University of Waterloo STAT 332 – W. H. Cherry

Assignment 10

A10 – 1. An auditor wished to estimate the number of incorrectly-filled travel vouchers out of $\mathbf{N} = 250$ vouchers. She used systematic selecting with $\mathbf{K} = 5$ (*i.e.*, she selected every 5th travel voucher), and found that 20 of those obtained were filled incorrectly.

(a) Estimate the total number of incorrectly-filled vouchers, and find an approximate 99% confidence interval for this total.

(b) State the assumption(s under which the confidence interval in (a) is likely to have an *actual* confidence leve;l not too far below its *stated* level.

A10 – 2. Suppose that a trust company has $\mathbf{N}$ mortgages numbered serially in the order in which they were granted over a period of 20 years. There is a generally increasing trend in the amount owing per mortgage because of the rising cost of housing over the 20 years. It is desired to estimate the total amount owing on all the mortgages using a sample survey of the mortgage records. Would you use *equiprobable* selecting (EPS) or *systematic* selecting (SyS) of the records? Justify your choice.

A10 – 3. A professional society publishes a list of its members. In a sample survey to estimate the average salary of members of the profession, this membership list is to be used as the basis for selecting a sample of about 5% of the 2,640 members of the profession. Discuss briefly the possible advantage(s) and disadvantage(s) of *equiprobable* selecting (EPS) or *systematic* selecting (SyS) of members in the two cases:

(a) where the list is in alphabetical order by last name;

(b) where the list is in order of length of membership of the society.

A10 – 4. A population of $\mathbf{N}$ elements has a response variate $\mathbf{Y}_i = i$, $i = 1, 2, \ldots, \mathbf{N}$; this is a case of a population with a 'linear trend' in its response. To assist with the questions which follow, there are two results at the right for the sum of the first $\mathbf{N}$ integers and for the sum of their squares.

$$\sum_{i=1}^{\mathbf{N}} i = \frac{\mathbf{N}(\mathbf{N}+1)}{2} \qquad \sum_{i=1}^{\mathbf{N}} i^2 = \frac{\mathbf{N}(\mathbf{N}+1)(2\mathbf{N}+1)}{6}$$

(a) Show that the population average is given by $\overline{\mathbf{Y}} = \frac{\mathbf{N}+1}{2}$,
and the population (data) standard deviation is given by $\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{Y}_i - \overline{\mathbf{Y}})^2} \equiv \sqrt{\frac{1}{\mathbf{N}-1}\left[\sum_{i=1}^{\mathbf{N}}\mathbf{Y}_i^2 - \mathbf{N}\overline{\mathbf{Y}}^2\right]} = \sqrt{\frac{\mathbf{N}(\mathbf{N}+1)}{12}}$.

(b) If $\mathbf{N} = n\mathbf{K}$, show that for n units obtained by EPS from the *un*stratified population: $[s.d.(\overline{Y}_{us})]^2 = \frac{(\mathbf{K}-1)(\mathbf{N}+1)}{12}$.

(c) Suppose that the population is arranged in order of increasing response variate value and then divided into n strata each of size $\mathbf{K}$, so that $\mathbf{W}_h = \mathbf{K}/\mathbf{N} = 1/n$ for $h = 1, 2, \ldots, n$.

(i) Explain why each stratum (data) standard deviation squared is given by $\mathbf{S}_h^2 = \frac{\mathbf{K}(\mathbf{K}+1)}{12}$.

(ii) Show that for a sample of one unit obtained from each stratum by EPS: $[s.d.(\overline{Y}_{st})]^2 = \frac{(\mathbf{K}-1)(\mathbf{K}+1)}{12n}$.

(d) For a 1-in-$\mathbf{K}$ sample obtained by SyS from this population with*out* stratification, it can be shown that $\overline{\mathbf{Y}}_i - \overline{\mathbf{Y}} = i - (\mathbf{K}+1)/2$ for $i = 1, 2, \ldots, \mathbf{K}$, where $\overline{\mathbf{Y}}_i$ is the average of the i*th* 1-in-$\mathbf{K}$ sample obtained by SyS; using this result, show that: $[s.d.(\overline{Y}_{sy})]^2 = \frac{\mathbf{K}(\mathbf{K}+1)}{12}$.

(e) It is clear from (b), (c) and (d) that, for this population with a linear trend, $s.d.(\overline{Y}_{st}) < s.d.(\overline{Y}_{sy}) < s.d.(\overline{Y}_{us})$.

(i) Explain why, for this population, both EPS *with* prior stratification and SyS with*out* stratification would be expected to yield more precise estimates of $\overline{\mathbf{Y}}$ than EPS with*out* stratification.

(ii) Can you explain why, for this population, EPS *with* prior stratification yields a more precise estimate of $\overline{\mathbf{Y}}$ than SyS with*out* stratification?

(f) Discuss briefly the implications of your answers in (e) for sample survey selecting *practice*.

Blank page