Figure 2.13. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Calculating (approximately) a Sample Size

An important consideration in the Plan of any sample survey is the *sample size*, because it determines key features of a survey such as its *cost* and the *precision* of the estimator(s) that lead to answers about respondent population attribute(s) of interest; precision of estimators is a major factor in determining the severity of limitations imposed on answers by sample error. Two approaches to choosing the sample size for a sample survey are:

- ∗ A specification, in the Plan for the survey (*i.e.*, in the Design stage of the FDEAC cycle), of the *desired precision* of the estimator(s); a statement of precision is essentially a specification of the value of the standard deviation of the sample average and, and, because $s.d.(\bar{Y})$ is given by equation (2.13.1) at the right, if values of $\mathbf{N}$ and $\mathbf{S}$ are available, n can be calculated.

$$s.d.(\bar{Y}) = \mathbf{S}\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} \qquad \text{-----(2.13.1)}$$

  - – In the usual situation where *several* population attributes are being estimated in the *same* survey, the *actual* sample size is determined by the attribute which yields the *largest* calculated value of n, as determined by the magnitude of the variation of each relevant response over the elements (or units) of the respondent population and the precision specified for the estimator of its attribute.

- ∗ A specification, in the Plan for the survey (*i.e.*, in the Design stage of the FDEAC cycle), of the of the *total cost* of the survey; the sample size can then be calculated using an expression like equation (2.13.2).

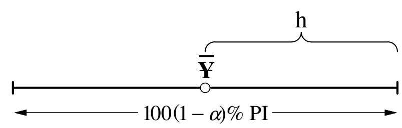
$$n = \frac{\text{total budget} - \text{fixed costs}}{\text{cost per unit surveyed}} \qquad \text{-----(2.13.2)}$$

Only the *first* of these two approaches to calculating sample size [under equiprobable selecting (EPS)], ispursued in this Figure 2.13, which deals successively with the expressions for n when estimating $\bar{\mathbf{Y}}$, ${}_T\mathbf{Y}$ and $\mathbf{P}$. Note that *all* the expressions developed in this Figure 2.13 are based on an assumption of *complete* response; if there will be *non*-response in the sample survey, the number of units (or elements) to be selected must be *in*creased appropriately from the calculated value of n.

1. Precision Specifed as a Probability Interval Half-width

Averages: To derive an expression for n under EPS, we consider first a specification of precision as a statement about the *half-width* (h, say) of a probability interval; the probability *value* of this interval should also be given but, if it is not, the default would usually be taken as 95%. [A probability interval is not to be confused with a regression *prediction* interval, often also abbreviated PI – see page 13.44 in Figure 13.7 of the STAT 221 Course Materials.]

In the diagram at the right, because the probability interval for $\bar{Y}$ under EPS is $\bar{\mathbf{Y}} \pm z_\alpha^* \times s.d.(\bar{Y})$, if h represents the value specified for the desired half-width of a $100(1-\alpha)\%$ probability interval, we must have:



$$h = z_\alpha^* \times s.d.(\bar{Y}) \qquad [e.g., z_\alpha^* = 1.95996 \simeq 1.96 \text{ for a 95\% probability interval } (\alpha = 0.05)];$$

$$\therefore\ [s.d.(\bar{Y})]^2 = \frac{h^2}{(z_\alpha^*)^2} \quad \text{so that:} \quad \mathbf{S}^2\left(\frac{1}{n} - \frac{1}{\mathbf{N}}\right) = \frac{\mathbf{S}^2}{n} - \frac{\mathbf{S}^2}{\mathbf{N}} = \frac{h^2}{(z_\alpha^*)^2}; \quad \text{hence:} \quad \frac{\mathbf{S}^2}{n} = \frac{\mathbf{S}^2}{\mathbf{N}} + \frac{h^2}{(z_\alpha^*)^2}. \qquad \text{-----(2.13.3)}$$

$$\therefore\ \frac{1}{n} = \frac{1}{\mathbf{N}} + \frac{h^2}{(z_\alpha^*)^2\mathbf{S}^2} = \frac{1}{\mathbf{N}}\left[1 + \mathbf{N}\left(\frac{h}{z_\alpha^*\mathbf{S}}\right)^2\right] \quad \text{so that:} \quad n = \mathbf{N}\left[1 + \mathbf{N}\left(\frac{h}{z_\alpha^*\mathbf{S}}\right)^2\right]^{-1}. \qquad \text{-----(2.13.4)}$$

$$\text{Alternatively, let } D = \frac{h}{z_\alpha^*} \quad \text{so that:} \quad \frac{1}{n} = \frac{D^2}{\mathbf{S}^2} + \frac{1}{\mathbf{N}} = \frac{D^2}{\mathbf{S}^2}\left[1 + \frac{1}{\mathbf{N}}\frac{\mathbf{S}^2}{D^2}\right] \quad \text{and:} \quad n = \frac{\mathbf{S}^2}{D^2}\left[1 + \frac{1}{\mathbf{N}}\frac{\mathbf{S}^2}{D^2}\right]^{-1}. \qquad \text{-----(2.13.5)}$$

thus, to a *first* approximation, the required sample size is:

$$n_0 = \frac{\mathbf{S}^2}{D^2} = \left(\frac{\mathbf{S}}{D}\right)^2 \quad \text{where:} \quad D = \frac{h}{z_\alpha^*}\ [= s.d.(\bar{Y})]; \qquad \text{-----(2.13.6)}$$

i.e., when estimating $\bar{\mathbf{Y}}$, the sample size must be roughly the square of the ratio by which the standard deviation of the sample average is to be reduced in comparison to the study (or respondent) population standard deviation.

The expression for n_0 gives a value that is too *large* because the term which is neglected, $[\ldots]^{-1}$, is *less* than 1, but it will be a reasonable approximate choice of sample size unless the provisional sampling fraction, $n_0/\mathbf{N}$, is substantial; with*out* approximation (*e.g.*, when $n_0/\mathbf{N}$ *is* substantial), we take:

$$n = n_0\left[1 + \frac{n_0}{\mathbf{N}}\right]^{-1}. \qquad \text{-----(2.13.7)}$$

Totals: When estimating ${}_T\mathbf{Y}$, we must have:

$$h = z_\alpha^* \times s.d.(\mathbf{N}\bar{Y}) = z_\alpha^* \times \mathbf{N} \times s.d.(\bar{Y}) = z_\alpha^* \times \mathbf{N}\mathbf{S}\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}};$$

reasoning as for the case of *averages* leads to a first approximation and an 'exact' result for the choice of sample size as:

$$n_0 = \frac{\mathbf{N}^2\mathbf{S}^2}{D^2} \quad \text{and} \quad n = n_0\left[1 + \frac{n_0}{\mathbf{N}}\right]^{-1}, \quad \text{where:} \quad D = \frac{h}{z_\alpha^*}. \qquad \text{-----(2.13.8)}$$

Proportions: When estimating $\mathbf{P}$, we must have:

$$h = z_\alpha^* \times s.d.(P) = z_\alpha^* \times \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}\left(\frac{1}{n} - \frac{1}{\mathbf{N}}\right)}; \qquad (\mathbf{Q} = 1 - \mathbf{P})$$

reasoning as for the case of *averages* leads to a first approximation and an 'exact' result for the choice of sample size as:

$$n_0 = \frac{\mathbf{P}\mathbf{Q}}{D^2} \quad \text{and} \quad n = n_0\left[1 + \frac{n_0 - 1}{\mathbf{N}}\right]^{-1}, \quad \text{where:} \quad D = \frac{h}{z_\alpha^*}. \qquad \text{-----(2.13.9)}$$

Although the three pairs of expressions in equations (2.13.6) to (2.13.9) above for n_0 and n are shown as being *exact*, they

(*continued overleaf*)

are based on the Central Limit Theorem *approximation* for the assumed normality of the random variables $\overline{Y}$ and P representing the average or proportion based on n units obtained by equiprobable selecting. However, more immediate difficulties in using these expressions *in practice* are that, when estimating $\overline{\mathbf{Y}}$ or ${}_T\mathbf{Y}$, knowledge of the value of $\mathbf{S}$ is needed and, when estimating $\mathbf{P}$, its value miust (already) be known. Ways of trying to deal with these difficulties are as follows:

- ∗ a suitable value for $\mathbf{S}$ or $\mathbf{P}$ may be available from a *previous survey*;
- ∗ it may be possible to estimate $\mathbf{S}$ or $\mathbf{P}$ from a *pilot survey* or *preliminary sample*;
- ∗ when estimating $\mathbf{P}$, a 'worst-case' calculation of sample size using $\mathbf{P} = \mathbf{Q} = ½$ may be acceptable;
- ∗ there may be cases where the approximate *range* of the values of the response is known and a value for $\mathbf{S}$ can be taken as *range*/k, where k would typically be taken as 4 or 5 or 6;
- ∗ it may occasionally be possible to use *probabilistic* considerations – for example, for a Poisson process, the standard deviation is the square root of the mean.

Three other matters are:

- – the first *three* of these five strategies are generally the more useful ones;
- – the requirement for knowledge in the Plan, at the *design* stage of the FDEAC cycle and *prior to* executing the sample survey, may make it difficult to choose a realistic value of n.
- – if we set $\mathbf{P} = \mathbf{Q} = ½$ and put $z_\alpha^* = 2$ (corresponding to a 95.5% or about a 95½% probability value), the expression (2.13.8) overleaf at the bottom of page 2.93 for n_0 when estimating a proportion becomes $n_0 = ½\times½/h^2/2^2 = 1/h^2$; values of n_0 for different values of h from this expression are shown in Table 2.13.1 at the right. Each value of n_0 represents a readily-calculated first approximation to the *maximum* (*i.e.*, the 'worst-case') sample size needed (assuming complete response) to estimate a proportion to within 100h percentage points with about 95½% probability. Except for a change to 95½% (rather than 95%) probability, Table 2.13.1 is based on essentially the *same* ideas as the information given in Note 11 and Table 2.10.3 at the upper right of page 2.85 in Figure 2.10.

Table 2.13.1

h	n_0	h	n_0
0.010	10,000	0.040	635
0.015	4,444	0.045	494
0.020	2,500	0.050	400
0.025	1,600	0.060	278
0.030	1,111	0.075	178
0.035	816	0.100	100

Example 2.13.1: Suppose it is known from previous investigations that the standard deviation of the weight gain over the period 0-4 weeks of life for a certain breed of chick is about 6 grams. For a new diet, it is desired to estimate ${}_T\mathbf{Y}$, the *total* weight gain over the period 0-4 weeks, for a population of 1,000 of the chicks.

(a) Find the number of chicks that should be selected equiprobably to be able to estimate ${}_T\mathbf{Y}$ to within 1 kg with 95% probability.

(b) What *assumption* about the effect of the new diet underlies this calculation of sample size?

Solution: (a) This question involves calculating a sample size when estimating ${}_T\mathbf{Y}$ with $\mathbf{S} = 6$ grams.

We have: $\mathbf{N} = 1{,}000$, $\mathbf{S} = 6$ grams; also, we set:

h = 1 kg = 1,000 grams, $z_\alpha^* = 1.95996$ for 95% probability, so that: $D^2 = \dfrac{h^2}{(z_\alpha^*)^2} = \dfrac{1{,}000^2}{1.95996^2} \simeq 260\,319.$

Hence, when estimating ${}_T\mathbf{Y}$: $n_0 = \dfrac{\mathbf{N}^2\mathbf{S}^2}{D^2} = \dfrac{1{,}000^2\times6^2}{260{,}319} = 138.29 \simeq 138;$

$$n = n_0\Big[1 + \frac{n_0}{\mathbf{N}}\Big]^{-1} = \frac{138.29}{1 + 138.29/1{,}000} = 121.49 \simeq 122;$$

i.e., a sample size of around 122 chicks is needed to estimate ${}_T\mathbf{Y}$ to within 1 kg

(b) The *assumption* is that the new diet does not change the s.d. of the chicks' weight gain over the 0-4 week period.

Example 2.13.2: An investigation is to be undertaken to estimate $\overline{\mathbf{Y}}$, the *average* amount of a population of 1,000 accounts; there has been no previous survey of the accounts but a person with considerable work experience in the area knows that their amounts lie generally within a \$100 range. Find the number of accounts that should be obtained by EPS to estimate $\overline{\mathbf{Y}}$ to within \$3 with 90% probability.

Solution: This question involves calculating a sample size when estimating $\overline{\mathbf{Y}}$; no value is given directly for $\mathbf{S}$ but we can obtain a *rough* value using the usual *range* of account values (\$100) given in the statement of the question and the *fourth* of the five suggestions (∗) listed above.

We have: $\mathbf{N} = 1{,}000$, $\mathbf{S} \simeq \frac{100}{5} = \$20;$

also, we set: h = \$3, $z_\alpha^* = 1.64485$ for 90% probability, so that: $D^2 = \dfrac{h^2}{(z_\alpha^*)^2} = \dfrac{3^2}{1.64485^2} = 3.326\,518.$

Figure 2.13. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Calculating (approximately) a Sample Size (continued 1)

Solution: (cont.) Hence, when estimating $\bar{\mathbf{Y}}$: $\quad n_0 = \dfrac{\mathbf{S}^2}{D^2} = \dfrac{20^2}{3.326\,518} = 120.25 \simeq 120;$

$$n = n_0[1 + \tfrac{n_0}{\mathbf{N}}]^{-1} = \frac{120.25}{1 + 120.25/1{,}000} = 107.34 \simeq 107;$$

i.e., a sample size of around 107 accounts is needed to estimate $\bar{\mathbf{Y}}$ to within \$3 with 90% probability.

NOTE: 1. The crude method used to obtain a value for $\mathbf{S}$ in Example 2.13.2 means that the resulting sample size of around 107 is only a *rough* guide-line; we could also have used:

$\mathbf{S} \simeq \frac{100}{4} = \$25,$ leading to: $n_0 \simeq 188,$ $n \simeq 158.$

OR: $\mathbf{S} \simeq \frac{100}{6} = \$16^2\!/_3,$ leading to: $n_0 \simeq 84,$ $n \simeq 77.$

Thus, it might be prudent in circumstances like those of Example 2.13.2 to carry out a *pilot* survey with perhaps 10 or 20 accounts (*i.e.*, 10-20% of the tentative *overall* sample size) to obtain a better value for $\mathbf{S}$ as the basis for calculating n.

Example 2.13.3: In a university with a total of 20,000 students, a sample survey is to be carried out to estimate $\mathbf{P}$, the proportion of students who find that the available athletic facilities are adequate.

(a) Find the number of students that should be selected equiprobably to be able to estimate $\mathbf{P}$ to within 5 percentage points with 99% probability.

(b) What *assumptions* underlie this calculation of sample size?

Solution: (a) This question involves calculating a sample size when estimating $\mathbf{P}$.

We have: $\mathbf{N} = 20{,}000;$ with no information bout a value for $\mathbf{P}$, we take $\mathbf{P} = \mathbf{Q} = ½;$ also, we set:

h = 5 percentage points = 0.05 and $z^*_\alpha = 2.57583$ for 99% probability, so that: $D^2 = \dfrac{h^2}{(z^*_\alpha)^2} = \dfrac{0.05^2}{2.57583^2} = 0.000\,376\,795\,42.$

Hence, when estimating $\mathbf{P}$: $\quad n_0 = \dfrac{\mathbf{PQ}}{D^2} = \dfrac{0.5\times 0.5}{0.000\,376\,795} = 663.49 \simeq 664$

$$n = n_0[1 + \tfrac{n_0 - 1}{\mathbf{N}}]^{-1} = \frac{663.49}{1 + 662.49/20{,}000} = 642.22 \simeq 642$$

i.e., a sample size of around 640 students is needed to estimate $\bar{\mathbf{Y}}$ to within 5 percentage points with 99% probability.

(b) Two *assumption* are:

- $\mathbf{P} = \mathbf{Q} = ½$, a 'worst case' assumption when there is no information about the value (anticipated) for $\mathbf{P}$;
- respondents give *truthfull* answers – there may be a endency for respondents to express *dis*satisfaction with the athletic facilities on the basis of unrealistic expectations of what is (ideally) desirable.

NOTE: 2. In place of the 'worst case' $\mathbf{P} = \mathbf{Q} = ½$, sample size calculations for two other $\mathbf{P}$ vslues are:

$\mathbf{P} = 0.3,\ \mathbf{Q} = 0.7$ leading to: $n_0 \simeq 557,$ $n \simeq 542.$

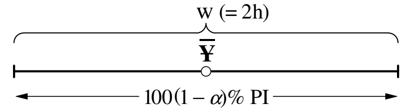
OR: $\mathbf{P} = 0.1,\ \mathbf{Q} = 0.9$ leading to: $n_0 \simeq 239,$ $n \simeq 236.$

Thus, as in Example 2.13.2, it might be prudent to carry out a *pilot* survey with perhaps 50 students {*i.e.*, 10% of the 'worst case' sample size) to obtain a better value for $\mathbf{P}$ as the basis for calculating n.

2. Two Other Specifications of Precision

The results developed on the first side (page 2.93) of this Figure 2.13 for calculating a sample size when estimating $\bar{\mathbf{Y}}$, ${}_T\mathbf{Y}$ or $\mathbf{P}$ under equiprobable selecting can also be used if the question is posed with either of two minor variations:

* the specification of precision is the (*full*) width (w) [rather than the *half*-width, h] of a probability interval – see the diagram at the right; the formulae on page 2.93 in equations (2.13.6) to (2.13.9) can be used in this situation by replacing h by w/2 when calculating D.



* the specification of precision is the *standard deviation* of the estimator rather than a statement about a probability interval; the previous formulae can be used here if *D* is taken as the specified standard deviation *instead* of h/z^*_α.

Example 2.13.4: Suppose it is known from previous investigations that the standard deviation of the weight gain over the period 0-4 weeks of life for a certain breed of chick is about 6 grams. For a new diet, it is desired to estimate ${}_T\mathbf{Y}$, the *total* weight gain over the period 0-4 weeks, for a study population of 1,000 of the chicks. Find the number of chicks that should be selected equiprobably to be able to estimate ${}_T\mathbf{Y}$ with a *standard deviation* of 1 kg.

(*continued overleaf*)

Solution: This question has the *same* context as Example 2.13.1 on page 2.98 – we have to calculate a sample size when estimating ${}_{T}\mathbf{Y}$; the *difference* from Example 2.13.1 is that the specification of precision is in terms of the *standard deviation* of the estimator, *not* the half-width of a probability interval. We proceed as in the solution of Example 2.13.1, except that the calculations are somewhat *simpler* because we are *given* the value of D, rather than having to calculate it from h and z_{α}^{*}.

We have: $\mathbf{N} = 1{,}000$, $\mathbf{S} = 6$ grams, and we set D = 1 kg = 1,000 grams.

Hence, when estimating ${}_{T}\mathbf{Y}$: $$n_0 = \frac{\mathbf{N}^2\mathbf{S}^2}{D^2} = \frac{1{,}000^2 \times 6^2}{1{,}000^2} = 36;$$

$$n = n_0[1 + \tfrac{n_0}{\mathbf{N}}]^{-1} = \frac{36}{1 + 36/1{,}000} = 34.75 \simeq 35;$$

i.e., a sample size of around 35 chicks is needed to estimate ${}_{T}\mathbf{Y}$ with a standard deviation of 1 kg.

NOTE: 2. The calculated sample size is much *smaller* in Example 2.13.4 than in Example 2.13.1, reflecting a specification of 1 kg being *less stringent* as a standard deviation than as a probability interval half-width. In fact, it is about one *half* as stringent (Why?), resulting in a sample size roughly one *quarter* (*i.e.*, one half *squared*) as large in Example 2.13.4. (Why?)

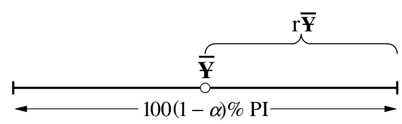
3. Relative Precision

The discussion so far in this Figure 2.13 is concerned with choosing a sample size on the basis of an *absolute* specification of precision; *i.e.*, the specification makes no reference to *the magnitude of the study population attribute being estimated.* An alternative is a *relative* specification of precision, which *does* involve the magnitude of the attribute.

NOTE: 3. The idea of *relative* (as compared to *absolute*) magnitude is illustrated by comparisons made throughout the articles about 1990 and 1995 Canadian income data from the 1996 census, reprinted in Figure 2.2f of the Course Materials. For instance, in the third paragraph of the middle column of the first side (page 2.31) of that Figure, it is stated that: *While average earnings fell in all education categories, the steepest decline – 8.2 per cent – was for people with less than a Grade 9 education*. The 8.2 per cent means that the decline in the relevant average income was an amount that was *8.2 one-hundredths of the magnitude of that average* – we would need to know the *value* of that average before we could find the *absolute* amount of the decline.

Averages: When estimating $\bar{\mathbf{Y}}$, a specification of *relative* precision starts algebraically as shown in equation (2.13.10) at the right and diagramatically below it, where some multiple (r) of $\bar{\mathbf{Y}}$ is now the desired half-width of a $100(1-\alpha)\%$ probability interval. Typically r would be a number appreciably smaller than 1; for instance, r would be 0.05 to estimate $\bar{\mathbf{Y}}$ to within 5% of its magnitude.

$$r\bar{\mathbf{Y}} = z_{\alpha}^{*} \times s.d.(\bar{Y}) \qquad \text{-----(2.13.10)}$$



Reasoning as for the case of the corresponding *absolute* specification given on the first side (page 2.93) of this Figure 2.13 leads to the first approximation and the 'exact' result for the calculated sample size at the right in equation (2.13.11).

$$n_0 = \frac{\mathbf{S}^2/\bar{\mathbf{Y}}^2}{r^2/(z_{\alpha}^{*})^2} = \left(\frac{\mathbf{S}/\bar{\mathbf{Y}}}{r/z_{\alpha}^{*}}\right)^2 \quad \text{and} \quad n = n_0[1 + \tfrac{n_0}{\mathbf{N}}]^{-1} \qquad \text{-----(2.13.11)}$$

Totals: When estimating ${}_{T}\mathbf{Y}$, we must have: $$r_{T}\mathbf{Y} = z_{\alpha}^{*} \times s.d.(\mathbf{N}\bar{Y}) \quad \text{or:} \quad r\mathbf{N}\bar{\mathbf{Y}} = z_{\alpha}^{*} \times \mathbf{N} \times s.d.(\bar{Y}); \qquad \text{-----(2.13.12)}$$
after cancelling the $\mathbf{N}$ from the left- and right-hand sides of equation (2.13.12), it becomes the *same* as equation (2.13.10), the starting point when estimating $\bar{\mathbf{Y}}$,. Thus, for a specification of *relative* precision, the expressions for n_0 and for n are the same when estimating *either* $\bar{\mathbf{Y}}$ or ${}_{T}\mathbf{Y}$.

Proportions: When estimating $\mathbf{P}$, we must have: reasoning as for the case of *averages* leads to a first approximation and an 'exact' result for the choice of sample size as in equation (2.13.13) at the right.

$$r\mathbf{P} = z_{\alpha}^{*} \times s.d.(P) = z_{\alpha}^{*} \times \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}(\tfrac{1}{n} - \tfrac{1}{\mathbf{N}})};$$

$$n_0 = \frac{\mathbf{Q}/\mathbf{P}}{(r/z_{\alpha}^{*})^2} \quad \text{and} \quad n = n_0[1 + \tfrac{n_0 - 1}{\mathbf{N}}]^{-1} \qquad \text{-----(2.13.13)}$$

Table 2.13.2 at the top of the facing page 2.101 summarizes the results derived in this Figure 2.13 for calculating a sample size to meet a specification of precision of an estimator expressed in terms of the half-width required for a probability interval; the population attribute in the first column, and the population referred to in the numerators of the six first-approximation expressions in the second and fourth columns, denote the respondent population or the study population, depending on context. Recall that a *coefficient of variation (c.v.) is the quotient of a standard deviation and the corresponding average or mean.*

(*continued*)

Figure 2.13. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Calculating (approximately) a Sample Size (continued 2)

Table 2.13.2

Population Attribute	Absolute Specification of Precision: First Approximation (n_0)	Absolute: 'Exact' Result (n)	Relative Specification of Precision: First Approximation (n_0)	Relative: 'Exact' Result (n)
$\overline{\mathbf{Y}}$	$\left(\dfrac{\text{s.d. of study population response}}{\textit{s.d.}\text{ of sample average}}\right)^2$	$n_0\left[1+\dfrac{n_0}{\mathbf{N}}\right]^{-1}$	$\left(\dfrac{\text{c.v. of study population response}}{\textit{c.v.}\text{ of sample average}}\right)^2$	$n_0\left[1+\dfrac{n_0}{\mathbf{N}}\right]^{-1}$
${}_T\mathbf{Y}$	$\mathbf{N}^2\left(\dfrac{\text{s.d. of study population response}}{\mathbf{N}\times\textit{s.d.}\text{ of sample average}}\right)^2$	$n_0\left[1+\dfrac{n_0}{\mathbf{N}}\right]^{-1}$	$\left(\dfrac{\text{c.v. of study population response}}{\textit{c.v.}\text{ of sample average}}\right)^2$	$n_0\left[1+\dfrac{n_0}{\mathbf{N}}\right]^{-1}$
$\mathbf{P}$	$\left(\dfrac{\text{s.d. of study population proportion}}{\textit{s.d.}\text{ of sample proportion}}\right)^2$	$n_0\left[1+\dfrac{n_0-1}{\mathbf{N}}\right]^{-1}$	$\left(\dfrac{\text{c.v. of study population proportion}}{\textit{c.v.}\text{ of sample proportion}}\right)^2$	$n_0\left[1+\dfrac{n_0-1}{\mathbf{N}}\right]^{-1}$

We see that the main difference between choice of sample size under absolute and relative specifications of precision is that the former entails knowledge of *standard deviations* whereas the latter entails knowledge of *coefficients of variation*. It is *not* the case that one of these requirements is *always* the more difficult to meet – in some situations, information on a standard deviation *by itself* may be more accessible whereas, in other situations, it may be easier to acquire a value for the size of a standard deviation *in relation to* the size of the corresponding average.

There are two other matters of interest in the case of *proportions*:

1. Units of *percentage points* reflect absolute precision, units of *percent* reflect relative precision.
2. The formulae for n_0 for absolute and relative specifications of precision have substantially different implications in the values they yield; as illustrated in the following Table 2.13.3, a requirement for *high relative* precision when estimating a *small* value of $\mathbf{P}$ may lead to (unrealistically) large values for the calculated sample size.

Table 2.13.3 For $\mathbf{P}$ =	0.02	0.05	0.10	0.30	0.50	0.70	0.90	0.95	0.98
(absolute) $\mathbf{P}$ estimated to within 5 percentage points: n_0 =	30	73	138	323	384	323	138	73	30
(relative) $\mathbf{P}$ estimated to within 5 percent: n_0 =	75,292	29,195	13,829	3,585	1,537	659	171	81	31 ;

the calculations for the table are based on setting $h = r = 0.05$ and $z^*_\alpha = 1.95996$ so that: $D^2 = h^2/(z^*_\alpha)^2 = 0.000\,6508 = r^2/(z^*_\alpha)^2$.

Example 2.13.5: Suppose it is known from previous studies of a certain breed of chick that, over the period 0-4 weeks of of life, the weight gain has an average of about 65 grams and a standard deviation of about 6 grams. For a new diet, it is required to estimate ${}_T\mathbf{Y}$, the *total* weight gain over the period 0-4 weeks, for a population of 1,000 of the chicks.

(a) Find the number of chicks that should be selected equiprobably to be able to estimate ${}_T\mathbf{Y}$ to within 2% of its value with 95% probability.

(b) What *assumption* about the effect of the new diet underlies this calculation of sample size?

Solution: (a) This question has the *same* context as Examples 2.13.1 on page 2.98 and 2.1.4 on pages 2.99 amd 2.100 – we have to calculate a sample size when estimating ${}_T\mathbf{Y}$; the *difference* from Example 2.13.1 is that the specification of precision is *relative*, *not* absolute. We proceed as in the solution of Example 2.13.1, except that the calculations involve the additional information in the statement of the question about the *average* weight gain on the old diet.

We have: $\mathbf{N} = 1{,}000$, $\overline{\mathbf{Y}} = 65$ g, $\mathbf{S} = 6$ grams;
also, we set: $r = 2\% = 0.02$, $z^*_\alpha = 1.95996$ for 95% probability.

Hence, when estimating ${}_T\mathbf{Y}$: $$n_0 = \left(\frac{\mathbf{S}/\overline{\mathbf{Y}}}{r/z^*_\alpha}\right)^2 = \left(\frac{6/65}{0.02/1.95996}\right)^2 = 81.83 \simeq 82;$$

$$n = n_0\left[1+\frac{n_0}{\mathbf{N}}\right]^{-1} = \frac{81.83}{1+81.83/1{,}000} = 75.64 \simeq 76;$$

i.e., a sample size of about 76 chicks is needed to estimate ${}_T\mathbf{Y}$ to within 2% of its value with 95% probability.

(b) One assumption underlying the calculation of sample size in (a) is that the coefficient of variation of weight gain is the *same* under the new diet as under the existing diet; if this assumption is *not* met, the calculated sample size will be inaccurate.

It is *un*likely that the standard deviation and the coefficient of variation of weight gain are *equally* sensitive to change in diet, so the assumptions in the (b) parts of Examples 2.13.1 and 2.13.5 are most likely *not* equivalent; this reminds us that *extra*-statistical knowledge may be needed to assess a modelling assumption.

(*continued overleaf*)

NOTE: 4. The value (82) of n_0 in Example 2.13.5 is somewhat *smaller* than that (138) in Example 2.13.1; this reflects the somewhat *less* stringent specification of precision in the former – an approximately *9*-fold reduction in the coefficient of variation – than in the latter – nearly a *12*-fold reduction in the standard deviation.

Example 2.13.6: In a university with a total of 20,000 students, a survey is to be carried out to estimate $\mathbf{P}$, the proportion of students who find that the available athletic facilities are adequate.

(a) Find the number of students that should be selected equiprobably to be able to estimate $\mathbf{P}$ to within 5 percentage points with 95% probability.

(b) Find the number of students that should be obtained by EPS to estimate $\mathbf{P}$ to within 5 percent with 95% probability.

(c) What *assumptions* underlie these calculations of sample size?

Solution: (a) This question has the *same* context as Example 2.13.3 on page 2.99 – we have to calculate a sample size with an *absolute* specification of precision when estimating $\mathbf{P}$; we are given no information about an anticipated magnitude for $\mathbf{P}$ and so we have no value for the population standard deviation. We are therefore in the position of having to do a 'worst case' calculation using $\mathbf{P} = \mathbf{Q} = \frac{1}{2}$.

We have: $\mathbf{N} = 20{,}000$, $\mathbf{P} = \mathbf{Q} = \frac{1}{2}$.;

also, we set: $h = 5\% = 0.05$, $z^*_\alpha = 1.95996$ for 95% probability,

so that: $$D^2 = \frac{h^2}{(z^*_\alpha)^2} = \frac{0.05^2}{1.95996^2} = 0.00\,650\,797\,075.$$

Hence, when estimating $\mathbf{P}$: $$n_0 = \frac{\mathbf{P}\mathbf{Q}}{D^2} = \frac{\frac{1}{2}\times\frac{1}{2}}{0.000\,650\,797} = 384.14 \simeq 385;$$

$$n = n_0\Big[1 + \frac{n_0 - 1}{\mathbf{N}}\Big]^{-1} = \frac{384.14}{1 + 383.14/20{,}000} = 376.92 \simeq 377;$$

i.e., at *worst*, a sample size of around 380 students is needed to estimate $\mathbf{P}$ to within 5 percentage points with 95% probability.

(b) This question involves calculating a sample size with a *relative* specification of precision when estimating $\mathbf{P}$; we are given no information about an anticipated magnitude for $\mathbf{P}$ and so, *un*like (a), we are *not* able to calculate the sample size.

(c) One assumption is that people tell the *truth* – in questions about adequacy in sample surveys like this, respondents have a tendency to ask for a level of facilities that, realistically, may not be achievable.

NOTE: 6. The reduction in the probability from 99% in Example 2.13.3 to 95% in this Example 2.13.6 results in a (substantial) reduction in the calculated sample size in part (a) from 642 to 377.

A final comment is that this Figure 2.13 is based on the *ideal* of being able to specify the precision required, or the maximum cost which can be sustained, in the Plan for a sample survey, and of then calculating an appropriate sample size; the goal is to avoid waste of resources either by investigating too *many* elements or by obtaining estimate(s) of too low a precision from investigating too *few*. In *practice*, this goal usually involves a number of *other* matters, including:

- ∗ quantifying the cost structure of sampling (*i.e.*, selecting, responding, measuring and estimating) under a given Plan;
- ∗ deciding on the relative importance of conflicting Plan requirements for a *variety* of population response variates that will usually be involved in a sample survey;
- ∗ assessing what precision would be acceptable for each response;
- ∗ finding sources for values to use for population attributes like standard deviations.

Such matters may introduce considerable complications into the (simplified) ideas presented in this Figure 2.13.

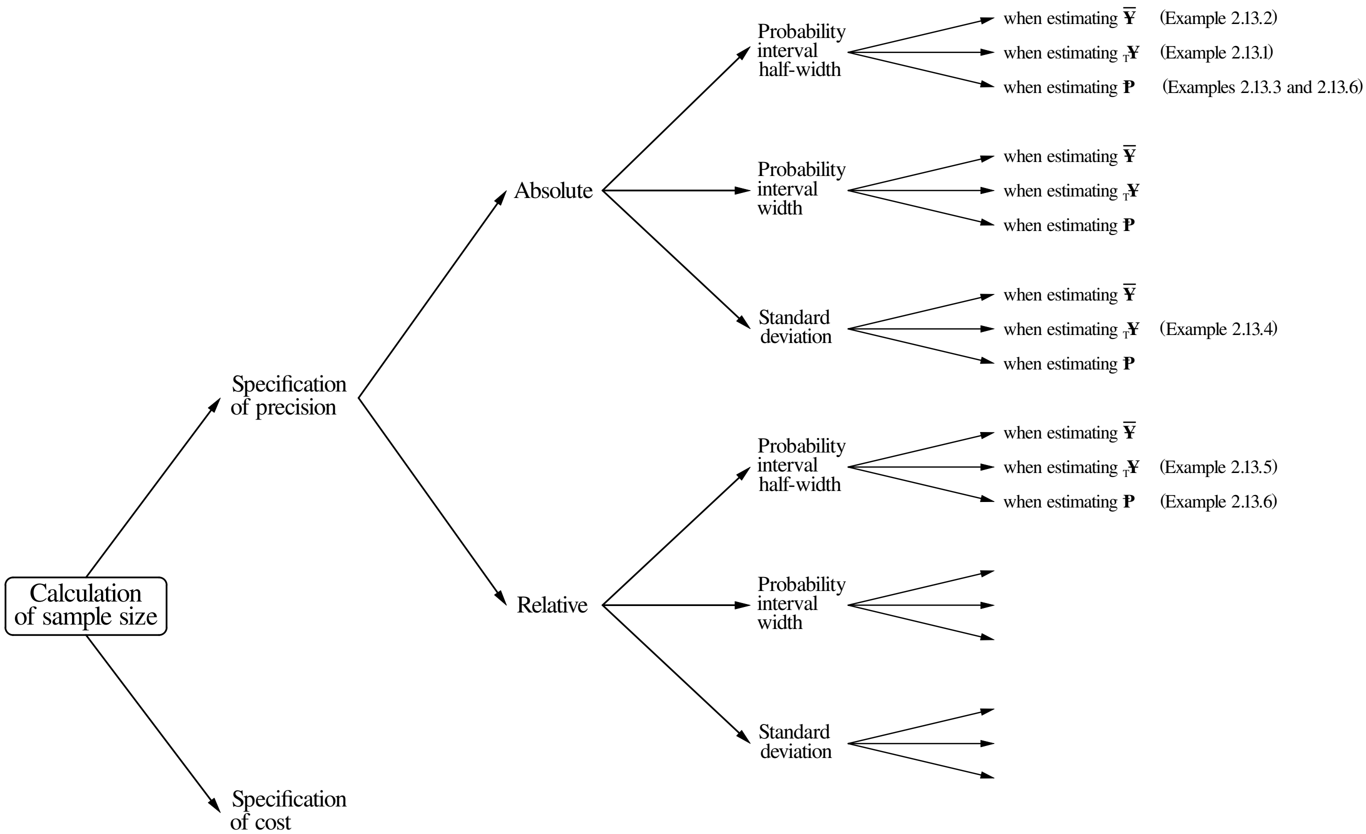
☐ Comment critically on the following statement: *The formula for the first approximation for the calculated sample size when estimating* $_T\mathbf{Y}$ *under an* absolute *specification of precision contains a term* $\mathbf{N}^2$ *in its numerator and so will often lead to* very *large values for* n_0.

(*continued*)

Figure 2.13. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Calculating (approximately) a Sample Size (continued 3)

4. Appendix: Tree Diagram Summary of Precision Specifications

The primary theme of this Figure 2.13 – using appropriate information to calculate (approximate) sample size, as one component of the Plan in the Design stage of the FDEAC cycle – is straight forward, but complications arise from the number of cases involved. A tree diagram given below provides an overview of these cases; for tree branches with blank ends, formulae for calculating sample size are not developed explicitly in this Figure 2.13 but they can be obtained readily from what *is* given.



NOTE: 5. Curiously, the emphasis in statistics seems to be on *absolute* precision – as evidenced by the ubiquitous occurrence of standard deviation – whereas, in practice, *relative* precision (or change) is often more relevant.

Blank page