Figure 8.17. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating a Proportion or a Frequency

The theory developed in Figure 8.11 is concerned with equiprobable selecting (EPS) when the response, $\mathbf{Y}$, is the *measured* value of some *quantitative* respondent population response variate, and we wish to estimate a respondent population attribute such as an average or total. In this Figure 8.17, we develop the corresponding theory for the *proportion* of respondent population elements with some *qualitative* characteristic. Examples of proportions of practical interest are:

- ⁎ the proportion (or percentage) of voters who support a given political party (for a political poll);
- ⁎ the proportion of TV viewers watching a particular program (for TV ratings);
- ⁎ the proportion of sales of an item of a specified brand (for market research);
- ⁎ the proportion of trees in a forest with trunks greater than 30 cm in diameter (for a lumber company).

With appropriate changes in notation, the theory of EPS for a proportion can be derived from the theory for an average given in Figure 8.11. We deal here with a response in *two* categories [C and ${}^{c}C$ (the complement of C), say] for the characteristic of the respondent population elements; in Figure 2.12 in STAT 332, the ideas are extended to a response in *more than two* categories.

1. Notation.

Define the indicator variate $\mathbf{V}$ such that: $\mathbf{V}_i = 1$ if respondent population element i has the characteristic C,
$\mathbf{V}_i = 0$ if respondent population element i has the characteristic ${}^{c}C$.

Table 8.17.1:

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

The response of element j obtained by EPS from the population of $\mathbf{N}$ elements can be modelled by the random variable V_j with probability function and mean as in equation (8.17.1) at the right. It follows that the random variable P is an unbiased estimator of $\mathbf{P}$ – see equation (8.17.2), the analogue of equation (8.11.8) on page 2.53 of Figure 8.11.

$$\begin{array}{c|cc} v_j & 1 & 0 \\ \hline f(v_j) & \mathbf{P} & 1-\mathbf{P} \end{array}; \qquad E(V_j) = 1\times\mathbf{P} + 0\times(1-\mathbf{P}) = \mathbf{P} \qquad \text{-----(8.17.1)}$$

$$E(P) = E(\tfrac{1}{n}\sum_{j=1}^{n} V_j) = \tfrac{1}{n}\sum_{j=1}^{n} E(V_j) = \tfrac{1}{n}\,nE(V_j) = \tfrac{1}{n}\,n\mathbf{P} = \mathbf{P} \qquad \text{-----(8.17.2)}$$

2. Estimating $\mathbf{P}$, the Respondent Population Proportion

We want both a value (or *point estimate*) for this population attribute *and* a measure of the uncertainty of the estimate, for which we use a confidence interval in STAT 220.

Using the indicator variate $\mathbf{V}$, we have shown above that a proportion becomes a special case of an average. Because it was established in equation (8.11.8) at the top of page 2.53 in Figure 8.11 that the random variable $\overline{Y}$ representing the sample *average* under EPS is an unbiased estimator of $\overline{\mathbf{Y}}$, the respondent population average, the random variable P representing the sample *proportion* under EPS is an unbiased estimator of $\mathbf{P}$, the respondent population proportion [as in equation (8.17.2) above].

To find the standard deviation of $\mathbf{P}$, we note first that because the population of $\mathbf{V}$s consists only of 1s and 0s, both population sums $\Sigma\mathbf{V}_i$ and $\Sigma\mathbf{V}_i^2$ are equal to $\mathbf{N}\mathbf{P}$, the number of elements with characteristic C in the respondent population; the corresponding *sample* sums are both equal to np (= np); we also use the notation $\mathbf{Q} = 1-\mathbf{P}$.

Hence: $$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{V}_i - \overline{\mathbf{V}})^2} = \sqrt{\frac{1}{\mathbf{N}-1}\Big[\sum_{i=1}^{\mathbf{N}}\mathbf{V}_i^2 - \mathbf{N}\overline{\mathbf{V}}^2\Big]} = \sqrt{\frac{1}{\mathbf{N}-1}[\mathbf{N}\mathbf{P} - \mathbf{N}\mathbf{P}^2]} = \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}(1-\mathbf{P})} = \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}}, \qquad \text{-----(8.17.3)}$$

so that: $$s.d.(P) = \mathbf{S}\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}(\frac{1}{n} - \frac{1}{\mathbf{N}})} \ \Big[= \sqrt{\frac{\mathbf{N}-n}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}\frac{1}{n}} = \sqrt{\frac{\mathbf{N}-n}{\mathbf{N}}\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}\frac{1}{n}}\Big] \quad \text{(= the s.d. of } P\text{).} \qquad \text{-----(8.17.4)}$$

Also: $$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(v_j - \overline{v})^2} = \sqrt{\frac{1}{n-1}\Big[\sum_{j=1}^{n}v_j^2 - n\overline{v}^2\Big]} = \sqrt{\frac{1}{n-1}[np - np^2]} = \sqrt{\frac{n}{n-1}p(1-p)} = \sqrt{\frac{n}{n-1}pq} \quad (q = 1-p), \qquad \text{-----(8.17.5)}$$

so that: $$\hat{s.d.}(P) = s\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} = \sqrt{\frac{n}{n-1}pq(\frac{1}{n} - \frac{1}{\mathbf{N}})} \ \Big[= \sqrt{\frac{\mathbf{N}-n}{\mathbf{N}}pq\frac{1}{n-1}} = \sqrt{\frac{\mathbf{N}-n}{\mathbf{N}}\frac{n}{n-1}pq\frac{1}{n}}\Big] \quad \text{(= the \textit{estimated} s.d. of } P\text{).} \qquad \text{-----(8.17.6)}$$

If $\mathbf{N} \gg$ n, equation (8.17.6) yields the approximate equation (8.17.7). $$\hat{s.d.}(P) \simeq \sqrt{\frac{pq}{n-1}} \qquad \text{-----(8.17.7)}$$

The *coefficient of variation* (*c.v.*) of P [a measure of *relative* precision] is: $$c.v.(P) \equiv \frac{s.d.(P)}{\mathbf{P}} = \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}(\frac{1}{\mathbf{P}} - 1)(\frac{1}{n} - \frac{1}{\mathbf{N}})}. \qquad \text{-----(8.17.8)}$$

If $\mathbf{N} \gg$ n, equation (8.17.8) yields the approximate equations (7.17.9). $$c.v.(P) \simeq \sqrt{\frac{1}{n}(\frac{1}{\mathbf{P}} - 1)} \ ; \quad \hat{c.v.}(P) \simeq \sqrt{\frac{1}{n}(\frac{1}{p} - 1)} \qquad \text{-----(8.17.9)}$$

(*continued overleaf*)

NOTES: 1. The rightmost two expressions in brackets [ ] in equation (8.17.4) and in equation (8.17.6) have the symbols re-arranged only to illustrate discussion in Notes 7 and 8 below and on page 8.95; the *working* expressions in these two equations are their *second* square roots on their right-hand sides.

2. Equation (8.17.4) shows that the magnitude of the standard deviation of P is determined by the values of two respondent attributes, $\mathbf{PQ}$ and $\mathbf{N}$, and n, the sample size; because typically $\mathbf{N} \gg$ n, n usually dominates $\mathbf{N}$ in *their* effect on the standard deviation of P.
   - The weak (often negligible) effect of $\mathbf{N}$ on precision of sample estimates (under EPS) is contrary to many people's intuition; *i.e.*, it seems counterintuitive that *population size* (like 30 million Canadians or 300 million Americans) usually makes little or no contribution to the limitation imposed on an Answer by sample error.
   - $\mathbf{PQ} = \mathbf{P}(1-\mathbf{P})$ is effectively, of course, just the attribute $\mathbf{P}$.

3. The expression for $s.d.(P)$ in equation (8.17.4) overleaf on page 8.93 is potentially useful in three ways:
   - to assess the precision of the estimator P [if $\mathbf{N}$, n and $\mathbf{P}$ are known];
   - to calculated the sample size needed to attain a specified precision [if $\mathbf{N}$ and $\mathbf{P}$ are known];
   - to compare the precision of P with that of *other* estimators of $\mathbf{P}$ – recall Note 4 on page 8.53 in Figure 8.11.

4. When n $= \mathbf{N}$, $s.d.(P) = 0$, reminding us that, at least in principle, sample error is eliminated in a census of the re-pondent population.
   - Of course, study error, non-response error and measurement error must still be adequately managed in a census.

5. The random variable S^2 [with value s^2 taken as (the value of) the *square* of the sample (data) standard deviation under EPS] is an *un*biased estimator of $\mathbf{S}^2$, the *square* of the respondent population (data) standard deviation $\mathbf{S}$] [as shown near the bottom of page 8.53 in Figure 8.11], so it follows that $\mathrm{n}PQ/(\mathrm{n}-1)$ is an unbiased estimator of $\mathbf{NPQ}/(\mathbf{N}-1)$.
   - This unbiasedness, a staple of survey sampling theory, is of interest statistically but is of limited relevance in the context of this Figure 8.17, because the unbiasedness is lost as a consequence of the *square roots* in the rightmost terms in equations (8.17.5) and (8.17.3) – see also Appendix 5 on pages 8.58 and 8.59 in Figure 8.11.

6. Equations (8.17.8) and (8.17.9) overleaf at the bottom of page 8.93 show that the *relative* precision *de*creases [*i.e.*, $s.d.(P)$ becomes *larger* relative to $\mathbf{P}$] as $\mathbf{P}$ *de*creases, reminding us that (very) large sample sizes are needed to estimate small proportions precisely.

7. In real-world sample surveys to estimate a proportion or frequency, the sample is selected probabilistically from a (finite respondent) population. This raises the possibility of two *probability models*, which are often presented as an urn containing balls of two colours, red and black say. [Statistically, the urn is the population with the balls as its elements, the two colours being the response variate 'values' C and ${}^{c}C$.]
   - For EPS with*out* replacement, the model is *hypergeometric*, whose probability function for the *number* of red balls in n selections (*i.e.*, in a sample of size n) is equation (8.17.10) at the right; $\mathbf{N}$ is the number of balls in the urn, $\mathbf{R}$ are red (and $\mathbf{N}-\mathbf{R}$ are black). Unsurprisingly, the mean and (probabilistic) standard deviation of the hypergeometric distribution are, respectively, n$\mathbf{P}$ and equation (8.17.4) overleaf on page 8.93 with an additional n outside the square root to deal with a frequency rather than a proportion; our $\mathbf{P}$ is (of course) $\mathbf{R}/\mathbf{N}$ and the terms multiplying $\mathbf{PQ}$ under the square root may by written $(\mathbf{N}-\mathrm{n})/(\mathbf{N}-1)$ as in the third expression in equation (8.17.4).

$$f(r) = \frac{\binom{\mathbf{R}}{r}\binom{\mathbf{N}-\mathbf{R}}{\mathrm{n}-r}}{\binom{\mathbf{N}}{\mathrm{n}}}, \qquad r = 0, 1, 2, ..., \mathrm{n} \qquad \text{-----(8.17.10)}$$

   - It is sometimes mathematically convenient to assume selecting is *with* replacement, with the appropriate probability model then being *binomial* instead of hypergeometric – selecting balls with replacement from the urn is an instance of Bernoulli trials; the probability function for the *number* of red balls in n selections (*i.e.*, in a sample of size n) is equation (8.17.11) at the right – the binomial parameter (a *Greek* letter) π is the probability of 'success', here, selecting a red ball so $\pi = \mathbf{R}/\mathbf{N}$ (our $\mathbf{P}$).

$$g(r) = \binom{\mathrm{n}}{r}\pi^{r}(1-\pi)^{\mathrm{n}-r}, \qquad r = 0, 1, 2, ..., \mathrm{n} \qquad \text{-----(8.17.11)}$$

   The mean and standard deviation of the binomial distribution are nπ and $\sqrt{\mathrm{n}\pi(1-\pi)}$ for the *number* of red balls in n selections, and π and $\sqrt{\pi(1-\pi)/\mathrm{n}}$ for the *proportion* of red balls in n selections. Comparing this second standard deviation with the third expression in equation (8.17.4) on page 8.93, we see that selecting with*out* replacement introduces the additional term $(\mathbf{N}-\mathrm{n})/(\mathbf{N}-1)$ – mathematically, we can think of this term as dealing with the (slight) *lack* of *probabilistic independence* when selecting with*out* replacement compared to selecting *with* replacement.

   Other comparisons are pursued in Note 8 below and on the facing page 8.95.

8. The order of topics in most current introductory statistics texts means that, before students can encounter ideas (involving $\mathbf{N}$) like $s.d.(\overline{Y})$ in equation (8.11.9) near the top of page 8.53 in Figure 8.11 or $s.d.(P)$ in equation (8.17.4) overleaf on page 8.93, they are presented with two (likely new) ideas.

(*continued*)

Figure 8.17. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating a Proportion or a Frequency (continued 1)

NOTES: (cont.) 8. ● For (much beloved of statisticians) n probabilistically independent identically distributed (abbreviated iid) random variables (Y_j, say) each with mean μ and standard deviation σ, the random variable $\overline{Y}$ that is their *average* has the *same* mean μ but a (smaller) standard deviation of $\sigma/\sqrt{n}$ [equation (8.17.12)].

$$s.d.(\overline{Y}) = \sigma/\sqrt{n} \equiv \sigma\sqrt{\frac{1}{n}} \quad \text{-----(8.17.12)}$$

● The standard deviation for a binomial *proportion* also has $\sqrt{n}$ in its denominator, as discussed on the facing page 8.94 in the second bullet (●) of Note 7.

However, with the accolades it tends to receive from statisticians, this useful insight that averages are *less* variable than their components by a factor of $\sqrt{n}$ can become an ideal that other relevant statistical theory should try to emulate; in the case of survey sampling theory, pursuit of this 'ideal' is detrimental to statistics and its teaching in three ways.

∗ Equations (8.3.9) and (8.3.17) near the top and bottom of page 8.53 in Figure 8.11 are written in their second (equivalent) forms shown as equations (8.17.13) and (8.17.14) at the right, as are equations (8.17.4) and (8.17.6) from page 8.93, where $f = n/\mathbf{N}$; equation (8.17.6) is even rewritten as (the inelegant) equation (8.17.15) to yield (the mandatory?) $1/n$ under the square root.

$$s.d.(\overline{Y}) = \mathbf{S}\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} \equiv \sqrt{(1-f)}\,\mathbf{S}\sqrt{\frac{1}{n}} \quad \text{----(8.17.13)}$$

$$\hat{s.d.}(\overline{Y}) = s\sqrt{\frac{1}{n} - \frac{1}{\mathbf{N}}} \equiv \sqrt{(1-f)}\,s\sqrt{\frac{1}{n}} \quad \text{----(8.17.14)}$$

$$s.d.(P) = \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}(\frac{1}{n} - \frac{1}{\mathbf{N}})} \equiv \sqrt{(1-f)\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}\frac{1}{n}} \quad \text{-----(8.17.4)}$$

$$\hat{s.d.}(P) = \sqrt{\frac{n}{n-1}pq(\frac{1}{n} - \frac{1}{\mathbf{N}})} \equiv \sqrt{(1-f)pq\frac{1}{n-1}} \quad \text{-----(8.17.6)}$$

$$= \sqrt{(1-f)\frac{n}{n-1}pq\frac{1}{n}} \quad \text{----(8.17.15)}$$

$$s.d.(\overline{Y})_{\text{EPSWIR}} = \sqrt{\frac{\mathbf{N}-1}{\mathbf{N}}}\,\mathbf{S}\sqrt{\frac{1}{n}} \quad \text{----(8.17.16)}$$

– $(1-f)$ in the five second forms makes *less* clear the relative roles of n and $\mathbf{N}$ (see Note 2 near the top of the facing page 8.94) and in equations (8.17.13) and (8.17.14) makes it appear (wrongly) that the (data) s.d.s $\mathbf{S}$ [and s] are equivalent to the (probabilistic) s.d. σ – comparing equations (8.17.12) at the right above and (8.17.16) shows that the *actual* equivalence involves a premultiplier (although its value is typically *very* close to 1) for $\mathbf{S}$ – see also equation (8.11.6) on page 8.52.

○ Equation (8.17.16) is derived from the *first* term of equation (8.11.9) near the top of page 8.53 in Figure 8.11 when the *second* covariance term is zero due to probabilistic independence under equiprobable selecting *with* replacement (EPSWIR).

○ We note that the *reciprocal* of the premultiplier of $\mathbf{S}$ in equation (8.17.16) occurs in equation (8.17.3) to convert the probabilistic (binomial proportion) s.d. $\mathbf{P}\mathbf{Q}$ to the *data* s.d. needed when selecting is with*out* replacement.

○ When $n = 1$, the RHS of equation (8.17.16) is that of equation (8.11.6) on page 8.52 of Figure 8.11.

○ When $n = 1$, equations (8.17.13) and (8.17.16) become the *same*, reminding us that selecting without and with replacement no longer differ for a sample of size one. Regrettably, this common special case of equations (8.17.13) and (8.17.16) has been made part of the (unprofitable) $n-1$ *vs* n saga – see Statistical Highlights #94 and #100 (the lower half of page HL94.9, the middle of page HL100.8).

∗ $1-f$ is called the *finite population correction* (abbreviated f.p.c.); such a name conveys the bizarre misconception that we should think of the real world as a 'correction' to a model.

– As these Materials maintain throughout, populations in statistics are real-world entities with *finite* numbers of elements and they should *not* be confused with (sometimes useful) infinite *models*.

∗ The idea of an infinite 'population' is, in part, a confusion of selecting with*out* replacement, which ends when all elements have been selected, and selecting *with* replacement which can, in principle, continue indefinitely. However, as the discussion in Note 7 on the facing page 8.94 illustrates, an urn model has a *finite* number of elements even when selecting is with replacement (and so can continue indefinitely) under a binomial model.

3. Confidence Intervals for $\mathbf{P}$, the Respondent Population Proportion

On the basis of approximate normality of the distribution of P as a consequence of the Central Limit Theorem, equation (8.17.6) developed on page 8.93 leads, reasoning as in Figure 8.11, to an interval involving random variables:

$$I = [P - z_\alpha^*\sqrt{\frac{n}{n-1}PQ(\frac{1}{n} - \frac{1}{\mathbf{N}})},\ \ P + z_\alpha^*\sqrt{\frac{n}{n-1}PQ(\frac{1}{n} - \frac{1}{\mathbf{N}})}],$$

such that $\Pr(I \ni \mathbf{P}) \simeq 100(1-\alpha)\%$, where z_α^* is the $100(1-\alpha/2)th$ percentile of the $N(0, 1)$ distribution. For calculating an approximate $100(1-\alpha)\%$ *confidence interval* for $\mathbf{P}$, we use:

$$p \pm z_\alpha^*\sqrt{\frac{n}{n-1}pq(\frac{1}{n} - \frac{1}{\mathbf{N}})} = [p - z_\alpha^*\sqrt{\frac{n}{n-1}pq(\frac{1}{n} - \frac{1}{\mathbf{N}})},\ p + z_\alpha^*\sqrt{\frac{n}{n-1}pq(\frac{1}{n} - \frac{1}{\mathbf{N}})}]. \quad \text{----(8.17.17)}$$

NOTES: 9. The theory developed in Figure 8.11 for averages and in this Figure 8.17 for proportions are similar; a *difference* is the (appropriate approximation for the) distributions of $\overline{Y}$ and P, the random variables representing, respectively, the sample average and the sample proportion under equiprobable selecting.

● As discussed in more detail in Note 12 overleaf on page 8.96, use of the normal (rather than the t_{n-1}) distribu-

NOTES: (cont.) 9. ● tion in the foregoing expressions (8.17.17) for a confidence interval for **P** is based on the normal approximation to the binomial distribution (used to model *counted* data).

– The only approximate normality of the distribution of P is one reason why the confidence interval expressions (8.17.17) are only approximate.

10. The theoretical foundation for the confidence intervals expressions (8.11.18) and (8.11.19) on page 8.54 in Figure 8.11 and hence of (8.17.17) overleaf on page 8.95, involves *explicitly* unit inclusion probabilities based on EPS; there is thus *no* basis for the (not uncommon) use of these expressions to calculate confidence intervals for samples selected by *other* (non-probability) methods (accessibility, haphazard, judgement, quota, systematic, volunteer, etc.).

11. The value of a sample attribute determines two characteristics of the approximate confidence interval for **P** – the sample proportion defines both its *centre* and its *width*; *both* characteristics (centre *and* width) of a confidence interval may be adversely affected by *compromised* selecting or measuring processes.

- Estimating the population proportion by the sample proportion in estimating the standard deviation of P is another reason why the confidence interval expressions (8.17.17) are approximate.

12. In principle, the normal approximation involved in the confidence interval expressions (8.17.17) overleaf on page 8.95 could be avoided by using the 'exact' hypergeometric model (8.17.10) discussed in Note 7 on page 8.94. Although charts and tables have been published for selected values of **N**, n and p (see Cochran, page 57, for references), which allow hypergeometric confidence limits for **P** to be found, Barnett (page 44) states that their practical utility is limited by factors such as the accuracy with which the charts can be read and often a need for complicated inverse interpolation; however, these difficulties could now likely be mitigated by suitable computer software.

- Provided that n is much smaller than both **R** and **N** − **R**, this hypergeometric distribution can be approximated by a *binomial* distribution with parameters n and **R**/**N**, denoted Bi(n, **R**/**N**). However, computational difficulties again make it inconvenient to find confidence intervals for **P** from this binomial approximation. Cochran (pp. 59, 60) gives examples involving calculations for both the hypergeometric and binomial distributions.
- For practical convenience, it is usual to take the approximation one stage further to the *normal* approximation to the binomial distribution; this approximation is shown at the right in equation (8.17.18).

$$P \doteqdot N[\mathbf{P}, \sqrt{\frac{\mathbf{N}}{\mathbf{N}-1}\mathbf{P}\mathbf{Q}(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})}] \quad \text{----(8.17.18)}$$

– The accuracy of the normal approximation will be reasonable provided that:

∗ n ≪ **R**, **N** − **R**; *i.e.*, the sample size is much smaller than the numbers of elements in the respondent population with and without the characteristic C;

∗ n**P** and n**Q** are not too small (say at least 30), because investigating a characteristic that occurs in very few or in nearly all elements of the population generally needs a very large sample size to estimate its proportion with reasonable precision. Table 8.17.2 at the lower right, taken in part from Cochran (page 58), gives working guidelines for deciding when it is reasonable to use the normal approximation; np is the number of selected units in the *smaller* of the categories defined by C and ^cC, and n is the required sample size. Cochran states that the rules by which his table is constructed ensure that, at the 95% confidence level, the true frequency with which the confidence limits fail to cover **P** is not greater than 5.5%; also, the probability that the upper limit is below **P** is between 2.5% and 3.5%, and the probability that the lower limit exceeds **P** is between 2.5% and 1.5%. The last column of Table 8.17.2 shows that, under the rules used to construct the table, the (estimated) mean of the binomial distribution, np, lies between about 5½ and 8½ (estimated) standard deviations above the lower limit (0) of the distribution.

Table 8.17.2

p	np	n	$p/\sqrt{pq/\mathrm{n}}$
0.5	15	30	5.48
0.4	20	50	5.77
0.3	24	80	5.86
0.2	40	200	7.07
0.1	60	600	8.16
0.05	70	1400	8.58

Thus, the normal approximation, in conjunction with equation (8.17.6) near the bottom of page 8.93, leads to the (approximate) confidence interval expressions (8.17.17) overleaf near the bottom of page 8.95, where **P** and **Q** are estimated using p and q taken as the sample p and q values.

– When n is near the limit where it is reasonable to use the normal approximation, a slight improvement can be affected by incorporating a **continuity correction** of 1/2n; the expression for calculating the confidence interval for **P** is then equation (8.17.19) shown at the right. Without the continuity correction, the normal approximation usually gives a confidence interval that is too narrow (Cochran, page 58).

$$p \pm z^*_\alpha[\sqrt{\frac{\mathrm{n}}{\mathrm{n}-1}pq(\frac{1}{\mathrm{n}} - \frac{1}{\mathbf{N}})} + 1/2\mathrm{n}] \quad \text{----(8.17.19)}$$

13. As indicated previously in Note 3 on page 8.94, the expression for $s.d.(P)$ in equation (8.17.4) can be used to assess precision. If n ≪ **N** so that 1/**N** is negligible compared to 1/n and in the 'worst-case' [*i.e.*, maximum $s.d.(P)$ ≡ minimum precision] situation where **P** = **Q** = ½, Table 8.17.3 at the upper right of the facing page 8.97 shows both

Figure 8.17. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating a Proportion or a Frequency (continued 2)

NOTES: (cont.) 13. the resulting standard deviations and 95% confidence interval half-widths (in %) for selected values of n.

- Numbers such as those in the third and sixth columns are often the basis for the statements used in the media to report the precision of polls. For example, for polls taken before elections in Canada, which typically have around 1,500 respondents, the 'margin of error' is commonly given as plus or minus 2½ percentage points. However, in some instances, *larger* (more conservative?) margins of error are given than those in Table 8.17.3; *e.g.*, plus or minus 4 percentage points for (a sample size of) about 1,000 respondents.

Table 8.17.3

n	$\sqrt{\frac{1}{4n}}$	$196\sqrt{\frac{1}{4n}}$	n	$\sqrt{\frac{1}{4n}}$	$196\sqrt{\frac{1}{4n}}$
50	.0707	13.9%	1,000	.0158	3.1%
100	.05	9.8	1,500	.0129	2.5
200	.0354	6.9	2,000	.0112	2.2
250	.0316	6.2	2,500	.01	2.0
300	.0289	5.7%	3,000	.0091	1.8%
400	.025	4.9	3,500	.0085	1.7
500	.0224	4.4	4,000	.0079	1.5
750	.0183	3.6	5,000	.0071	1.4

Example 8.17.1: A small university has 300 final-year undergraduates; in a poll of 100 of these students obtained by equiprobable selecting, 15 said they intended to go on to graduate school. Find an approximate 95% confidence interval for the proportion of all such students who intend to go on to graduate school.

Solution: We have: $\mathbf{N} = 300$, $n = 100$, $p = 0.15$ $(= p)$, $z^*_\alpha = 1.95996$ for 95% confidence,

so that: $\hat{s.d.}(P) = \sqrt{\frac{n}{n-1}pq(\frac{1}{n} - \frac{1}{\mathbf{N}})} = \sqrt{\frac{100}{99}(0.15 \times 0.85)(\frac{1}{100} - \frac{1}{300})} = 0.029\,301\,635.$

Hence, an approximate 95% confidence interval for $\mathbf{P}$, the respondent population proportion of all such students who intend to go on to graduate school, is:

$p \pm 1.95996 \times \hat{s.d.}(P) = 0.15 \pm 1.95996 \times 0.029\,3016 \implies (0.0926,\ 0.2074)$ or about (9, 21%).

NOTES: 14. Despite the high sampling fraction of one third, the confidence interval is relatively *wide* – *i.e.*, the small sample size of 100 means that sample error imposes appreciable limitation on the answer involving a (fairly small) respondent population proportion of (we estimate) around 15%.

15. With a sample size of 100, the continuity correction of $\pm 1/2n$ is ± 0.005; the confidence intervals without and with the correction are given in Table 8.17.4 at the right below. We see that, in this situation, the difference is unlikely to be practically important – we would typically give a final value for both intervals as about (9, 21%).

Table 8.17.4: Calculated	Confidence interval Value	Interval width	Interval width $\div p$
with*out* continuity correction	(0.0926, 0.2074)	0.1148	77%
with continuity correction	(0.0876, 0.2124)	0.1248	83%

Example 8.17.2: A sociological survey is carried out in a small town containing 621 households; 60 of the households obtained by equiprobable selecting show that 11 contain at least one person over the age of 65. Find an approximate 90% confidence interval for the proportion of households in the town that contain at least one person over the age of 65.

Solution: We have: $\mathbf{N} = 621$, $n = 60$, $p = \frac{11}{60} = 0.18\dot{3}$ $(= p)$, $z^*_\alpha = 1.64485$ for 90% confidence,

so that: $\hat{s.d.}(P) = \sqrt{\frac{n}{n-1}pq(\frac{1}{n} - \frac{1}{\mathbf{N}})} = \sqrt{\frac{60}{59}(\frac{11}{60} \times \frac{49}{60})(\frac{1}{60} - \frac{1}{621})} = 0.047\,879\,847.$

Hence, an approximate 90% confidence interval for $\mathbf{P}$, the respondent population proportion of households with at least one person over the age of 65, is:

$p \pm 1.64485 \times \hat{s.d.}(P) = 0.18\dot{3} \pm 1.64485 \times 0.047\,8798 \implies (0.1046,\ 0.2621)$ or about (10, 27%).

NOTES: 16. Despite the appreciable sampling fraction of about ten per cent, the confidence interval is, as in Example 8.17.1, relatively *wide* – *i.e.*, the small sample size of 60 means that sample error imposes appreciable limitation on the answer involving a (fairly small) respondent population proportion of (we estimate) around 18%.

17. With a sample size of 60, the continuity correction of $\pm 1/2n$ is ± 0.0083; the confidence intervals without and with the correction are given in Table 8.17.5 at the right. We see that, in this situation, the difference is unlikely to be practically important – we would typically give a final value for both intervals as about (10, 27%).

Table 8.17.5: Calculated	Confidence interval Value	Interval width	Interval width $\div p$
with*out* continuity correction	(0.1046, 0.2621)	0.1575	86%
with continuity correction	(0.0963, 0.2704)	0.1741	95%

(*continued overleaf*)

4. Confidence Intervals for $\mathbf{NP}$, the Respondent Population Number or Frequency

Under the assumption that the population size, $\mathbf{N}$, is a *known constant*, the theory of equiprobable selecting for estimating a population number or frequency is a straight-forward extension of that for a proportion. Because the population frequency is $\mathbf{NP}$, its estimator is $\mathbf{N}P$; the standard deviation of this estimator is then $\mathbf{N} \times s.d.(P)$. Hence, we obtain an interval involving random variables:

$$I = [\mathbf{N}P - z_\alpha^* \sqrt{\tfrac{\mathrm{n}}{\mathrm{n}-1}\mathbf{N}^2 PQ(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})},\ \ \mathbf{N}P + z_\alpha^* \sqrt{\tfrac{\mathrm{n}}{\mathrm{n}-1}\mathbf{N}^2 PQ(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})}],$$

such that $\Pr(I \ni \mathbf{NP}) \simeq 100(1-\alpha)\%$, where z_α^* is the $100(1-\alpha/2)th$ percentile of the $N(0, 1)$ distribution. For calculating an approximate $100(1-\alpha)\%$ *confidence interval* for $\mathbf{NP}$, we use:

$$\mathbf{N}p \pm z_\alpha^* \sqrt{\tfrac{\mathrm{n}}{\mathrm{n}-1}\mathbf{N}^2 pq(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})} = [\mathbf{N}p - z_\alpha^* \sqrt{\tfrac{\mathrm{n}}{\mathrm{n}-1}\mathbf{N}^2 pq(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})},\ \ \mathbf{N}p + z_\alpha^* \sqrt{\tfrac{\mathrm{n}}{\mathrm{n}-1}\mathbf{N}^2 pq(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})}]. \quad \text{----(8.17.19)}$$

The discussion near the end of Note 12 on page 8.96 shows that including a continuity correction of $\mathbf{N}/2\mathrm{n}$ in the confidence interval expressions (8.17.19) may be desirable when the sample size is near the limit for reasonable use of the normal approximation to the binomial distribution.

Example 8.17.3: An auditor obtains, by equiprobable selecting, 40 accounts from 1,200 accounts of a certain firm; she finds that, for 16 of the 40 accounts, the underlying documentation is not in compliance with stated procedures. Find an interval estimate of the *number* of the firm's accounts not in compliance.

Solution: We have: $\mathbf{N} = 1{,}200$, $\mathrm{n} = 40$, $\mathrm{p} = \frac{16}{40} = 0.4\ (= p)$, $z_\alpha^* = 1.95996$ for 95% confidence,

so that: $$\hat{s.d.}(\mathbf{N}P) = \sqrt{\tfrac{\mathrm{n}}{\mathrm{n}-1}\mathbf{N}^2 pq(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})} = \sqrt{\tfrac{40}{39}(1{,}200^2 \times 0.4 \times 0.6)(\tfrac{1}{40} - \tfrac{1}{1{,}200})} = 92.553\,518.82$$

Hence, an approximate 95% confidence interval for $\mathbf{NP}$, the number of the firm's accounts not in compliance with stated procedures, is:

$$\mathbf{N}p \pm 1.95996 \times \hat{s.d.}(\mathbf{N}P) = 1{,}200 \times 0.4 \pm 1.95996 \times 92.553\,519 = 480 \pm 181.401\,195 \implies (298.60,\ 661.40)$$

or about (290, 670) accounts.

NOTES: 18. When the confidence *level* is not specifed in the question, we take the default as 95%. Also, the two significant figures given for the rounded end points of the final interval reflect the fact that there are only two significant digits in the number of accounts in the sample which were not in compliance with stated procedures.

19. The confidence interval is relatively *wide* – *i.e.*, the small sample size of 40 accounts means that sample error imposes appreciable limitation on the answer, despite a (statistically favourable) respondent population proportion of (we estimate) around 40%.

20. With a sample size of 40, the continuity correction of $\mathbf{N}/2\mathrm{n}$ is ± 15; the confidence intervals without and with the correction are given in Table 8.17.6 at the right below. We see that, with such a small sample size, the difference in the intervals may be large enough to be of practical importance.

Table 8.17.6: Confidence interval		Interval width	Interval width $\div\ \mathbf{N}p$
Calculated	**Value**		
with*out* continuity correction	(298, 662)	364	76%
with continuity correction	(283, 677)	394	82%

Example 8.17.4: Inspection of 200 items obtained by equiprobable selecting from a production line yields 3 defectives.
(a) If the line produces 10,000 items a day, find an approximate 99% confidence interval for the number of defectives produced in a day.
(b) Comment briefly on the use of such a sample survey to measure the defect rate of this production process.

Solution: (a) We have: $\mathbf{N} = 10{,}000$, $\mathrm{n} = 200$, $\mathrm{p} = \frac{3}{200} = 0.015\ (= p)$, $z_\alpha^* = 2.57583$ for 99% confidence,

so that: $$\hat{s.d.}(\mathbf{N}P) = \sqrt{\tfrac{\mathrm{n}}{\mathrm{n}-1}\mathbf{N}^2 pq(\tfrac{1}{\mathrm{n}} - \tfrac{1}{\mathbf{N}})} = \sqrt{\tfrac{200}{199}(10{,}000^2 \times 0.015 \times 0.985)(\tfrac{1}{200} - \tfrac{1}{10{,}000})} = 85.300\,238\,29.$$

Hence, an approximate 99% confidence interval for $\mathbf{NP}$, the number of defectives produced in a day, is:

$$\mathbf{N}p \pm 2.57583 \times \hat{s.d.}(\mathbf{N}P) = 10{,}000 \times 0.015 \pm 2.57583 \times 85.3002 = 150 \pm 219.719 \implies (-69.719,\ 369.719)$$

or about (0, 370) defectives.

(b) The confidence interval for $\mathbf{NP}$, even in its rounded final form of (0, 370) defectives, has a width nearly 2½ times the magnitude of $\mathbf{NP}$, reflecting the *severe* limitation imposed on the answer by sample error. Example 8.17.4 again reminds us that:

- appreciable (*i.e.*, resource-intensive) sample sizes may be needed to estimate a (small) proportion with adequate precision – that is, to make acceptable in the Question context the limitation imposed on an answer by sample error;

(*continued*)

Figure 8.17. UNSTRATIFIED POPULATIONS: One-Stage EPSWOR of Individual Elements Estimating a Proportion or a Frequency (continued 3)

Solution: (cont.) (b) ● the *smaller* the magnitude of the population proportion $\mathbf{P}$, the *larger* the sample size needed to attain a given level of precision specified *relative to* the magnitude of $\mathbf{P}$.

The typical sample size of around 1,500 respondents in national polls is used in part to provide what is deemed to be an acceptable width (*viz*., about 2.5 percentage points) for a 'worst case' 95% confidence interval and in part to give still acceptable (although lower) precision for poll results for subdivisions of the national data such as males and females or regions (*e.g*., in Canada, Atlantic Canada, Quebec, Ontario, the Prairie provinces, British Columbia) – recall also Table 8.17.3 at the upper right of page 8.97.

NOTES: 21. The unrealizable *negative* lower end point for the confidence interval on page 8.98 has been rounded to zero (the nearest realizable value) for the final statement of the interval. The negative value reminds us to assess:
- the *precision* of the estimate of $\mathbf{NP}$.
- the accuracy of the hypergeometric-binomial-normal approximation sequence – see Note 7 on page 8.94.

22. With a sample size of 200, the continuity correction of $\pm\mathbf{N}/2n$ is ± 25; the confidence intervals without and with the correction are given in Table 8.17.7 at the right. We see that the difference in the intervals might be large enough to be of practical importance.

Table 8.17.7:

Confidence interval Calculated	Value	Interval width	Interval width $\div\mathbf{N}p$
with*out* continuity correction	(0, 370)	370	247%
with continuity correction	(0, 395)	395	263%

23. A summary of information relevant to precision from the four foregoing Examples is given in Table 8.17.8 at the right; the Examples are ordered by sample size. The last column is the estimated *cofficient of variation*, given by $\hat{s.d.}(P)/p$ [equation (8.17.9)], a measure of *relative precision* of the estimator P of $\mathbf{P}$; the standard deviation is a measure of *absolute* precision. These coefficient of variation values are of interest in light of the comment by Dr. Glenn in the left-hand column near the top of page 8.62 in Figure 8.11 about a c.v. of no more than 30% being needed for a proportion to be estimated with acceptable sample error.

Table 8.17.8

Example	n	p	$\hat{s.d.}(P)$	$\hat{c.v.}(P)$
8.17.3	40	0.4	0.0771	19%
8.17.2	60	$0.18\dot{3}$	0.0479	26%
8.17.1	100	0.15	0.0293	20%
8.17.4	200	0.015	0.00853	57%

NOTES: 24. The theory in this Figure 8.17 for estimating a proportion in two categories under EPS involves *four* standard deviations – two data s.d.s $\mathbf{S}$ and s, two probabilistic s.d.s $s.d.(P)$ and $\hat{s.d.}(P)$ [the last one is the *estimated* standard deviation of P (when $\mathbf{P}$ and $\mathbf{Q}$ are estimated by p and q)]; recognizing and maintaining the distinctions among these s.d.s [and the (model) random variable S and its value s] is part of understanding and applying the theory.

25. The standard deviation of P is sometimes called the *standard error* of P (*e.g*., Barnett, p. 45); while you should recognize this useage when you encounter it, it has been avoided in these Course Materials because the term is not always used with the same meaning by different authors (see also Cochran, pages 24, 25-27 and 53).

26. A more general reminder about terminology concerns three other terms:
- when you encounter the term *standard deviation* or *standard error*, be sure you understand whether it refers to:
  - the variation of data (and whether the data are from a *sample* or a *census*), **OR**
  - a random variable [and whether it is an individual random variable or a (linear) combination (*e.g*., an average, sum or difference)], **OR**
  - a parameter of a response model or probability model.
- when you encounter the word *bias*, remember that it is defined only in the context of *repetition* of a *process* – for instance, selecting or measuring.
  - *Estimating* bias *differs* from, for example, (real-world) measuring *inaccuracy* in that it *de*creases with *in*creasing sample size and so may not be of much practical concern in actual sample surveys.

  Recall that bias (and rms error) are discussed in Appendix 5, Appendix 6 and Appendix 7 on pages 8.58 to 8.61 in Figure 8.11.

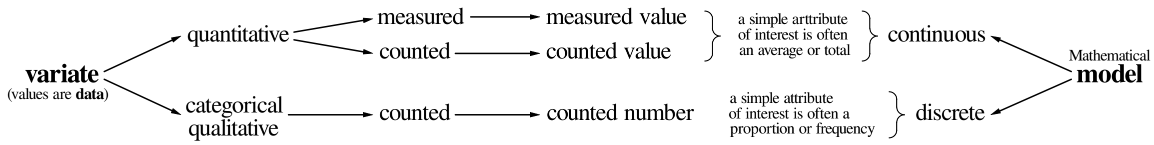
5. Appendix 1: Terminology Review [optional reading]

∗ **Variate:** a characteristic associated with each *element* of a population/process.
- **Response variate ($\mathbf{Y}$):** a variate defined in the Formulation stage of the FDEAC cycle; an Answer gives some attribute(s) of the response variate over the target population/process.
- **Explanatory variate ($\mathbf{Z}$):** a variate, defined in the Formulation stage of the FDEAC cycle, that accounts, at least in part, for changes from element to element in the value of a response variate.

(*continued overleaf*)

- **Focal (explanatory) variate ($\not{X}$):** for a Question with a *causative* aspect, the focal variate is the *explanatory* variate whose relationship to the *response* variate is involved in the Answer(s) to the Question(s).

Distinctions associated with *other* adjectives which qualify 'variate' – **quantitative**, **categorical** (or **qualitative**), **measured** and **counted**, **continuous** and **discrete** – are shown in the following schema.



- An illustration of the distinction between a counted *value* [a real number, modelled by a **continuous** variate] and a counted *number* [an integer, modelled by a **discrete** variate] is:
  - to assess the effectiveness of an insecticide, the number of insects could be counted on a defined part of each of a sample of plants from *un*treated and treated crops – a measure of effectiveness would be the decrease in the *average* number of insects per plant;
  - to assess gender balance in an area of employment, the number of men and women employed in the area could be counted – a (simple) measure of the balance would be the *proportion* of each sex in the area (but see also Simpson's Paradox, Figure 9.8).

  The quantitative *measured* and quantitative *counted* distinction blurs progressively as counts become larger in magnitude; it is also affected by the limited resolving power (finite precision) of real measuring processes.

  Continuous and discrete variates are *model* concepts because real measuring instruments with finite precision yield only *discrete* values.
- *Quantitative* variate values can become (ordinal) *categorical* – *e.g.*, ages can be classified into age *groups*; *we* take *qualitative* to mean nominal (*non*-ordinal) categorical – *e.g.*, marital status or skin colour.
- A *binary* variate is a categorical variate in *two* categories.

6. Appendix 2: Notation Review [optional reading]

Random variables (discrete or continuous) are a component of the probability *models* used in statistics; *our* concern is more commonly with *discrete* random variables which take on a *finite* number of values (which are often numerical but may be categorical). *We* denote random variables by upper-case *italic* letters, usually from near the end of the alphabet; we try to keep upper-case italic letters near the start of the alphabet for *events* (bearing in mind the choice of a letter which needs to be *evocative*).

- Formally, a random variable is an entity that assigns a probability to each point of the sample space.
- Informally, a random variable is an entity that takes on different values according to chance.

Both statements mean that we can specify a random variable by a table that gives its values and their probabilities. For example, for selecting equiprobably two balls from an urn containing three red balls and three black balls, the probability functions for the number of red balls in the sample are shown at the right for selecting without and with replacement.

r	0	1	2
$f(r)$	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$

r	0	1	2
$f(r)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

The notation in these tables is a lower case *italic* letter (r) for the random variable *values* and f for the **probability function** that gives the *probabilities*. Other examples of probability functions are the hypergeometric and binomial p.f.s in equations (8.17.10) and (8.17.11) on the lower half of page 8.94 and the p.f. of the random variable V_j in equation (8.17.1) near the middle of page 8.93.

Our use of f (rather than p) for a probability function allows us to keep the letter p for a *proportion*, with four meanings as shown at the right in Table 8.17.9. The distinctions we can maintain with this notation are easily lost when the predominant use of letter p is (the random variable look-alike) P for probability, as is common practice in introductory statistics texts. In particular, our distinguishing p and p reminds us of the statistical requirement for the Plan for an investigation to make it reasonable to treat sample proportions as values of a (model) random variable P.

Table 8.17.9

QUANTITY	SYMBOL
Population proportion	Ᵽ
Sample proportion	p
Random variable	P
A value of P	p
Model parameter	π
Probability	Pr

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

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

Appendix 8 of Figure 8.11 continued from page 8.62

It is clear from a few quick calculations that all of the "answers" cited on page 8.62 are based on *two* observations. One was the white female in Region 2, the other the black male in Region 3, both of whom admitted to being regular users of heroin. These were "projected" into 1,440 and 1,320 respectively, and the rest was simple arithmetic. However, arithmetic cannot erase the very tenouus nature of the answers. The only valid answer is that, in the general population of North Carolina, two people were ready to admit to the regular use of heroin. One was a white female in Region 2; the other was a black male in Region 3. What does this tell us about regular users of heroin in the state? Very little, except that out of 2,007 people interviewed, two were willing to admit being regular users. Apparently one person in 1,000 interviewed is willing to admit to regular use of heroin. The number who are *actually* regular users and their sociodemographic characteristics remain as unknown as they were before the survey.