variable Z is $N(0,1)$ and the random variable U is K with ν degrees of freedom, and if the random variable T is the quotient of Z and U, then T is t with ν degrees of freedom – see equation (HL2.39).

$$\text{If } Z \sim N(0,1),\ U \sim K_\nu \text{ and } T = \frac{Z}{U}, \quad \text{then: } T \sim t_\nu. \qquad \text{-----(HL2.39)}$$

i.e., a t_ν random variable is the quotient of probabilistically independent $N(0,1)$ and K_ν random variables.

The distribution of $\tilde{\mu}$: the random variable $\tilde{\mu}$ (which can also be denoted $\overline{Y}$) is the *estimator* of the model parameter μ, the

(*continued overleaf*)

structural component of the response model (HL2.13) on page HL2.7; μ is also the mean of the normal model for the distribution of the Y_js. From equation (HL2.13), we have:

$$R_j \sim N(0, \sigma), \quad Y_j \sim N(\mu, \sigma), \quad \tilde{\mu} \equiv \overline{Y} \sim N(\mu, \sigma/\sqrt{n}) \text{ so: } \frac{\tilde{\mu}-\mu}{\sigma/\sqrt{n}} \equiv \frac{\overline{Y}-\mu}{\sigma/\sqrt{n}} \sim N(0, 1). \qquad \text{-----(HL2.40)}$$

Also, equation (HL2.38) is:

$$\tilde{\sigma}/\sigma \sim K_{n-1}. \qquad \text{-----(HL2.38)}$$

Dividing the two random variables on the LHSs of equations (HL2.40) and (HL2.38) and using equation (HL2.39), we obtain:

$$\frac{\tilde{\mu}-\mu}{\tilde{\sigma}/\sqrt{n}} \equiv \frac{\overline{Y}-\mu}{S/\sqrt{n}} \sim t_{n-1}. \qquad \text{-----(HL2.41)}$$

Equation (HL2.41) is the basis for the expression (HL2.12) on page HL2.6 for a (realized) CI for μ representing $\overline{\mathbf{Y}}$.

NOTES: 15. Table HL2.8 at the right gives the mean and standard deviation of the χ^2, K and t distributions, both as a general expression in terms of ν and for ν = 5, 10, 25 and 50. These values (rounded to 2 decimal places in four of the sets) are of interest in relation to the three sets of four p.d.f. diagrams at the bottom of page HL2.6 and overleaf on page HL2.15.

Table HL2.8:

		Mean					Standard deviation			
Df:	ν =	5	10	25	50	ν =	5	10	25	50
χ^2	ν	5	10	25	50	$\sqrt{2\nu}$	3.16	4.47	7.07	10
K	$\sqrt{\frac{2}{\nu}}\frac{\Gamma[(\nu+1)/2]}{\Gamma[\nu/2]}$	0.95	0.98	0.99	1.00	$\sqrt{1-\text{mean}^2}$	0.31	0.22	0.14	0.10
t	0	0	0	0	0	$\sqrt{\nu/(\nu-2)}$	1.29	1.12	1.04	1.02

9. Appendix 4: A Confidence Interval for σ representing $\mathbf{S}$ (and for σ^2 representing $\mathbf{S}^2$)

We obtain the CI expression for the (probabilistic) standard deviation of the model (HL2.13) on page HL2.7 from the theory in Appendix 3, using the *same* reasoning as in equations (HL2.5) to (HL2.9) on page HL2.3 in the derivation of a CI for μ representing $\overline{\mathbf{Y}}$.

For a random variable U with a K_{n-1} distribution, we can write equation (HL2.42), where the values of ${}_l k^*_{n-1,1-\alpha}$ and ${}_r k^*_{n-1,1-\alpha}$ are looked up from the table of the K distribution in STAT 231 Table 4 and their suffices l and r refer to the left and right tails of the distribution – if $\alpha = 0.05$, the RHS probability is 0.95. and the k^*s are the 2.5 and 97.5 percentiles of the K_{n-1} distribution.

$$\Pr({}_l k^*_{n-1,1-\alpha} < U \le {}_r k^*_{n-1,1-\alpha}) = 1-\alpha \qquad \text{-----(HL2.42)}$$

Using equation (HL2.38) in equation (HL2.42), we have:

$$\Pr({}_l k^*_{n-1,1-\alpha} < \tilde{\sigma}/\sigma \le {}_r k^*_{n-1,1-\alpha}) = 1-\alpha. \qquad \text{-----(HL2.43)}$$

Rearranging the terms of equation (HL2.43) and changing the wording to that appropriate for a CI, we obtain equation (HL2.44):

$$\textit{a (realized) } 100(1-\alpha)\% \textit{ CI for } \sigma \textit{ representing } \mathbf{S} \textit{ is } (\hat{\sigma}/{}_r k^*_{n-1,1-\alpha}, \hat{\sigma}/{}_l k^*_{n-1,1-\alpha}) \equiv (s/{}_r k^*_{n-1,1-\alpha}, s/{}_l k^*_{n-1,1-\alpha}). \qquad \text{-----(HL2.44)}$$

The *smaller lower* limit of this CI involves the *larger* ${}_r k^*$ in its denominator, the *larger upper* limit involves the *smaller* ${}_l k^*$.

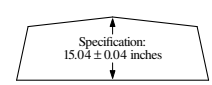
As n increases, the width of the CI of equation (HL2.44) *de*creases with the decreasing standard deviation of the K_{n-1} distribution, reflecting [as with the CI for μ of equation (HL2.12)] decreasing estimating imprecision with increasing sample size.

Statistics' unfortunate traditional emphasis on variance (rather than standard deviation) means that, elsewhere, the CI for σ of equation (HL2.44) above may instead be given as a CI for σ^2 representing $\mathbf{S}^2$. The theory overleaf in Appendix 3 leads to the expression (HL2.45) by the same reasoning as above

$$\left(\sum_{j=1}^{n}[y_j-\overline{y}]^2/{}_r x^*_{n-1,1-\alpha}, \sum_{j=1}^{n}[y_j-\overline{y}]^2/{}_l x^*_{n-1,1-\alpha}\right) \equiv \left([n-1]s^2/{}_r x^*_{n-1,1-\alpha}, [n-1]s^2/{}_l x^*_{n-1,1-\alpha}\right) \qquad \text{-----(HL2.45)}$$

but, instead of the K_{n-1} distribution of equation (HL2.38), it is based on the χ^2_{n-1} distribution of the LHS of equation (HL2.37) written in terms of Y not R using equation (HL2.35); the values of ${}_l x^*_{n-1,1-\alpha}$ and ${}_r x^*_{n-1,1-\alpha}$ are looked up (*e.g.*, in Table 5 on pages Ap. 7 and Ap. 8 in STAT 231 Appendix B) as the relevant percentiles from the left and right tails of the χ^2_{n-1} distribution.

Example HL2.3: In an automobile assembly process, windows which do not open involve a sheet of glass, with a flexible sealing ring around it, being pressed into the corresponding opening in the metal vehicle body. For an efficient process, the fit in the body opening of the glass plus ring is critical: an *over*size glass may be difficult to fit, an *under*size glass may leak when the finished vehicle is exposed to rain. Many dimensions of the glass, ring and opening are critical for a proper fit; the data (and their stemplot) at the right are for *one* such dimension for a rear window of a pickup truck – the distance from the bottom of the cab rear opening to the highest point in the middle of its top side. The 30 cabs in the sample were obtained over a 5-day week (when about 7,200 vehicles were produced) by selecting 6 per day – one about every hour over one shift.



Day	Cab rear opening height (inches)					
1	15.058	15.066	15.046	15.044	15.066	15.026
2	15.038	15.028	15.042	15.048	15.040	15.070
3	15.028	15.046	15.068	14.998	15.024	15.006
4	15.034	15.040	15.064	15.050	15.032	15.016
5	15.024	15.026	15.032	15.046	15.020	15.018

14.99	8
15.00	6
15.01	68
15.02	04466 88
15.03	2248
15.04	00246 668
15.05	08
15.06	4668
15.07	0

For these data: sum = 451.144, average = $\overline{y} \equiv \hat{\mu} = 15.038\,1\dot{3}$ in., $[\sum_{j=1}^{n}[y_j-\overline{y}]^2 = 0.009\,863\,47]$

sum of squares = 6,784.373 488, (data) standard deviation = $s \equiv \hat{\sigma} = 0.018\,442\,333$ in.

Calculating (realized) CIs for μ and σ from these data is shown in Table HL2.9 at the right. Using ${}_r x^*_{29,0.95} = 45.7223$ and ${}_l x^*_{29,0.95} = 16.0471$ from the χ^2 distribution Table 5 in Appendix B, equation (HL2.45) gives a 95% CI for σ^2 as (0.000 2157, 0.000 61466) in., whose square root is the 95% CI for σ in Table HL2.9 (apart from rounding error).

Table HL2.9:

		CI for μ [equation (HL2.12)]			CI for σ [equation (HL2.44)]			
$1-\alpha$	Level	$t_{29,1-\alpha}$	CI	Width	${}_r k^*_{29,1-\alpha}$	${}_l k^*_{29,1-\alpha}$	CI	Width
0.90	90%	1.69913	(15.032, 15.044)	0.012	1.211 40	0.781 430	(0.0152, 0.0236)	0.0084
0.95	95	2.04523	(15.031, 15.045)	0.014	1.255 64	0.743 873	(0.0147, 0.0248)	0.0101
0.99	99	2.75639	(15.029, 15.047)	0.018	1.343 38	0.672 647	(0.0137, 0.0274)	0.0137

University of Waterloo W. H. Cherry

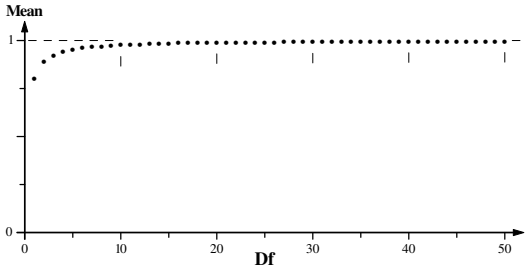
CONFIDENCE INTERVALS: Quantifying Sampling and Measuring Imprecision (continued 8)
(for some types of Questions)

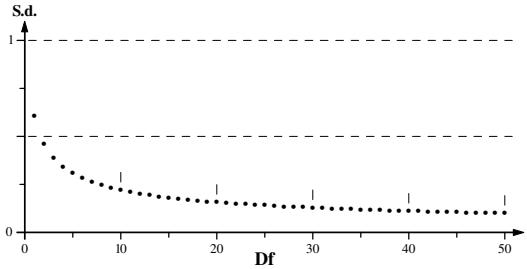
DATA SOURCE: R.J. MacKay and R.W. Oldford. 2004 Course Notes for Statistics 231, Chapter 13, page 2.
The mean and standard deviation of the *K* distribution in Table HL2.8 on the facing page HL2.16 were kindly provided by Prof. MacKay, who also calculated the more extensive values in Table HL2.10 below.

Table HL2.10: *K* distribution

Df	Mean	Df	Mean	Df	Mean	Df	Mean	Df	Mean
1	0.797 885	11	0.977 559	21	0.988 170	31	0.991 969	41	0.993 922
2	0.886 227	12	0.979 406	22	0.988 705	32	0.992 219	42	0.994 066
3	0.921 318	13	0.980 971	23	0.989 193	33	0.992 454	43	0.994 203
4	0.939 986	14	0.982 316	24	0.989 640	34	0.992 675	44	0.994 335
5	0.951 533	15	0.983 484	25	0.990 052	35	0.992 884	45	0.994 460
6	0.959 369	16	0.984 506	26	0.990 433	36	0.993 080	46	0.994 580
7	0.965 030	17	0.985 410	27	0.990 786	37	0.993 267	47	0.994 695
8	0.969 311	18	0.986 214	28	0.991 113	38	0.993 443	48	0.994 806
9	0.972 659	19	0.986 934	29	0.991 418	39	0.993 611	49	0.994 911
10	0.975 350	20	0.987 583	30	0.991 703	40	0.993 770	50	0.995 013

Df	S.d.	Df	S.d.	Df	S.d.	Df	S.d.	Df	S.d.
1	0.602 810	11	0.210 660	21	0.153 361	31	0.126 479	41	0.110 090
2	0.463 251	12	0.201 903	22	0.149 878	32	0.124 503	42	0.108 780
3	0.388 811	13	0.194 152	23	0.146 621	33	0.122 617	43	0.107 515
4	0.341 214	14	0.187 230	24	0.143 569	34	0.120 815	44	0.106 294
5	0.307 547	15	0.180 998	25	0.140 699	35	0.119 089	45	0.105 113
6	0.282 155	16	0.175 349	26	0.137 994	36	0.117 436	46	0.103 970
7	0.262 138	17	0.170 197	27	0.135 440	37	0.115 849	47	0.102 865
8	0.245 839	18	0.165 474	28	0.133 022	38	0.114 325	48	0.101 793
9	0.232 237	19	0.161 123	29	0.130 730	39	0.112 860	49	0.100 755
10	0.220 663	20	0.157 099	30	0.128 552	40	0.111 449	50	0.099 747





10. Appendix 5: The International Reference Kilogram

The article EM0303 reprinted below provides background information for the introductory discussion on page HL2.1.

EM0303: The Globe and Mail, May 27, 2003, pages A1, A14

Shrinking kilogram sows mass confusion in labs

Quest for exact kilogram involves counting atoms

BY OTTO POHL
BRAUNSCHWEIG, GERMANY

The kilogram is getting lighter, scientists say, sowing potential confusion over a range of scientific endeavours.

The kilogram is defined by a platinum-iridium cylinder, cast in England in 1889. No one knows why it is shedding weight, at least in comparison with other reference weights, but the change has spurred an international search for a more stable definition.

"It's certainly not helpful to have a standard that keeps changing," said Peter Becker, a scientist at the Federal Standards Laboratory in Braunschweig, an institution of 1,500 scientists dedicated entirely to improving the ability to measure things precisely.

The kilogram's apparent loss of 50 micrograms (less than the weight of a grain of salt) is enough to distort scientific calculations.

Mr. Becker is leading a team of international researchers seeking to redefine the kilogram as a number of atoms of a selected element. Other scientists are developing a competing technique to define the weight using a complex mechanism known as the watt balance.

The final recommendation will be made by the International Committee on Weights and Measures, created by international treaty in 1875. The agency keeps the international reference kilogram in a heavily guarded safe in a château outside Paris.

It is visited once a year, under strict security, by the only three people to have keys to the safe.

..... creating reliable measurements is such painstaking work.

It is visited once a year, under strict security, by the only three people to have keys to the safe. The weight change has been noted on the occasions it has been removed for measurement.

The race is well under way to determine a new standard, although at a measured pace, since creating reliable measurements is such painstaking work.

The kilogram is the only one of the seven base units of measurement that still retains its 19th-century definition. Over the years, scientists have redefined units such as the metre (first based on the earth's circumference) and the second (conceived as a fraction of a day). The metre is now the distance light travels in 1/299,792,458th of a second, and a second is the time it takes for a cesium atom to vibrate 9,192,631,770 times. Each can be measured with remarkable precision, and, equally important, can be reproduced anywhere. The kilogram was conceived as the mass of a litre of water, but measuring that accurately proved very difficult. Instead, an English goldsmith was hired to make a platinum-iridium cylinder that would be used to define the kilogram.

A total of 80 copies of the reference kilogram have been created and distributed to signatories of the metric treaty. The sometime colourful history of these small metal cylinders underscores how long the world has used the same definition of the kilogram.

Some of the metal plugs were issued to countries that later vanished, such as Serbia and the Dutch East Indies. The Japanese had to surrender theirs after the Second World War. Germany has acquired several,

(*continued overleaf*)

including the one issued to Bavaria in 1889 and the one that belonged to East Germany.

To update the kilogram, Germany is working with scientists from countries including Australia, Italy and Japan to produce a perfectly round one-kilogram silicon crystal. The idea is that by knowing exactly what atoms are in the crystal, how far apart they are and the size of the ball, the number of atoms in the ball can be calculated. That number then becomes the definition of a kilogram.

To separate the three isotopes of silicon, Mr. Becker and his team are turning to old nuclear-weapons factories from the former Soviet Union, where centrifuges once used to make highly enriched uranium are able to produce the required purity of silicon.

"We need so many nines," Mr. Becker said, and the Soviet uranium processors are one of the only places to get them. "With the Russians, we're getting about four of them" – 99.99 per cent pure silicon 28.

A test crystal has been produced, and Dr. Arnold Nicolaus, another scientist at the German standards laboratory, is responsible for measuring whether it is perfectly round. He has measured the crystal in a half- million places to determine its shape.

It's probably the roundest item ever made by hand. "If the earth were this round, Mount Everest would be four metres tall," Dr. Nicolaus said.

An intriguing characteristic of this smooth ball is that there is no way to tell whether it is spinning or at rest. Only if a grain of dust lands on the surface is there something for the eye to track.

But scientists from the United States, England, France and Switzerland say the challenge of calculating the exact number of atoms in a silicon crystal is too imprecise with today's technology, so they are refining a technique to calculate the kilogram using voltage.

"Measuring energy is easier than counting atoms," said Dr. Richard Steiner, a scientist at the National Institute of Standards and Technology in Gaithersburg, Md.

New York Times Service

2010-06-20