Assignment 3

A3 – 1. Some populations (*e.g.*, the records of the number of days absent over a specified time period for each of the employees of a company or organization) consist of a proportion $\mathbf{P}$ of elements with non-zero $\mathbf{Y}$-values and a proportion $\mathbf{Q}$ ($\mathbf{Q} = 1 - \mathbf{P}$) of elements with zero $\mathbf{Y}$-values. It is sometimes possible to identify accurately and cheaply the $\mathbf{NQ}$ elements of the latter type and, if it is desired to estimate $\overline{\mathbf{Y}}$ (the population average) from a sample survey, it may then be more efficient to obtain, by equiprobable selecting (EPS), n units (or elements) from the $\mathbf{N}_{\text{nz}} = \mathbf{NP}$ elements with *non*-zero values, rather than from the whole respondent (or study) population of $\mathbf{N}$ elements.

(a) If $\overline{\mathbf{Y}}_{\text{nz}}$ is the average of the $\mathbf{NP}$ population elements with non-zero $\mathbf{Y}$-values, show that $\overline{\mathbf{Y}} = \mathbf{P}\overline{\mathbf{Y}}_{\text{nz}}$.

(b) If $\mathbf{S}_{\text{nz}}$ is the (data) standard deviation of the $\mathbf{NP}$ population elements with non-zero $\mathbf{Y}$-values, establish the result at the right, where $\mathbf{S}$ is the (whole) population (data) standard deviation.

$$\mathbf{S}_{\text{nz}} = \sqrt{\frac{\mathbf{N}_{\text{nz}}}{\mathbf{N}_{\text{nz}}-1}\left[\frac{\mathbf{N}-1}{\mathbf{N}}\cdot\frac{\mathbf{S}^2}{\mathbf{P}} - \frac{\mathbf{Q}}{\mathbf{P}^2}\overline{\mathbf{Y}}^2\right]}.$$

(c) If $\mathbf{N}_{\text{nz}} \gg 1$, write the expression in (b) for $\mathbf{S}_{\text{nz}}$ in *simpler but approximate* form.

(d) Suppose that $\bar{y}_{\text{nz}}$ is the average of n units, obtained by equiprobable selecting, with non-zero $\mathbf{Y}$-values. Suggest an appropriate estimator of $\overline{\mathbf{Y}}$ based on $\bar{y}_{\text{nz}}$, and give its standard deviation in terms of $\mathbf{N}$, n, $\mathbf{P}$, $\mathbf{Q}$, $\overline{\mathbf{Y}}$ and $\mathbf{S}$. Indicate briefly how this estimator is an improvement over the average of n units, obtained by equiprobable selecting, from *all* $\mathbf{N}$ respondent (or study) population units (or elements).

A3 – 2. *The Globe and Mail* of September 16, 1993, reported the results of a poll which asked the question: *Which leader do you think would make the best Prime Minister of Canada?* Of those who responded, 39% said Kim Campbell and 19% said Jean Chretien. The report contained a common newspaper statement of survey precision: *The results are based on 1,446 personal interviews and are considered accurate within 2.6 percentage points, 19 times in 20.*

(a) Show how the figure of '2.6 percentage points' is obtained and explain briefly what it means.

(b) Would it be correct to say instead '2.6 percent'? Explain briefly.

(c) In assessing the limitations imposed by error of the results of the poll, a student wrote: *The article does not say what economic class the respondents were from or how old they were, but it does not really have to because the poll is accurate and right 19 times in 20.* Comment.

(d) The student also wrote: *The article does not tell us how many people refused the interview but it seems it does not matter – the people who were interviewed and their age and economic status must be positively correlated with regard to the final results because the poll is 'on the nose' 19 times in 20.* Comment.

A3 – 3. A dentist is interested in assessing the effectiveness of a new toothpaste, and so she obtains permission to involve a group of $\mathbf{N} = 1{,}000$ school children in an investigation. Pre-investigation records of *all* the children show an average of 2.2 cavities every six months. After three months of the investigation, the dentist obtains, by equiprobable selecting, 10 children from the group, and obtains the results tabulated above. From an analysis of these data, answer the dentist's question about the toothpaste; show your calculations and reasoning clearly.

Child (i)	1	2	3	4	5	6	7	8	9	10
Number of cavities in 3 months (y_i)	0	4	2	3	2	0	3	4	1	1 .

A3 – 4. One hundred households, obtained by equiprobable selecting from a village with $\mathbf{N} =$ 850 households, showed 30 households possessing transistor radios; further, a complete record of the number of members ($\mathbf{Y}$) in these 30 households gave the data shown above.

$$\sum_{j=1}^{n} y_j = 140 \quad , \quad \sum_{j=1}^{n} y_j^2 = 976$$

(a) Estimate both the proportion and number of households in the village which possess transistor radios, and give an approximate 99% confidence interval for each estimate.

(b) Estimate both the average size, and its standard deviation, of households in the village possessing transistor radios. What assumption about one of the estimates from (a) is needed in the calculations in (b)?

A3 – 5. This question refers to the article EM9514 *Soveeignty a winner, poll finds, if Quebec gets link to Canada*, which is reprinted in Figure 2.2c on page 2.17 of the Course Materials.

In the middle and right-hand columns on the first side of Figure 2.2c, two non-overlapping 'half-samples' of 501 people are mentioned, with those in the first half-sample being asked about sovereignty and those in the second asked about sovereignty-association. From the information given in the article, find approximate 90% confidence intervals for:

(a) the proportion of Quebeckers in the respondent population who are in favour of sovereignty;

(b) the proportion of Quebeckers in the respondent population who are in favour of sovereignty-association;

(c) the *difference* [(b) – (a)] in the two proportions in (a) and (b).

(d) Explain briefly whether it is a reasonable Answer from this poll that a greater proportion of Quebeckers support sovereignty-association than support sovereignty.

(*continued overleaf*)

A3 – 6. Suppose that there are more than two candidates for an election. Candidate *A* has the support of a proportion $\mathbf{P}_A$ of decided voters, while candidate *B* has the support of a proportion $\mathbf{P}_B$. We want to estimate the *difference* $\mathbf{P}_A - \mathbf{P}_B$ and to find a confidence interval for this difference.

(a) Define:

$$\mathbf{V}_i = \begin{cases} 1 & \text{if decided voter i intends to vote for candidate } A, \\ -1 & \text{if decided voter i intends to vote for candidate } B, \\ 0 & \text{if decided voter i intends to vote otherwise,} \end{cases}$$

for each of the $\mathbf{N}$ decided voters in the population. What population attribute of $\mathbf{V}_1, \mathbf{V}_2, \ldots.., \mathbf{V}_\mathbf{N}$ is equal to $\mathbf{P}_A - \mathbf{P}_B$? Justify your answer.

(b) For $\mathbf{V}_i$ defined in (a), write the standard deviation $\mathbf{S}$ (at the right) in terms of $\mathbf{P}_A$ and $\mathbf{P}_B$.

$$\mathbf{S} = \sqrt{\frac{1}{\mathbf{N}-1}\sum_{i=1}^{\mathbf{N}}(\mathbf{V}_i - \overline{\mathbf{V}})^2}$$

(c) Suggest an estimator of $\mathbf{P}_A - \mathbf{P}_B$. How does it relate to $V_1, V_2, \ldots.., V_n$, random variables representing the *sample* analogues of the population $\mathbf{V}_i$s?

(d) Based on your answers to (a) and (b), give an expression for the standard deviation of the estimator in (c).

(e) Give an expression for the *estimated* standard deviation corresponding to (d).

(f) Suppose that the Liberals have the support of 36% of 2,500 decided voters, obtained by equiprobable selecting, while the Conservatives have 31% support in this sample. Find an approximate 95% confidence interval for the Liberals' lead over the Conservatives.

(g) Comment briefly on the interpretation of the confidence interval found in (f).

1995-04-20