University of Waterloo

STAT 332 – W. H. Cherry

Tutorial 6 Outline Solution

T6 – 1. This question involves equiprobable selecting (EPS) of clusters of size L = 10, to estimate $\overline{\mathbf{Y}}$, the (respondent) population average number of newspapers purchased per household.

We have: $\mathbf{M} = 400$, $\mathrm{m} = 4$; also: $\mathbf{N} = 4{,}000$, $\mathrm{n} = 40$.

(i) The four selected cluster averages are: $\bar{\mathrm{y}}_1 = 1.9\ (= \bar{y}_1)$; $\bar{\mathrm{y}}_2 = 2.0\ (= \bar{y}_2)$; $\bar{\mathrm{y}}_3 = 1.6\ (= \bar{y}_3)$; $\bar{\mathrm{y}}_4 = 2.0\ (= \bar{y}_4)$;
from these averages we obtain:

$\bar{\mathrm{y}}_{ec} = 1.875\ (= \bar{y}_{ec})$ newspapers, $\mathrm{s}_{ec} = 0.189\,296\,944\ (= s_{ec})$ newspapers, $\hat{s.d.}(\bar{Y}_{ec}) = 0.094\,174\,041$ newspapers.

With a sample size as small as 4 (clusters), we are concerned about the adequacy of the normal approximation for the distribution of $\bar{Y}_{ec}$ as a consequence of the Central Limit Theorem; this imposes a limitation on the answer of not knowing how close the *actual* confidence level is to the stated level of 90%. Also, with $\mathbf{S}_{ec}$ estimated by s_{ec}, we use the t_3 distribution value of 2.35336 on page 2.51 in Figure 2.4 to find an approximate 90% confidence interval for $\overline{\mathbf{Y}}$, the average number of newspapers putchased per household, as:

$1.875 \pm 2.35336 \times 0.094\,174 = 1.875 \pm 0.221\,625 \implies (1.6534,\ 2.0966)$ or about $(1.6,\ 2.1)$ newspapers.

(ii) Regarding the data as referring to 40 (individual) households obtained by EPS, and using the numerical summaries given at the right of the table in the statement of the question, we have:

$\bar{\mathrm{y}} = 1.875\ (= \bar{y})$ newspapers, $\mathrm{s} = 1.066\,686\,699\ (= s)$ newspapers, $\hat{s.d.}(\bar{Y}_{in}) = 0.167\,812\,567$ newspapers.

With a sample size of n = 40 (elements), we are appreciably less concerned than in (i) about the adequacy of the normal approximation (as a consequence of the Central Limit Theorem) for the distribution of $\bar{Y}_{in}$; hopefully, the *actual* confidence level of the answer *is* (reasonably) close to the stated level of 90%. For 39 degrees of freedom, the relevant entry from the t table on page 2.51 in Figure 2.4 is 1.68488, so that an approximate 90% confidence interval for $\overline{\mathbf{Y}}$, the average number of newspapers purchased per household, is:

$1.875 \pm 1.68488 \times 0.167\,813 = 1.875 \pm 0.282\,744 \implies (1.5923,\ 2.1577)$ or about $(1.5,\ 2.2)$ newspapers.

(b) We see that, when the clustering is ignored, the *point* estimate of the average number of newspapers purchased per household is the *same* and, *un*usually, there is a slight *in*crease in the width of the 90% confidence interval. This *de*crease in the precision of the answer in (ii) is the net result of three factors:

- there is a five- to six-fold *in*crease in the sample standard deviation in (ii) when the clustering is ignored;
- this substantial increase in s compared with s_{ec} is *partially* offset in (ii) by the increase from 4 to 40 in n, which occurs as its *square root* in the *denominators* of the expressions for the estimated standard deviations of the estimators of $\overline{\mathbf{Y}}$; however, $\hat{s.d.}(\bar{Y}_{in})$ in (ii) is still nearly 80% larger than is $\hat{s.d.}(\bar{Y}_{ec})$ in (i);
- the *smaller* critical value of the $t_{\mathrm{n}-1}$ distribution corresponding to the increase in n from 4 in (i) to 40 in (ii) finally results in a 90% confidence interval in (ii) that is only a little less than 30% wider [about 0.57 newspapers] than the interval in (i) [about 0.44 newspapers].

From the (slight) *de*crease in precision of the answer in (ii) where clustering is ignored, we infer that the clusters in this question are somewhat *more* heterogeneous than the (respondent) population with respect to the numbers of newspapers purchased by households. There is support for this inference in the data table in the statement of the question – we see substantial variation of the y_js with*in* each of the four selected clusters (*heterogeneity within clusters*), but these clusters are have relatively *similar* averages (*homogeneity among clusters*). We recall from Figure 2.14 that these are the conditions (*un*usual in practice) that *favour* the precision of selecting clusters compared with selecting individual elements.

T6 – 2. (a) The symbols have the following meanings:

- R is a random variable representing the ratio of the average of the Y_js to the average of the X_js in a sample obtained by EPS;
- $\mathbf{R}$ is a respondent population attribute – the ratio of the average of a response variate $\mathbf{Y}$ to the average of another response variate $\mathbf{X}$ – *i.e.*, $\mathbf{R} = \overline{\mathbf{Y}}/\overline{\mathbf{X}}$;
- n is the number of units in the sample – the *sample size*, a sample attribute;
- $\mathbf{N}$ is the number of elements (or units) in the respondent population – the *respondent population size*, a respondent population attribute;
- $\mathbf{X}_\mathrm{i}$ is the value of a response variate of interest for the ith element (or unit) of the respondent population and $\overline{\mathbf{X}}$ is the average of the $\mathbf{X}_\mathrm{i}$s – $\overline{\mathbf{X}}$ is a respondent population attribute;
- $\mathbf{Y}_\mathrm{i}$ is the value of (another) response variate of interest for the ith element (or unit) of the respondent population and $\overline{\mathbf{Y}}$ is the average of the $\mathbf{Y}_\mathrm{i}$s – $\overline{\mathbf{Y}}$ is a respondent population attribute.

The response variates $\mathbf{Y}$ and $\mathbf{X}$ are also of interest for the *study* population elements (or units) and, often, for the *target* population elements.

(*continued overleaf*)

T6 – 2. (cont.) (b) The *estimating bias* of R is the *difference* between the mean of R and the true value of the population attribute that is estimated by R; *i.e.*, *e.b.*(R) = $E(R) - \mathbf{R}$.

We can think of the estimating bias of R as approximately the difference between the *average* of the estimates obtained in repeated selecting and the true value of $\mathbf{R}$.

NOTE: 1. Saying that an estimator (such as $\overline{Y}$) is *unbiased* for $\overline{\mathbf{Y}}$ means that $E(\overline{Y}) - \overline{\mathbf{Y}} = 0$ or that $E(\overline{Y}) = \overline{\mathbf{Y}}$.

(c) Two factors (mainly) and a third factor (slightly) affect the magnitude of the relative estimating bias:

n, the sample size, appears as its *first power* in the *denominator* of the rightmost brackets of the expression for *r.e.b.*(R); hence, the relative estimating bias gets *smaller* roughly in *direct proportion* to any *in*crease in the sample size – see also the discussion of $\mathbf{N}$ below.

The sample size *is* under the control of the investigator(s) subject (of course) to the practical constraints on the resources available for a particular investigation.

how close the points represented by the population responses ($\mathbf{X}_i$, $\mathbf{Y}_i$) lie to a line with slope $\mathbf{R}$ through the origin – the *closer* they lie to this line, the *smaller* the middle term in brackets inside the summation in the expression for *r.e.b.*(R) and, hence, the *smaller* the estimating bias. The term $\mathbf{X}_i/\overline{\mathbf{X}}$ multiplying the brackets inside the summation means that departures of the $\mathbf{X}_i$ from the line make *more* contribution to *in*creasing the estimating bias of R than do departures of the $\mathbf{Y}_i$, and the larger the magnitude of an $\mathbf{X}_i$, the greater the effect of its departure from the line.

This linearity of the ($\mathbf{X}_i$, $\mathbf{Y}_i$) is *not* under the control of the investigator(s) – it is a characteristic of the (values of the) two response variates $\mathbf{X}$ and $\mathbf{Y}$ over the elements (or units) of the respondent population.

$\mathbf{N}$, the respondent population size, appears in three places in the expression for *r.e.b.*(R):

- in the first *denominator*, the term ($\mathbf{N} - 1$) does *not* confer any real dependence of the magnitude of the relative estimating bias of R on $\mathbf{N}$ because this term is offset by the summation, which contains $\mathbf{N}$ terms and so *in*creases in magnitude as $\mathbf{N}$ increases.
- in the second *denominator* of the rightmost brackets, but the algebraic form of this term means that the influence of the initial 1/n on the estimating bias of R (greatly) exceeds that of $1/\mathbf{N}$, in the usual situation where $n \ll \mathbf{N}$; thus, $\mathbf{N}$ usually has little or no practically important effect on the magnitude of the relative estimating bias of R.
- $\mathbf{N}$ is also the upper index of the summation defined by the sigma sign.

As an attribute of the respondent population, $\mathbf{N}$ is *not* under the control of the investigator(s), except in the sense that it can be made closer to the *study* population size by adequate management of non-response error.

(d) To obtain *accurate* answer(s) by proper management of bias (one part of the statistician's responsibilities), estimator(s) with *low* estimating bias are desirable. What is meant by *low*? This is where the *relative* bias is useful, because *low* means low *relative to* the magnitude of the attribute being estimated, here $\mathbf{R}$. Thus, if $\mathbf{R}$ were to have a large magnitude, the estimating bias of its estimator R could be appreciable in size but still be low *relative to* the magnitude of $\mathbf{R}$.

NOTES: 2. *Relative* (rather than *absolute*) measures of bias (and of precision) are particularly important in the context of estimating *totals*, which may have very large magnitudes.

3. The discussion in (d) is set in the context of *estimating* bias but it applies also to the *other* four categories of bias – studying, non-responding, selecting and measuring (although there is little consideration of these latter four in STAT 332, but see Section 14 on pages 2.48 and 2.49 of Figure 2.3).