Assignment 10 Outline Solution

A10 – 1. **(a)** This part of the question involves systematic selecting (SyS) for a *number* (or *frequency*) derived from the corresponding *proportion*. We assume that the vouchers were selected by a method essentially equivalent to equiprobable selecting (EPS) from the population of vouchers, so that $\mathbf{P}$, the respondent population proportion of incorrectly-filled vouchers, is estimated by p_{sy}, the value of the random aroiable P_{sy} repesenting the sample proportion of incorrectly-filled vouchers under SyS of vouchers.

We have: $\mathbf{N} = 250$, $\quad n = 50$, $\quad p_{sy} = 20/50 = 0.40$ or 40% $(= p_{sy})$, $\quad \mathbf{N}p_{sy} = 100$ vouchers, $\quad z^*_\alpha = 2.57583$ for 99% confidence,

so that: $\hat{s.d.}(P_{sy}) = \sqrt{\frac{n}{n-1} p_{sy} q_{sy} \left(\frac{1}{n} - \frac{1}{\mathbf{N}}\right)} = \sqrt{\frac{50}{49}(0.40 \times 0.60)\left(\frac{1}{50} - \frac{1}{250}\right)} = 0.062\,596\,863.$ $\qquad (q_{sy} = 1 - p_{sy} = 0.60).$

Hence, an approximate 99% confidence interval for $\mathbf{NP}$, the respondent population number of incorrectly-filled travel vouchers, is:

$\mathbf{N}p_{sy} \pm 2.57583 \times \mathbf{N}\hat{s.d.}(P_{sy}) = 100 \pm 2.57583 \times 15.649\,215\,93 \implies (59.69,\ 140.31)$ or about $(59,\ 141)$ vouchers.

(b) Two matters which could adversely affect the stated confidence level are:

- the sample size – with n as small as 50 (despite the relatively high sampling *fraction*), there is concern about the adequacy of the normal approximation for the distribution of P_{sy} as a consequence of the Central Limit Theorem;
- assuming that the order of the vouchers is such that SyS is equivalent to EPS – there is no information in the question to indicate whether this assumption is reasonable in this contex;t but the stated confidence level might be adversely affected if the vouchers were arranged in chronological order, say, or if the accuracy of filling had been affected by factors like a change in the personnel responsible or an alteration in the form or in the regulations governing the filling process.

A10 – 2. The situation outlined in this question can be thought of as involving a population with an *in*creasing approximately linear trend in the response of interest, when there are two potential advantages of SyS over EPS.

- *Convenience*: SyS is likely to be appreciably *easier* to implement than EPS; the former can be based on some measure of the *position* of the mortgage records in their storage location, whereas the latter involves generating from a table of equiprobable digits n numbers in the interval $[1, \mathbf{N}]$ and then identifying these particular records in the storage location.
- *Precision*: SyS will usually yield a sample that is more evenly spread over the population, which should increase the *precision* of the estimate; however, in the absence of relevant SyS theory, a (conservative) estimated standard deviation of the estimate would still often be calculated on the basis of EPS theory.

A10 – 3. This question involves approximately 1-in-20 SyS from the membership list.

(a) When the list is arranged alphabetically, there is no obvious reason, in most circumstances, why rate of remuneration should depend on last name, so that SyS should be closely equivalent to EPS but more convenient to carry out, for reasons analogous to those discussed above in the solution of question **A10 – 2**. SyS is thus the preferred method of selecting the sample from the *alphabetic* membership list.

(b) When the list is arranged by length of membership in the society, there is likely to be an association between age and position on the list – on average, *older* people will have been members for *longer.* However, there is *also* an association between age and salary level, with older people tending to have *higher* incomes because of factors such as seniority and experience. Thus, position on the list and salary would be expected to be related in this situation – the relationship might reasonably be thought of as producing an approximately linear trend (an *in*creasing trend if longer memberships are *later* on the list). In this situation, SyS is the preferred method, because it can have advantages over EPS not only of convenience but *also* of improved precision, as outlined in the second comment in the solution of question **A10 – 2** above and elaborated in question **A10 – 4** below and overleaf.

A10 - 4. **(a)** The population average is: $\overline{\mathbf{Y}} = \frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}} \mathbf{Y}_i = \frac{1}{\mathbf{N}}\sum_{i=1}^{\mathbf{N}} i = \frac{1}{\mathbf{N}} \bullet \frac{\mathbf{N}(\mathbf{N}+1)}{2} = \frac{\mathbf{N}+1}{2};$

also: $\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2 = \sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - 2\overline{\mathbf{Y}}\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i + \mathbf{N}\overline{\mathbf{Y}}^2 = \sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - \mathbf{N}\overline{\mathbf{Y}}^2$ $\qquad$ because $\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i = \mathbf{N}\overline{\mathbf{Y}}$,

so that the population (data) standard deviation is:

$$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\Big[\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - \mathbf{N}\overline{\mathbf{Y}}^2\Big]} = \sqrt{\frac{1}{\mathbf{N}-1}\Big[\sum_{i=1}^{\mathbf{N}} i^2 - \mathbf{N}\overline{\mathbf{Y}}^2\Big]} = \sqrt{\frac{1}{\mathbf{N}-1}\Big[\frac{\mathbf{N}(\mathbf{N}+1)(2\mathbf{N}+1)}{6} - \frac{\mathbf{N}(\mathbf{N}+1)(\mathbf{N}+1)}{4}\Big]}$$

$$= \sqrt{\frac{\mathbf{N}(\mathbf{N}+1)}{12} \bullet \frac{4\mathbf{N}+2-3\mathbf{N}-3}{\mathbf{N}-1}} = \sqrt{\frac{\mathbf{N}(\mathbf{N}+1)}{12}}.$$

(b) We have $\mathbf{N} = n\mathbf{K}$ and, from (a), $\mathbf{S}^2 = \frac{\mathbf{N}(\mathbf{N}+1)}{12}$.

(*continued overleaf*)

Assignment 10 Outline Solution (continued 1)

A10 – 4. (cont.) (b) $[s.d.(\overline{Y}_{us})]^2 = \mathbf{S}^2(\frac{1}{n} - \frac{1}{\mathbf{N}}) = \frac{\mathbf{N}(\mathbf{N}+1)}{12}(\frac{1}{n} - \frac{1}{n\mathbf{K}}) = \frac{n\mathbf{K}(\mathbf{N}+1)}{12} \bullet \frac{\mathbf{K}-1}{n\mathbf{K}} = \frac{(\mathbf{K}-1)(\mathbf{N}+1)}{12}.$

(c)(i) The (data) standard deviation expression in (a) is for the population of the first $\mathbf{N}$ (consecutive) integers. In (c), each of the n strata consists of $\mathbf{K}$ consecutive integers so that, by analogy, each stratum (data) standard deviation will be given by the *same* expression as in (a), except that $\mathbf{K}$ replaces $\mathbf{N}$ and the square root is omitted because (c) asks about the *squared* s.d.
[In contrast to the *equal* standard deviations of the n strata, their *averages* differ and increase progressively (by the same amount) from one stratum to the next.]

(ii) $[s.d.(\overline{Y}_{st})]^2 = \sum_{h=1}^{n}\mathbf{W}_h^2\mathbf{S}_h^2(\frac{1}{n_h} - \frac{1}{\mathbf{N}_h}) = \sum_{h=1}^{n}\frac{1}{n^2}\bullet\frac{\mathbf{K}(\mathbf{K}+1)}{12}(\frac{1}{1} - \frac{1}{\mathbf{K}}) = n\bullet\frac{1}{n^2}\bullet\frac{\mathbf{K}(\mathbf{K}+1)}{12}\bullet\frac{\mathbf{K}-1}{\mathbf{K}} = \frac{(\mathbf{K}-1)(\mathbf{K}+1)}{12n}.$

(d) $\overline{\mathbf{Y}}_1 = \frac{1}{n}[1 + (\mathbf{K}+1) + (2\mathbf{K}+1) + \ldots + ((n-1)\mathbf{K}+1)] = \frac{1}{n}[n + \mathbf{K}(1+2+\ldots+(n-1))] = \frac{1}{n}[n + \mathbf{K}\frac{(n-1)n}{2}] = 1 + \frac{\mathbf{K}(n-1)}{2},$

so that: $\overline{\mathbf{Y}}_i = i + \frac{\mathbf{K}(n-1)}{2}$ and: $\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}} = i + \frac{\mathbf{K}(n-1)}{2} - \frac{\mathbf{N}+1}{2} = i + \frac{n\mathbf{K} - \mathbf{K} - \mathbf{N} - 1}{2} = i - \frac{\mathbf{K}+1}{2};$

hence:

$$[s.d.(\overline{Y}_{sy})]^2 = \frac{1}{\mathbf{K}-1}\sum_{i=1}^{\mathbf{K}}(\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}})^2 = \frac{1}{\mathbf{K}-1}\sum_{i=1}^{\mathbf{K}}[i - \frac{\mathbf{K}+1}{2}]^2 = \sum_{i=1}^{\mathbf{K}}i^2 - 2\sum_{i=1}^{\mathbf{K}}i\bullet\frac{\mathbf{K}+1}{2} + \sum_{i=1}^{\mathbf{K}}\frac{(\mathbf{K}+1)^2}{4}]$$

$$= \frac{1}{\mathbf{K}-1}[\frac{\mathbf{K}(\mathbf{K}+1)(2\mathbf{K}+1)}{6} - 2\frac{\mathbf{K}(\mathbf{K}+1)}{2}\bullet\frac{\mathbf{K}+1}{2} + \mathbf{K}\frac{(\mathbf{K}+1)^2}{4}]$$

$$= \frac{\mathbf{K}(\mathbf{K}+1)}{12(\mathbf{K}-1)}[4\mathbf{K} + 2 - 6\mathbf{K} - 6 + 3\mathbf{K} + 3] = \frac{\mathbf{K}(\mathbf{K}+1)}{12}.$$

NOTE: The first expression above for $[s.d.(\overline{Y}_{sy})]^2$ in the third line of (d) looks like equation (2.14.3) (without the square root) near the bottom of page 2.105 in Figure 2.14; the (key) difference is that $\overline{\mathbf{Y}}_i$ in (d) is one of $\mathbf{K}$ *sample* averages, whereas in equation (2.14.3) it is one of $\mathbf{M}$ (equal-sized) *cluster* averages.

Numerical Illustrations for Question A10 – 4

$\mathbf{N} = 100,\ n = 10,\ \mathbf{K} = 10.$

Sampling protocol	No. of samples	$[s.d.(\overline{Y})]^2$	S.d. ratio
EPS unstratified	$\sim 1.73\times10^{13}$	$75.75 \simeq 8.703^2$	9.6
SyS unstratified	10	$9.1\dot{6} \simeq 3.028^2$	3.3
EPS stratified	10^{10}	$0.825 \simeq 0.908^2$	1
$\mathbf{N} = 10{,}000,\ n = 100,\ \mathbf{K} = 100.$			
EPS unstratified	$\sim 6.52\times10^{241}$	$82{,}508 \simeq 287^2$	99.5
SyS unstratified	10^2	$841.\dot{6} \simeq 29^2$	10.1
EPS stratified	10^{200}	$8.332 \simeq 2.89^2$	1

(e) Before addressing the two parts of (e), we note the matters in the following three bullets (●).

- The three different sampling protocols – EPS and SyS from an unstratified population, EPS from a stratified population – involve (vastly) different numbers of possible samples; as shown in the second column of the upper half of the illustrative table at the right above, there are about 17.3 trillion possible samples under EPS from an unstratified population, 10 billion possible samples under EPS from a stratified population (a factor of about 1,730 times fewer), and only 10 possible samples under SyS from an unstratified population.
  – For larger (more realistic) values of $\mathbf{N}$, n and $\mathbf{K}$ in the lower half of the table, the numbers of possible samples and their differences are (mostly incomparably) greater.
- The purpose of selecting equiprobably (and the reason for its emphasis in introductory statistics) is to provide a basis for assessing the likely size of sample error (e.g., as a confidence interval); from EPS theory, it emerges that three factors affect the sampling precision of estimating the (respondent) population average $\overline{\mathbf{Y}}$:
  – the population (data) standard deviation $\mathbf{S}$ (which typically cannot be altered by investigators);
  – the sample size n – sampling precision increases as $\sqrt{n}$ so that, in the theory, its 2-fold increase requires a 4-fold increase in sample size (sample size can usually be set by investigators, subject to resource constraints);
  – the population size $\mathbf{N}$ but, in the (common) situation where $n \ll \mathbf{N}$, the effect of $\mathbf{N}$ on sampling precision is practically unimportant.

  A consequence of the foregoing is that Plans of real sample sutveys incorporrate EPS into more complicated sampling protocols like (possibly multistage) selecting from a stratified population with a view to managing precision of estimates. There can also be other Design-stage considerations in developing the Plan such as:
  – allowing breakdown of answers by explanatory variates (like region or gender).
  – convenience (like availability of a *clustered* frame or a population structure conducive to *systematic* selecting.
- The first $\mathbf{N}$ positive integers in increasing order is an idealized population compared with the 'trends' in most real-world populations; this (unfortunately) restricts the broader usefulness of the answers to (e) which follow.

(i) When estimating a population average, precision increases as the selecting protocol becomes more likely to yield a sample that 'looks like' the population – in the situation of question **A10 – 4** with a (strict) linear trend, this means a sample *spread out* over the population.

(*continued*)

Assignment 10 Outline Solution (continued 2)

A10 – 4. (cont.) (e)(i)

- SyS and EPS from a stratified population (as the stratification is implemented here) achieve this goal by selecting one element from the first $\mathbf{K}$ integers, one from the second $\mathbf{K}$ integers, and so on up to the n*th* $\mathbf{K}$.
- By contrast, EPS from an unstratified population, by selecting *all* possible samples of size n, yields many samples containing elements from limited regions of the population, the two worst cases here being the first n integers and the last n. This is the source of the (relatively) poor precision of EPS here.

(ii) EPS from a stratified population achieves higher precision than SyS by managing the effect of $\mathbf{S}$ – the $\mathbf{K}$ strata are defined so that the $\mathbf{S}_h$ [which, rather than $\mathbf{S}$, are involved in the expression for $s.d.(\overline{Y}_{st})$] are (appreciably) *smaller* than $\mathbf{S}$; in contrast, it is this *population* standard deviation which influences $s.d.(\overline{Y}_{sy})$ for which, in the (artificial) situation of question **A10 – 4**, we can derive a symbolic expression.

- The *difference* between the (otherwise similar) sampling protocols of EPS from a stratified population and SyS from an *un*stratified population is:
  - in the former, the one element from the $\mathbf{K}$ elements of each stratum is obtained by EPS, WHEREAS:
  - in the latter, for each sample the element obtained from each successive group of $\mathbf{K}$ elements is at a *fixed* position within these n groups.

  This difference likely contributes to the higher precision of the stratified sampling protocol.
- The third and fourth columns of the illustrative table at the centre right of the facing page 0.26b show, in the upper half of the table for modest values of $\mathbf{N}$, n and $\mathbf{K}$, that the precision of the estimate of $\overline{\mathbf{Y}}$ increases by a factor of about 3 for the three successive sampling protocols; this factor increases to around 10 for the larger $\mathbf{N}$, n and $\mathbf{K}$ values in the lower half of the table.

NOTE: With overall similarity but one clear difference in the two sampling protocols of EPS from a stratified population and SyS from an *un*stratified population, it is noteworthy that, in the (artificial) situation of question **A10 – 4**, the expression for $[s.d.(\overline{Y}_{sy})]^2$ in (d) is the *same* as that for $\mathbf{S}_h^2$, the stratum (data) standard deviation squared in (c)(i).

(f) The answers in (e) have several implications for sample selecting practice.

- EPS from an unstratified population is usually *not*, by itself, a realistic sampling protocol.
  - Favourable conditions for the precision of this 'simple' sampling protocol are when the values of the response(s) of interest tend to occur 'haphazardly' among the population elements.
- The advantage of retaining EPS in a more complicated sampling protocol is that it provides a basis for assessing the likely size of sample error.
- Knowledge of population structure can lead to more effective managemnent of population variation so as to increase precision when estimating an average. For a population with a 'linear' trend:
  - EPS from the population stratified appropriately can yeild a practically useful increase in precision
  - the advantage of convenience for SyS must be weighed against its lack of (general) theory for assessing the likely size of sample error.
- It may also be possible to exploit similarly other population 'regularities' like a cyclical pattern in the response(s) of interest.

POSTSCRIPT: Under EPS from an unstratified population, the numbers of samples possible, in the second column of the illustrative table at the centre right of the facing page 0.26b, remind us that a simple combinatorial context can lead to unimaginably large numbers, shown again at the right.

$$\binom{100}{10} \simeq 1.731\,030\,934 \times 10^{13}$$
$$\binom{10{,}000}{100} \simeq 6.520\,847 \times 10^{241}$$

Suppose that the life of the universe is 15 billion years = 1.5×10^{10} years $\simeq 1.3 \times 10^{14}$ hours $\simeq 4.7 \times 10^{17}$ seconds.

- Under EPS, the number of possible (small) samples of size 10 from a (small) population of size 100 is only about an order of magnitude smaller than the age of the universe in hours.
- Under EPS, if a (different) sample of size 100 were to be selected every second from a population of size 10,000 for the lifetime of the universe, the resulting set of 4.7×10^{17} samples would be roughly (the unimaginable) 10^{214} times *smaller* than the set of all possible samples.
  - The situation is not materially altered by selecting *one billion* samples per second rather than one.

These matters provide perspective on statements made in other contexts about calculations so onerous that they could not be completed (by the current fastest computers) in the life of the universe.

Blank page