

Sample Surveys

Meei Ng and Owen Jones

July 2008

1 Introduction

A **census** is a complete enumeration of the population: data are collected from every unit in the population. In a **survey**, a subset of the population, called a **sample**, is taken.

Census are taken at regular but infrequent intervals, e.g. every 5 or 10 years. In between, surveys are used to update results. The selection and estimation procedures of official surveys are almost always based on previous census of that population.

A **probability sample** is such that one can calculate the probability of inclusion in the sample for each unit in the population. Then one can apply probability theory to estimate sampling error and find confidence intervals. Two types of **non-probability samples** are purposive sampling (units are chosen subjectively to be “representative” of the population) and quota sampling (enumerators are instructed to select units to fill a quota, i.e. a prescribed number of units of either sex, in certain age-groups, social classes, etc.) We shall consider only probability sampling.

Advantages of surveying:

- cheaper
- faster to get results
- can cover a wider scope (more questions)
- higher quality of data (can train personnel)
- more reliable

Disadvantages of surveying:

- do not have information on every unit
- subject to sampling error
- “rare” groups can be misrepresented, giving high sampling errors.

Inference may be enumerative or analytical. **Enumerative inference** seeks to describe the population under study, **analytical inference** seeks to explain it. For example, in a population of households, we might attempt to enumerate the mean number of persons per household, total number of persons in the population, mean household income, or the ratio of the total income to the total number of persons (i.e. per capita income). Enumerative inference typically concerns itself with means, totals, proportions and ratios. Analytical inference might seek to regress household income on such variables as number of employed adults, educational level of household head and variables indicative of geographical location. We shall be concentrating on enumerative inference.

1.1 Terminology

The broad aim of sample surveying is to obtain information about a population.

In a sample survey, there are two statistical aspects: the **selection procedure** that describes how the sample is to be selected and the **estimation procedure** that describes how various parameters are to be estimated and their sampling errors estimated. These two procedures constitute the **sampling strategy**.

- The **target population** is the entire set of individuals about which we require information. For example, all 18 year olds in Australia.
- The **study population** or **sampled population** is the basic finite set of individuals from which a sample is drawn. This may be the target population, or may be a *more limited, more accessible* population whose properties we hope can be extrapolated to the larger target population. For example, all 18 year olds in metropolitan Melbourne. Statistical inference can only be applied to the sampled population. Its extrapolation to the target population is a matter of judgement.
- The individual members of the population are called **elementary units** or just **units**.
- Often, we do not sample the elementary units directly, because it is too expensive to make a full list of them. Instead, we may make a list of larger units to which the elementary units belong (e.g. households, schools, blocks of buildings). These are called **sampling units**. The full set of sampling units should cover the whole sampled population and there should be no overlap. They are also called **enumeration units** or **listing units**.
- A **sampling frame** is a list of all the sampling units.

1.2 Notation

- N denotes the *population size*.
- X, Y and Z denote some variables of interest. X_i, Y_i and Z_i represent their values for unit $i, i = 1, 2, \dots, N$.

Note that these are variables but *not* random variables.

- The *population aggregate* or *total* for these variables are denoted by τ_X, τ_Y and τ_Z respectively. That is,

$$\tau_X = \sum_{i=1}^N X_i \quad \tau_Y = \sum_{i=1}^N Y_i \quad \tau_Z = \sum_{i=1}^N Z_i$$

- The *population mean* for these variables are denoted by μ_X, μ_Y and μ_Z respectively. That is,

$$\mu_X = \frac{1}{N} \sum_{i=1}^N X_i \quad \mu_Y = \frac{1}{N} \sum_{i=1}^N Y_i \quad \mu_Z = \frac{1}{N} \sum_{i=1}^N Z_i$$

- The *proportion of a population* having a particular attribute is denoted by P . It can be regarded as a population mean for a 0-1 variable:

$$P = \mu_Z = \frac{1}{N} \sum_{i=1}^N Z_i$$

where $Z_i = 1$ if unit i has the attribute and 0 otherwise.

- The *population variance* of Y is

$$\sigma_Y^2 = \frac{1}{N} \sum_{i=1}^N (Y_i - \mu_Y)^2 = \frac{1}{N} \sum_{i=1}^N Y_i^2 - \mu_Y^2.$$

Its square root is called the *standard deviation* of Y .

- A related measure is

$$S_Y^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - \mu_Y)^2 = \frac{N}{N-1} \sigma_Y^2.$$

The use of S_Y^2 and similar quantities simplifies the algebraic expressions for sample estimates and sampling errors. For this reason, it is sometimes suggested that S_Y^2 should be called the population variance of Y . This is not universally accepted. We shall use S_Y^2 whenever convenient, but retain σ_Y^2 as the population variance.

- *Coefficient of variation, CV:*

$$CV_Y = \frac{\sigma_Y}{\mu_Y}$$

It gives a measure of variability that is independent of the unit of measurement.

- The *population covariance* of two variables X and Y is

$$\sigma_{XY} = \frac{1}{N} \sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y) = \frac{1}{N} \sum_{i=1}^N X_i Y_i - \mu_X \mu_Y = \sigma_{YX}.$$

We similarly introduce

$$S_{XY} = \frac{1}{N-1} \sum_{i=1}^N (X_i - \mu_X)(Y_i - \mu_Y) = \frac{N}{N-1} \sigma_{XY}.$$

- The *correlation coefficient* of X and Y is

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{S_{XY}}{S_X S_Y}.$$

This is a dimensionless number between -1 and 1 .

- The *ratio* of two population totals or means is

$$\theta = \frac{\tau_Y}{\tau_X} = \frac{\mu_Y}{\mu_X}.$$

Example: if the sampling unit = household, Y = total expenditure per week, and X = household size, then θ = weekly expenditure per capita.

- We use n to denote a sample size. The fraction, $f = n/N$ is called the *sampling fraction*.

1.3 References

- Scheaffer, R.L., Mendenhall, W. and Ott, L (1990) *Elementary Survey sampling, 4th Edition*, PWS-Kent.
- Som, R.K. (1996) *Practical Sampling Techniques*, Marcel Dekker.
- Foreman E.K. (1991) *Survey Sampling Principles*, Marcel Dekker.
- Rao, P.S.R.S (2000) *Sampling Methodologies with Applications*, Chapman & Hall/CRC
- Cochran, W.G. (1977) *Sampling Techniques, 3rd Edition*, Wiley.

2 Simple Random Samples

A **sample random sample** (SRS) is one drawn in such a way that all possible samples of the same size have the same probability of being the selected sample. We assume that units are drawn *without replacement*, i.e. the same unit may not appear more than once.

Total number of possible samples = $\binom{N}{n}$ where n = sample size. The probability that a particular sample is selected = $1/\binom{N}{n}$.

2.1 Sample notation

- $\{y_1, y_2, \dots, y_n\}$ and $\{x_1, x_2, \dots, x_n\}$ denote the observations of Y and X variables on the units in the sample.

- Sample means:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

When we want to indicate the sample size used, we write \bar{y}_n , etc.

- Sample variances:

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2, \quad s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- Sample covariance:

$$s_{xy} = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x}).$$

2.2 Estimation of μ_Y

We estimate μ_Y using $\hat{\mu}_Y = \bar{y}_n$.

- $E(\hat{\mu}_Y) = \mu_Y$ (unbiased)
- $\text{Var}(\hat{\mu}_Y) = \frac{S_Y^2}{n}(1-f)$.

We estimate S_Y^2 using $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$. We show below that it is unbiased. Hence, an unbiased estimator of the variance of \bar{y} is

- $\hat{\text{Var}}(\bar{y}) = \frac{s_y^2}{n}(1-f)$.

There are three chief uses for $\hat{\text{Var}}(\hat{\mu}_Y)$:

- to compare the precision of SRS with other methods of sampling;
- to estimate the sample size needed in a planned survey;
- to estimate the precision attained by the completed survey (e.g. error bounds and confidence intervals)

We give an example then the derivations.

2.3 Example

A small village has six households:

Household:	a	b	c	d	e	f
Household size:	8	6	3	5	4	4

For a sample of size 2 the possible distinct SRS are

$$\begin{aligned}
 &\{a, b\}, \quad \{a, c\}, \quad \{a, d\}, \quad \{a, e\}, \quad \{a, f\} \\
 &\{b, c\}, \quad \{b, d\}, \quad \{b, e\}, \quad \{b, f\}, \quad \{c, d\} \\
 &\{c, e\}, \quad \{c, f\}, \quad \{d, e\}, \quad \{d, f\}, \quad \{e, f\}.
 \end{aligned}$$

In practice, the units are drawn one at a time, with each unit already drawn *removed* from the population before the next one is drawn. Under the selection procedure,

$$\Pr(\text{a particular sample is drawn}) = \frac{1}{N} \frac{1}{N-1} \cdots \frac{1}{N-n+1} \times n! = \frac{1}{\binom{N}{n}}.$$

Let Y = household size. The populations values of Y are: 8, 6, 3, 5, 4, 4.

$$\begin{aligned}
 \mu_Y &= \frac{8 + 6 + 3 + 5 + 4 + 4}{6} = 5 \\
 \sigma_Y^2 &= \frac{1}{6} \sum_{i=1}^6 (Y_i - \mu_Y)^2 = \frac{9 + 1 + 4 + 0 + 1 + 1}{6} = \frac{16}{6} = 2\frac{2}{3} \\
 S_Y^2 &= \frac{1}{5} \sum_{i=1}^6 (Y_i - \mu_Y)^2 = \frac{16}{5} = 3.2
 \end{aligned}$$

Now we take a sample of size 2:

All possible samples of size 2 from Small Village

Sample no.	Sample	$\{y_1, y_2\}$	\bar{y}	s_y^2	$\hat{\text{Var}}(\bar{y})$
1	{a, b}	{8, 6}	7	2.0	0.6667
2	{a, c}	{8, 3}	5.5	12.5	4.1667
3	{a, d}	{8, 5}	6.5	4.5	1.5000
4	{a, e}	{8, 4}	6	8.0	2.6667
5	{a, f}	{8, 4}	6	8.0	2.6667
6	{b, c}	{6, 3}	4.5	4.5	1.5000
7	{b, d}	{6, 5}	5.5	0.5	0.1667
8	{b, e}	{6, 4}	5	2.0	0.6667
9	{b, f}	{6, 4}	5	2.0	0.6667
10	{c, d}	{3, 5}	4	2.0	0.6667
11	{c, e}	{3, 4}	3.5	0.5	0.1667
12	{c, f}	{3, 4}	3.5	0.5	0.1667
13	{d, e}	{5, 4}	4.5	0.5	0.1667
14	{d, f}	{5, 4}	4.5	0.5	0.1667
15	{e, f}	{4, 4}	4	0	0
mean			5.0	3.2	16/15
variance			16/15		

We note that

- Our estimates depend on the sample.

- On average,

$$E(\bar{y}) = \frac{7 + 5.5 + \cdots + 4}{15} = 5.0 = \mu_Y.$$

That is, it is unbiased, confirming the theoretical result.

- The variance of \bar{y} is, by direct evaluation,

$$\begin{aligned} \text{Var}(\bar{y}) &= E[(\bar{y} - E(\bar{y}))^2] = E[(\bar{y} - 5.0)]^2 \\ &= \frac{2^2 + .5^2 + \cdots + 1^2}{15} = \frac{16}{15} \end{aligned}$$

From theoretical result,

$$\text{Var}(\bar{y}) = \frac{S_Y^2}{n}(1 - f) = \frac{3.2}{2}\left(1 - \frac{2}{6}\right) = \frac{16}{15}$$

Thus, the result is verified.

- The mean of the values in the last column is $16/15$, which is the value of $\text{Var}(\bar{y})$ calculated above, showing that $\hat{\text{Var}}(\bar{y})$ is an unbiased variance estimator.

2.4 Derivations

For the population,

$$\begin{aligned} \mu_Y &= \frac{1}{N} \sum_{\ell=1}^N Y_\ell \\ \sigma_Y^2 &= \frac{1}{N} \sum_{\ell=1}^N (Y_\ell - \mu_Y)^2 = \frac{1}{N} \left(\sum_{\ell=1}^N Y_\ell^2 - N\mu_Y^2 \right) = \frac{1}{N} \sum_{\ell=1}^N Y_\ell^2 - \mu_Y^2. \end{aligned}$$

Hence,

$$\sigma_Y^2 + \mu_Y^2 = \frac{1}{N} \sum_{\ell=1}^N Y_\ell^2 \tag{1}$$

$$\mu_Y^2 = \left(\frac{1}{N} \sum_{\ell=1}^N Y_\ell \right)^2 = \frac{1}{N^2} \left(\sum_{\ell=1}^N Y_\ell^2 + 2 \sum_{\ell < m} Y_\ell Y_m \right)$$

$$\begin{aligned} 2 \sum_{\ell < m} Y_\ell Y_m &= N^2 \mu_Y^2 - \sum_{\ell=1}^N Y_\ell^2 = N^2 \mu_Y^2 - N(\sigma_Y^2 + \mu_Y^2) \\ &= N(N-1)\mu_Y^2 - N\sigma_Y^2 \end{aligned} \tag{2}$$

Indicator variable: let $I_\ell = 1$ if Y_ℓ is in the sample and 0 otherwise. Then I_ℓ is a random variable and

$$\begin{aligned} \bar{y} &= \frac{1}{n} \sum_{\ell=1}^N Y_\ell I_\ell. \\ s_y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right) \\ &= \frac{1}{n-1} \sum_{i=1}^n y_i^2 - \frac{n}{n-1} \bar{y}^2 \\ &= \frac{1}{n-1} \sum_{i=1}^N Y_\ell^2 I_\ell - \frac{n}{n-1} \left(\frac{1}{n} \sum_{\ell=1}^N Y_\ell I_\ell \right)^2 \end{aligned}$$

First order inclusion probabilities:

$$\Pr(I_\ell = 1) = \frac{n}{N}, \text{ by symmetry, whence } E(I_\ell) = \frac{n}{N}.$$

Second order inclusion probabilities:

$$\Pr(I_\ell = 1, I_m = 1) = \binom{N-2}{n-2} / \binom{N}{n} = \frac{n(n-1)}{N(N-1)}, \quad \text{for } \ell \neq m.$$

Hence,

$$E(I_\ell I_m) = \frac{n(n-1)}{N(N-1)}, \quad \text{for } \ell \neq m.$$

2.4.1 Calculating $E(\bar{y})$

$$\begin{aligned} E(\bar{y}) &= E\left(\frac{1}{n} \sum_{\ell=1}^N Y_\ell I_\ell\right) = \frac{1}{n} \sum_{\ell=1}^N Y_\ell E(I_\ell) \\ &= \frac{1}{n} \sum_{\ell=1}^N Y_\ell \frac{n}{N} = \frac{1}{N} \sum_{\ell=1}^N Y_\ell = \mu_Y \quad (\text{Q.E.D.}) \end{aligned}$$

2.4.2 Calculating $\text{Var}(\bar{y})$

$$\bar{y}^2 = \left(\frac{1}{n} \sum_{\ell=1}^N Y_\ell I_\ell\right)^2 = \frac{1}{n^2} \left(\sum_{\ell=1}^N Y_\ell^2 I_\ell + 2 \sum_{\ell < m} Y_\ell Y_m I_\ell I_m\right).$$

Hence,

$$\begin{aligned} E(\bar{y}^2) &= \frac{1}{n^2} \left(\sum_{\ell=1}^N Y_\ell^2 E(I_\ell) + 2 \sum_{\ell < m} Y_\ell Y_m E(I_\ell I_m)\right) \\ &= \frac{1}{n^2} \left(\sum_{\ell=1}^N Y_\ell^2 \frac{n}{N} + 2 \sum_{\ell < m} Y_\ell Y_m \frac{n(n-1)}{N(N-1)}\right) \end{aligned} \quad (3)$$

Substituting (1) and (2) into (3), we have

$$\begin{aligned} E(\bar{y}^2) &= \frac{1}{n}(\sigma_Y^2 + \mu_Y^2) + \frac{(n-1)}{nN(N-1)}[N(N-1)\mu_Y^2 - N\sigma_Y^2] \\ &= \frac{1}{n}(\sigma_Y^2 + \mu_Y^2) + \frac{n-1}{n}\mu_Y^2 - \frac{n-1}{n(N-1)}\sigma_Y^2 \\ &= \mu_Y^2 + \frac{N-n}{n(N-1)}\sigma_Y^2 \end{aligned}$$

Thus

$$\begin{aligned} \text{Var}(\bar{y}) &= E(\bar{y}^2) - E(\bar{y})^2 = \mu_Y^2 + \frac{N-n}{n(N-1)}\sigma_Y^2 - \mu_Y^2 \\ &= \frac{N\sigma_Y^2}{n(N-1)} \frac{N-n}{N} = \frac{S_Y^2}{n}(1-f) \quad (\text{Q.E.D.}) \end{aligned} \quad (4)$$

2.4.3 Calculating $E(s_Y^2)$

$$\begin{aligned} s_Y^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n-1} \left(\sum_{i=1}^n y_i^2 - n\bar{y}^2 \right) \\ &= \frac{1}{n-1} \sum_{\ell=1}^N Y_\ell^2 I_\ell - \frac{n}{n-1} \bar{y}^2 \end{aligned}$$

Hence,

$$\begin{aligned} E(s_Y^2) &= \frac{1}{n-1} \sum_{\ell=1}^N Y_\ell^2 E(I_\ell) - \frac{n}{n-1} E(\bar{y}^2) \\ &= \frac{1}{n-1} \sum_{\ell=1}^N Y_\ell^2 \frac{n}{N} - \frac{n}{n-1} E(\bar{y}^2) \\ &= \frac{n}{n-1} (\sigma_Y^2 + \mu_Y^2) - \frac{n}{n-1} \left(\mu_Y^2 + \frac{N-n}{n(N-1)} \sigma_Y^2 \right) \\ &= \frac{n}{n-1} \left(1 - \frac{N-n}{n(N-1)} \right) \sigma_Y^2 = \frac{N}{N-1} \sigma_Y^2 = S_Y^2 \quad (\text{Q.E.D.}) \end{aligned}$$

2.5 Sampling with replacement

Suppose that we sample *with* replacement. In this case the successive draws are independent, so we can use our knowledge of independent random variables.

- y_1, y_2, \dots, y_n are iidrv, with

$$E(y_i) = \mu_Y \quad \text{and} \quad \text{Var}(y_i) = \sigma_Y^2.$$

-

$$E(\bar{y}) = E(y_i) = \mu_Y.$$

-

$$\text{Var}(\bar{y}) = \frac{\sigma_Y^2}{n}.$$

-

$$E(s_y^2) = \sigma_Y^2, \quad \text{for} \quad s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

We note that for sampling without replacement $\hat{\text{Var}}(\bar{y}) = (s_y^2/n)(1-f)$ while for sampling with replacement we just have s_y^2/n . The factor $1-f$ is called the *finite population correction* (fpc). As $1-f < 1$, sampling without replacement is more efficient than sampling with replacement, especially when the sampling fraction f is relatively large. However, computation of population estimates is greatly simplified for sampling with replacement. For this reason, even if sampling is done without replacement, the sample is sometimes treated as if it was made with replacement provided the sample size is small relative to the population, e.g. $f < 0.1$, or preferably, < 0.05 .

2.6 Confidence intervals

There is a central limit theorem for large samples (Hajek, 1960). Under certain conditions, as $N \rightarrow \infty$, $n \rightarrow \infty$ and $N - n \rightarrow \infty$, \bar{y} tends to a normal distribution. See Cochran §2.13 for a precise statement. See also Sugden, Smith and Jones, 2000.

Thus, a $100(1 - \alpha)\%$ confidence interval for μ_Y is

$$\bar{y}_n \pm z_{1-\alpha/2} \times \text{s.e.}(\bar{y}_n)$$

where $z_{1-\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the $N(0, 1)$ distribution.

If the sample size is less than 50, the percentiles may be taken from a t -distribution with $(n - 1)$ degrees of freedom. The t approximation holds exactly only if the observations y_i are normally distributed and N is infinite. Moderate departures from these conditions do not affect the results greatly.

For small samples with very skewed distribution, special methods are needed.

2.7 Determining sample size

The larger the sample size, the more precise the estimates. On the other hand, too large a size is wasting resources. The required sample size is determined by

- the sampling strategy (sampling design and estimation);
- the required precision;
- the variability of the population.

We shall consider SRS and estimation of mean, and suppose that the required precision is expressed as the variance of \bar{y} or equivalently the width of a confidence interval. If we require $\text{Var } \bar{y} = D^2$ then rearranging (4) one obtains

$$\begin{aligned} D^2 &= \frac{N\sigma_Y^2}{n(N-1)} \frac{N-n}{N} \\ n(N-1)D^2 &= \sigma_Y^2(N-n) \\ n((N-1)D^2 + \sigma_Y^2) &= N\sigma_Y^2 \\ n &= \frac{N\sigma_Y^2}{(N-1)D^2 + \sigma_Y^2} \end{aligned}$$

The variability of the population, σ_Y^2 , is usually estimated from a previous study or some similar study done elsewhere. If the range of the values of Y is known (or can be guessed), then

$$\sigma_Y = \frac{\text{range}}{4}$$

is sometimes used as a value for calculating sample size. Alternatively note that n increases as N increases, so an upper bound on the required sample size can be obtained by sending $N \rightarrow \infty$, to get $n_0 = \sigma_Y^2/D^2$. (You can think of this as ignoring the fpc.)

2.8 Exercises

1. A simple random sample of 30 households was drawn from a city area comprising 14,848 households. The numbers of persons per household in the sample were as follows:

5 6 3 3 2 3 3 3 4 4 3 2 7 4 3
5 4 4 3 3 4 3 3 1 2 4 3 4 2 4

Estimate the total number of people in the area and compute the probability that this estimate is within $\pm 10\%$ of the true value.

2. Signatures to a petition were collected on 676 sheets. Each sheet had enough space for 42 signatures, but on many sheets a smaller number of signatures had been collected. The number of signatures per sheet were counted on a random sample of 50 sheets (about a 7% sample), with the result shown below.

Signatures	42	41	36	32	29	27	23	19	16	15	14
Frequency	23	4	1	1	1	2	1	1	2	2	1
Signatures	11	10	9	7	6	5	4	3			
Frequency	1	1	1	1	3	2	1	1			

Estimate the number of signatures to the petition and give 80% confidence limits.

How large must the sample be if the total number of signatures is to be estimated with a margin of error of 1000, apart from a 1 in 20 chance? Assume that the population variance is the value of s^2 in the sample.

3. Let P be the proportion of a population with a particular attribute. Using a SRS, we estimate P using the sample proportion p . Show that

$$\text{Var}(p) = \frac{1}{n} \frac{NP(1-P)}{N-1} (1-f)$$

and that

$$\hat{\text{Var}}(p) := \frac{p(1-p)}{n-1} (1-f)$$

is an unbiased estimator of $\text{Var}(p)$.

Hint: let $Y_i = 1$ if unit i has the attribute and 0 otherwise.

4. (a) If $N = 400$, $\mu_X \approx 20$ and $\sigma_X \approx 10$, how large a sample is required to estimate μ_X within 10%, apart from a chance of about one in twenty?
- (b) In a simple random sample of 200 from a population of 2000, 120 were in favour of a particular proposal. Specify approximate 95% confidence limits for the proportion of the population in favour of the proposal.
5. In a district containing 4000 houses the percentage of owned houses is to be estimated with a s.e. of not more than 2% and the percentage of two-car households with a s.e. of not more than 1%. (The figures 2 and 1% are the absolute values, not the coefficients of variation.) The true percentage of owners is thought to lie between 45 and 65% and the percentage of two-car households between 5 and 10%. How large a sample is necessary to satisfy both of them?
6. For a population of $N = 5$ elements, the values taken by the variable Y are 8, 3, 1, 11, 4 and 7. By considering all possible samples of $n = 3$ elements, verify that for simple random sampling
- $E(\bar{Y}) = \mu_Y$;
 - $\text{Var}(\bar{Y}) = \frac{1}{n} \frac{N}{N-1} \sigma^2 (1-f)$;
 - $E(s^2) = \frac{N}{N-1} \sigma^2$.

3 Systematic Sampling

How to take a systematic sample. Suppose $N/n = k$ is an integer.

- Arrange the population units in a sequence, e.g. natural order of arrival or in space; on a list; each given a number.
- Choose a random starting number r from $\{1, 2, \dots, k\}$.
- Select the r -th unit and every k -th one after it, i.e. units corresponding to $r, r + k, r + 2k, \dots, r + (n - 1)k$.

k is called the *sampling interval*. r is called the *random start*.

Possible approaches if $N/n = k$ is not an integer:

- Stop when the list ends (Foreman, 1991). This would result in some systematic samples having 1 unit less than other samples. As the starting point is random, the sample size is then random and has to be factored into the analysis.
- Arrange the units of the population around a circle. Choose a random starting number r from $\{1, \dots, N\}$. Select unit r and every k -th after it until n units are obtained. (Cochran, 1977).
- (ABS manual for Statistical training: Survey Methods 1)
 - Divide the interval $[0, n)$ into N equal subintervals and allocate each to a unit in the population.
 - Choose a uniformly distributed random number r from $[0, 1)$.
 - Select the population units that correspond to the subintervals that contain $r, r + 1, r + 2, \dots$

For simplicity we will assume that $N/n = k$ is an integer.

3.1 Estimation of μ_Y

An estimator for μ_Y is the sample mean (what else?):

$$\hat{\mu}_{sy} = \bar{y}.$$

To investigate the sampling distribution of $\hat{\mu}_{sy}$, we arrange the population units thus:

	Sample number					
	1	2	...	i	...	k
	y_1	y_2	...	y_i	...	y_k
	y_{k+1}	y_{k+2}	...	y_{k+i}	...	y_{2k}

	$y_{(n-1)k+1}$	$y_{(n-1)k+2}$...	$y_{(n-1)k+i}$...	y_{nk}
Means	\bar{y}_1	\bar{y}_2	...	\bar{y}_i	...	\bar{y}_k

Each column is a possible systematic sample. The last row of sample means constitute the distribution of $\hat{\mu}_{sy}$, with probability $1/k$ for each.

Let y_{ij} denote the j -th member of the i -th systematic sample.

3.1.1 $E(\hat{\mu}_{sy}) = \mu_Y$.

$$\begin{aligned} E(\hat{\mu}_{sy}) &= \frac{1}{k} \sum_{i=1}^k \bar{y}_i = \frac{1}{k} \sum_{i=1}^k \left(\frac{1}{n} \sum_{j=1}^n y_{ij} \right) \\ &= \frac{1}{nk} \sum_{i=1}^k \sum_{j=1}^n y_{ij} \\ &= \mu_Y. \end{aligned}$$

3.1.2 $\text{Var}(\hat{\mu}_{sy}) = \frac{1}{k} \sum_{i=1}^k (\bar{y}_i - \mu_Y)^2$.

This is just the usual definition of the variance of a distribution.

Using the sum-of-squares decomposition it can be shown that

$$\text{Var}(\hat{\mu}_{sy}) = \frac{N-1}{N} S^2 - \frac{N-k}{N} S_{wsy}^2 \quad (5)$$

where S_{wsy}^2 is the *within samples* sum of squares:

$$S_{wsy}^2 = \frac{1}{N-k} \sum_{i=1}^k \sum_{j=1}^n (y_{ij} - \bar{y}_i)^2.$$

(And $\text{Var}(\hat{\mu}_{sy})$ is the *between samples* sum of squares.)

3.1.3 $\hat{\text{Var}}(\hat{\mu}_{sy})$

Estimation of the variance is not possible, as we only have one observation of $\hat{\mu}_{sy}$. In practice, the SRS variance estimator is used:

$$\hat{\text{Var}}(\hat{\mu}_{sy}) = \frac{s_y^2}{n} (1-f).$$

This can be badly biased.

The performance of a systematic sample, compared with a SRS, will depend on the ordering of the population.

Let $\hat{\mu}_{srs}$ be our estimator under the assumption that the sample is a SRS. (The estimator is still given by \bar{y} , but its distribution changes depending on how the sample is collected.) Taking N large, so that $f \approx 0$ and $(N-1)/N \approx 1$, we see from equation (5) that

$$\text{Var}(\hat{\mu}_{sy}) < \text{Var}(\hat{\mu}_{srs}) \iff S_{wsy}^2 > S^2.$$

- If the population is in *random order*, then $\bar{y}_i \approx \bar{y}$ so $S_{wsy}^2 \approx S^2$ and

$$\text{Var}(\hat{\mu}_{sy}) \approx \text{Var}(\hat{\mu}_{srs}).$$

This justifies the using $\hat{\text{Var}}(\hat{\mu}_{srs})$ to approximate $\hat{\text{Var}}(\hat{\mu}_{sy})$.

- If the population is *ordered* more or less according to the values of Y_i , then a systematic sample is heterogeneous and $S_{wsy}^2 > S^2$. Thus,

$$\text{Var}(\hat{\mu}_{sy}) < \text{Var}(\hat{\mu}_{srs}).$$

That is, a systematic sample is more precise than a SRS in this case.

- If the population is *periodic*, and the period is in sync with the sampling interval k , then a systematic sample will be homogeneous and $S_{wsy}^2 < S^2$. Thus,

$$\text{Var}(\hat{\mu}_{sy}) > \text{Var}(\hat{\mu}_{srs}).$$

So in this case a systematic sample is less precise than a SRS.

3.2 Exercises

1. A population of 360 households (numbered 1 to 360) in Baltimore is arranged alphabetically in a file by the surname of the head of the household. Households in which the head is non-white occur at the following numbers:

28, 31–33, 36–41, 44, 45, 47, 55, 56, 58, 68, 69, 82, 83, 85,
86, 89–94, 98, 99, 101, 107–110, 114, 154, 156, 178, 223, 224, 296,
298–300, 302–304, 306–323, 325–331, 333, 335–339, 341, 342.

(The nonwhite households show some “clumping” because of an association between surname and colour.)

Compare the precision of a 1-in-8 systematic sample with a SRS of the same size for estimating the proportion of households in which the head is nonwhite.

Note that to calculate the variance of the systematic sample you will need to consider all 8 possible samples. It would be a good idea to put the data in a spreadsheet and arrange the population into columns and rows as we did in the lecture notes.

2. A neighbourhood contains three compact communities, consisting, respectively, of people of Anglo-Saxon, Polish and Italian descent. There is an up-to-date directory. In it the persons in a household are listed in the following order: husband, wife, children (by age), others. Houses are listed in order along streets. The average number of persons per house is five.

The choice is between a systematic sample of every fifth person in the directory and a 20% SRS. For which of the following variables do you expect the systematic sample to be more precise?

- Proportion of people of Polish descent,
- proportion of males,
- proportion of children.

Give reasons.

4 Stratified Random Sampling

Suppose the population is divided into non-overlapping subpopulations, called *strata*. Assuming that we know the size N_h of each stratum h , how do we allow for this structure when we try to estimate the population mean?

If an SRS is taken from each stratum, independently, the sampling procedure is called *stratified random sampling*.

4.1 Notation

For stratum $h = 1, 2, \dots, L$,

Population:	N_h = subpopulation size, μ_h = subpopulation mean (of some variable), τ_h = subpopulation total (of above variable), σ_h^2 = subpopulation variance (of above variable), $S_h^2 = N_h \sigma_h^2 / (N_h - 1)$, P_h = subpopulation proportion (having some attribute), $W_h = N_h / N$ stratum weight,
Sample:	n_h = sample size, \bar{y}_h = sample mean (of above variable), s_h^2 = sample variance (of above variable), p_h = sample proportion (having above attribute), $f_h = n_h / N_h$ sampling fraction in stratum.

As before, μ , τ , P , N etc. refer to the mean, total, proportion, size etc. of the whole population, and \bar{y} , n , etc. refer to the mean, size, etc. of the whole sample.

4.2 Estimation

We are interested in estimating μ , the whole population mean. (Estimators for τ and P follow directly from an estimator for μ .)

Within each stratum, the mean μ_h can be estimated as before:

Estimator of μ_h is	$\hat{\mu}_h = \bar{y}_h$ (unbiased).
Its variance is	$\text{Var}(\bar{y}_h) = \frac{S_h^2}{n_h} (1 - f_h)$.
Variance estimator is	$\hat{\text{Var}}(\bar{y}_h) = \frac{s_h^2}{n_h} (1 - f_h)$ (unbiased).

The whole population mean μ is estimated by the *weighted mean*

$$\hat{\mu}_{st} = \frac{1}{N} \sum_{h=1}^L N_h \bar{y}_h = \sum_{h=1}^L W_h \bar{y}_h$$

Note that this is not the sample mean of the whole sample, which is

$$\bar{y} = \frac{1}{n} \sum_{h=1}^L n_h \bar{y}_h.$$

If in every stratum,

$$\frac{n_h}{n} = \frac{N_h}{N}$$

then, $\hat{\mu}_{st} = \bar{y}$. This is described as stratification with *proportional allocation* of n_h . It results in a *self-weighting* sample, which can be very convenient.

4.2.1 Properties of $\hat{\mu}_{st}$

- $E(\hat{\mu}_{st}) = \mu$, therefore unbiased.
Note that $E\bar{y} \neq \mu$ in general.
- $\text{Var}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 \text{Var}(\bar{y}_h) = \sum_{h=1}^L W_h^2 S_h^2(1 - f_h)/n_h$.
- $\hat{\text{Var}}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 s_h^2(1 - f_h)/n_h$.
This is unbiased since s_h^2 is an unbiased estimator of S_h^2 .

4.2.2 Ignoring finite population

If sampling fraction $f_h = n_h/N_h$ is negligible in all strata, then

$$\text{Var}(\hat{\mu}_{st}) = \sum_{h=1}^L \frac{W_h^2 S_h^2}{n_h};$$

and

$$\hat{\text{Var}}(\hat{\mu}_{st}) = \sum_{h=1}^L \frac{W_h^2 s_h^2}{n_h}.$$

4.2.3 Proportional allocation

If $n_h/n = N_h/N$ then

$$n_h = n \times \frac{N_h}{N} = nW_h \quad \text{and} \quad f_h = \frac{n_h}{N_h} = \frac{n}{N} = f$$

so

$$\text{Var}_{prop}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{nW_h}(1 - f) = \frac{1 - f}{n} \sum_{h=1}^L W_h S_h^2.$$

4.2.4 Confidence interval

A $100(1 - \alpha)\%$ confidence interval for μ is

$$\hat{\mu}_{st} \pm t_{n-L}(1 - \alpha/2) \times s.e.(\hat{\mu}_{st})$$

where the factor $t_{n-L}(1 - \alpha/2)$ is from the student- t distribution with $n - L$ degrees of freedom. This assumes that each \bar{y}_h is approximately normally distributed. We lose one degree of freedom each time we substitute \bar{y}_h for μ_h , when using s_h^2 to approximate S_h^2 .

4.3 Example (Scheaffer et al, Ex 5.1 and 5.2)

An advertising firm, interested in determining how much to emphasise television in a certain county, decides to conduct a sample survey to estimate the average number of hours per week that households within the county watch TV. The county contains two towns, town A and town B , and a rural area. Town A is built around a factory, and most households contain factory workers with school children. Town B is an exclusive suburb of a city in a neighbouring county and contains older residents with few children at home. There are 155 households in town A , 62 in town B , and 93 in the rural area.

- (a) Discuss the merits of stratified random sampling in this situation.

- (b) Suppose a survey is carried out. The advertising firm has enough money to interview 40 households and decides to select 20 from town *A*, 8 from town *B* and 12 from the rural area. (This is in fact proportional allocation.) The results are shown below.

Television viewing time, in hours per week											
Stratum 1				Stratum 2				Stratum 3			
Town A				Town B				Rural area			
35	28	26	41	27	4	49	10	8	15	21	7
43	29	32	37	15	41	25	30	14	30	20	11
36	25	29	31					12	32	34	24
39	38	40	45								
28	27	35	34								

Estimate the average television viewing time and give a 95% confidence interval, in hours per week, for

- (i) all households in the county,
(ii) all households in town *B*.

Solution:

- (a) – The population falls into three natural groupings according to geographical location. Use these as strata has administrative convenience in selecting samples and carrying out fieldwork.
– Each group is reasonably homogenous and should have similar behaviours. So, we expect variability to be small within a group and large between groups: the ideal situation where precision can be gained by stratification.
– The advertising firm may wish to produce estimates of average TV viewing time of each town separately. Stratification allows this.
- (b) (i) From the data,

	Stratum 1	Stratum 2	Stratum 3
Sample:	$n_1 = 20$	$n_2 = 8$	$n_3 = 12$
	$\bar{y}_1 = 33.900$	$\bar{y}_2 = 25.125$	$\bar{y}_3 = 19.000$
	$s_1^2 = 35.358$	$s_2^2 = 232.411$	$s_3^2 = 87.636$
Population:	$N_1 = 155$	$N_2 = 62$	$N_3 = 93$

Total population size = $N = 155 + 62 + 93 = 310$.

$$\begin{aligned} \hat{\mu}_{st} &= \frac{1}{N}(N_1\bar{y}_1 + N_2\bar{y}_2 + N_3\bar{y}_3) \\ &= \frac{1}{310}[(155)(33.900) + (62)(25.125) + (93)(19.00)] \\ &= 27.675 \end{aligned}$$

Its variance estimate is

$$\begin{aligned} s^2(\hat{\mu}_{st}) &= \sum_{h=1}^3 W_h^2 \frac{s_h^2}{n_h} (1 - f_h) \\ &= \frac{1}{N^2} \sum_{h=1}^3 \frac{N_h^2 (1 - f_h) s_h^2}{n_h} \\ &= \frac{1}{310^2} \left[\frac{(155)^2 (.871)(35.358)}{20} + \frac{(62)^2 (0.871)(232.411)}{8} + \frac{(93)^2 (0.871)(87.636)}{12} \right] \\ &= 1.9675 \end{aligned}$$

For a 95% confidence interval we use $t_{72-3}(0.975) = 2.00$, giving

$$27.675 \pm 2.00 \times \sqrt{1.9675} = (24.870, 30.480).$$

A normal approximation to the student- t distribution would give a very similar answer.

(ii) This is a straight SRS problem.

$$\begin{aligned}\hat{\mu}_2 &= \bar{y}_2 = 25.125 \\ s^2(\hat{\mu}_2) &= \frac{s_2^2}{n_2}(1 - f_2) = \frac{232.411^2}{8}(1 - 8/62) = 10.0\end{aligned}$$

Using $t_7(0.975) = 2.365$, a 95% confidence interval for the average TV viewing time in town B is

$$25.125 \pm 2.365 \times \sqrt{10} = (17.6, 32.6).$$

4.4 Determining sample size

As with SRS, the sample size depends on the required precision and the population variability. In the case of stratified sampling, we also need to know the planned allocation of the sample to the strata. Let

$$w_h = \frac{n_h}{n}.$$

This is the fraction of the sample allocated to stratum h .

Suppose V is the required value of the variance of the estimate of μ . For example, $V = (B/2)^2$ where B is the required error bound, or $V = (W/z_{1-\alpha/2})^2$ where W is the width of a $100(1-\alpha)\%$ confidence interval of μ .

We have

$$\text{Var}(\hat{\mu}_{st}) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h}(1 - f_h) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^L W_h^2 \frac{S_h^2}{N_h}.$$

Substituting $V = \text{Var}(\hat{\mu}_{st})$, $n_h = nw_h$ and $W_h = N_h/N$,

$$\begin{aligned}V &= \sum_{h=1}^L \frac{N_h^2}{N^2} \frac{S_h^2}{nw_h} - \sum_{h=1}^L \frac{N_h^2}{N^2} \frac{S_h^2}{N_h}; \\ N^2V &= \frac{1}{n} \sum_{h=1}^L (N_h^2 S_h^2 / w_h) - \sum_{h=1}^L N_h S_h^2 \\ n &= \frac{\sum_{h=1}^L (N_h^2 S_h^2 / w_h)}{N^2V + \sum_{h=1}^L N_h S_h^2}\end{aligned}$$

4.5 Optimal allocation of sample

The sampler has to choose n_h , or rather, the weighting $w_h = n_h/n$. He/she may choose to

- minimize $\text{Var}(\hat{\mu}_{st})$ for a fixed cost of sampling, or
- minimize the cost for a specified $\text{Var}(\hat{\mu}_{st})$.

4.5.1 Theorem

Suppose that the sample cost has the form

$$C = c_0 + \sum c_h n_h$$

where the term c_0 represents an overhead cost and c_h represents the cost per unit in stratum h . Then in stratified random sampling, the cost C is minimized for a specified variance $\text{Var}(\hat{\mu}_{st})$, and the variance $\text{Var}(\hat{\mu}_{st})$ is minimized for a fixed cost C if

$$n_h \propto W_h S_h / \sqrt{c_h}.$$

That is,

$$\frac{n_h}{n} = \frac{W_h S_h / \sqrt{c_h}}{\sum (W_h S_h / \sqrt{c_h})}. \quad (6)$$

For a derivation, see Cochran (1977) page 97.

The implication of this result are clear. In a given stratum, take a larger sample if

- the stratum is larger ($W_h = N_h/N$ is larger);
- the stratum is more variable (S_h is larger);
- sampling is cheaper (c_h is smaller).

4.5.2 Determining total sample size

If the cost C is fixed, we substitute n_h from (6) into the cost function and solve for n :

$$n = \frac{(C - c_0) \sum (N_h S_h / \sqrt{c_h})}{\sum (N_h S_h \sqrt{c_h})}.$$

If $V = \text{Var}(\hat{\mu}_{st})$ is fixed, substitute n_h from (6) into the formula for the variance:

$$\begin{aligned} \text{Var}(\hat{\mu}_{st}) &= \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h} (1 - f_h) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h} \left(1 - \frac{n_h}{N_h}\right) \\ &= \sum_{h=1}^L W_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^L W_h \frac{S_h^2}{N}. \end{aligned} \quad (7)$$

So,

$$\begin{aligned} V &= \sum_{h=1}^L W_h^2 \frac{S_h^2}{n} \frac{\sum (W_h S_h / \sqrt{c_h})}{W_h S_h / \sqrt{c_h}} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \\ &= \frac{\sum (W_h S_h / \sqrt{c_h})}{n} \sum_{h=1}^L W_h S_h \sqrt{c_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2 \\ n &= \frac{(\sum W_h S_h / \sqrt{c_h}) (\sum W_h S_h \sqrt{c_h})}{V + (1/N) \sum W_h S_h^2}. \end{aligned}$$

4.5.3 Neyman allocation

If $c_h = c$ for all h , then formula (6) gives

$$n_h = n \times \frac{W_h S_h}{\sum (W_h S_h)} = n \times \frac{N_h S_h}{\sum (N_h S_h)},$$

which minimizes $V = \text{Var}(\hat{\mu}_{st})$ when sampling costs are the same and n is fixed. This is called *Neyman allocation* (or sometimes *optimal allocation*).

The minimum variance can be found by substituting n_h into (7):

$$\text{Var}_{opt}(\hat{\mu}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}.$$

4.6 Comparison with SRS

We shall compare the efficiency of simple random sampling and stratified random sampling with proportional allocation and Neyman allocation.

- For SRS,

$$\text{Var}(\hat{\mu}_{srs}) = \frac{S^2}{n}(1-f).$$

- For stratified random sampling with proportional allocation, we have

$$\text{Var}_{prop}(\hat{\mu}_{st}) = \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2.$$

- For stratified random sampling with Neyman allocation, we have

$$\text{Var}_{opt}(\hat{\mu}_{st}) = \frac{(\sum W_h S_h)^2}{n} - \frac{\sum W_h S_h^2}{N}.$$

4.6.1 Comparing $\text{Var}(\hat{\mu}_{srs})$ with $\text{Var}_{prop}(\hat{\mu}_{st})$:

Recall that

$$\text{total SS} = \text{within group SS} + \text{between group SS}.$$

When we apply this to strata (groups), we have

$$(N-1)S^2 = \sum_{h=1}^L (N_h - 1)S_h^2 + \sum_{h=1}^L N_h(\mu_h - \mu)^2.$$

Dividing by N gives

$$(1 - \frac{1}{N})S^2 = \sum_{h=1}^L (W_h - \frac{1}{N})S_h^2 + \sum_{h=1}^L W_h(\mu_h - \mu)^2.$$

If $1/N$ is negligible then

$$\begin{aligned} \text{Var}(\hat{\mu}_{srs}) &= \frac{S^2}{n}(1-f) \\ &\approx \frac{1-f}{n} \sum_{h=1}^L W_h S_h^2 + \frac{1-f}{n} \sum_{h=1}^L W_h(\mu_h - \mu)^2 \\ &= \text{Var}_{prop}(\hat{\mu}_{st}) + \frac{1-f}{n} \sum_{h=1}^L W_h(\mu_h - \mu)^2 \end{aligned}$$

Since the second term on the right is ≥ 0 , we have shown that, if $1/N$ is negligible,

$$\text{Var}(\hat{\mu}_{srs}) \geq \text{Var}_{prop}(\hat{\mu}_{st})$$

If $1/N$ is not negligible, it is mathematically possible for proportional allocation to give a higher variance than SRS. This can happen when the between strata mean square is smaller than the within strata mean square. But this is not very likely in practice.

4.6.2 Comparing $\text{Var}_{prop}(\hat{\mu}_{st})$ with $\text{Var}_{opt}(\hat{\mu}_{st})$:

By definition, $\text{Var}_{opt}(\hat{\mu}_{st}) \leq \text{Var}_{prop}(\hat{\mu}_{st})$. We have

$$\begin{aligned}\text{Var}_{prop}(\hat{\mu}_{st}) - \text{Var}_{opt}(\hat{\mu}_{st}) &= \frac{1}{n} [\sum W_h S_h^2 - (\sum W_h S_h)^2] \\ &= \frac{1}{n} [\sum W_h (S_h - \bar{S})^2]\end{aligned}$$

where $\bar{S} = \sum W_h S_h$ is the weighted mean of the S_h . From this we see that if the strata have very different variances, then Neyman allocation is a lot better than proportional allocation. However if all stratum variances are equal, then the proportional allocation agrees with the optimal Neyman allocation.

4.6.3 Comparing $\text{Var}(\hat{\mu}_{srs})$ with $\text{Var}_{opt}(\hat{\mu}_{st})$:

If $1/N$ is negligible, we have

$$\begin{aligned}\text{Var}(\hat{\mu}_{srs}) &= \text{Var}_{prop}(\hat{\mu}_{st}) + \frac{1-f}{n} \sum_{h=1}^L W_h (\mu_h - \mu)^2 \\ &= \text{Var}_{opt}(\hat{\mu}_{st}) + \frac{1}{n} [\sum W_h (S_h - \bar{S})^2] + \frac{1-f}{n} \sum_{h=1}^L W_h (\mu_h - \mu)^2\end{aligned}$$

We see that when changing from SRS to stratified random sampling with optimal allocation, there are two sources of reduction in the variance: one is due to the difference between the stratum means (last term), and one is due to difference between stratum variances (middle term).

The ideal variate for stratification is Y itself. This is impossible in practice, since we do not know the values of Y . However, large gains in precision can be made in situations that satisfy the following three conditions:

- There is a good measure of size for all units in the population.
- There is good correlation between the size measure and the Y variate.
- The units in the population vary widely in size.

In such as case we can stratify by the size measure, and it will increase the precision of estimation of the mean of Y .

4.7 Post-stratification

With some variables suitable for stratification, the stratum to which a unit belongs may not be known until after the data are collected. For example, sex, age, educational level, etc. The stratum size N_h and the weight W_h may be known quite accurately from official statistics.

The procedure is to take a SRS and then classify the units into strata. This is called *post-stratification*. The estimate of population mean μ can be found as before using the stratum weights $W_h = N_h/N$. It is still unbiased, but its variance will be different.

Let m_h be the number of units in stratum h , then m_h is a random variable. To see that $\hat{\mu}_{st}$ is unbiased under post-stratification, note that conditioned on $m = (m_1, \dots, m_L)$ our sample is a vanilla stratified random sample, so conditioned on m , $\hat{\mu}_{st}$ is unbiased. Thus

$$E\hat{\mu}_{st} = EE(\hat{\mu}_{st} | m) = E\mu = \mu.$$

To calculate $\text{Var}_p(\hat{\mu}_{st})$ we use

$$\text{Var}_p(\hat{\mu}_{st}) = E\text{Var}_p(\hat{\mu}_{st} | m) + \text{Var} E(\hat{\mu}_{st} | m).$$

The subscript p indicating post-stratification.

Since $E(\hat{\mu}_{st} | m) = \mu$ we have for the second term on the RHS, $\text{Var} E(\hat{\mu}_{st} | m) = 0$

For the first term on the RHS we have, conditioning on $m = (m_1, \dots, m_L)$,

$$\text{Var}_p(\hat{\mu}_{st} | m) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{m_h} (1 - f_h) = \sum_{h=1}^L W_h^2 \frac{S_h^2}{m_h} - \frac{1}{N} \sum_{h=1}^L W_h S_h^2$$

To calculate $E\text{Var}_p(\hat{\mu}_{st} | m)$ we need $E(1/m_h)$, for which we use a Taylor approximation. For any function f and random variable X with mean μ we have

$$\begin{aligned} Ef(X) &\approx E(f(\mu) + (X - \mu)f'(\mu) + (X - \mu)^2 f''(\mu)/2) \\ &= f(\mu) + \text{Var}(X)f''(\mu)/2. \end{aligned}$$

Now, since $m_h \sim \text{hypergeometric}(N, N_h, n)$, we have $Em_h = nN_h/N = nW_h$ and $\text{Var} m_h = nW_h(1 - W_h)(1 - (n-1)/(N-1))$. Putting $f(X) = 1/X$ and $X = m_h$ we get (ignoring the term $1 - (n-1)/(N-1)$)

$$E\left(\frac{1}{m_h}\right) \approx \frac{1}{nW_h} + \frac{1 - W_h}{n^2 W_h^2}.$$

Note that we actually have $\text{Pr}(m_h = 0) > 0$, so $E(1/m_h) = \infty$. To avoid this we should, strictly speaking, redefine our estimator to exclude strata that are unrepresented and then consider $E(1/m_h | m_h > 0)$. However any benefit from doing this is out-weighed by the inaccuracy of the Taylor expansion, so we don't bother.

We thus have

$$\text{Var}_p(\hat{\mu}_{st}) \approx \frac{1 - f}{n} \sum W_h S_h^2 + \frac{1}{n^2} \sum (1 - W_h) S_h^2$$

The first term is $\text{Var}_{prop}(\hat{\mu}_{st})$. The second term shows the increase in the variance due to post-stratification. This is small in comparison if n is large.

4.8 Exercises

- The households in a town are to be sampled in order to estimate the average amount of assets per household that are readily convertible into cash. The households are stratified into a high-rent and a low-rent stratum. A house in the high-rent stratum is thought to have about nine times as much in assets as one in the low-rent stratum, and S_r^2 is expected to be proportional to μ_r , for $r = 1, 2$.

There are 4,000 households in the high-rent stratum and 20,000 in the low-rent stratum.

- How would you distribute a sample of 1,000 households between the two strata?
 - How should the sample be distributed if the aim of the survey were to estimate the difference between assets per household in the two strata?
- A sampler proposes to take a stratified random sample. He/she expects that his/her field costs will be of the form $\sum c_r n_r$. His/her advance estimates of relevant quantities for the two strata are as follows:

stratum	N_r/N	S_r	c_r
1	0.4	10	\$4
2	0.6	20	\$9

- Find the values of n_1/n and n_2/n that minimise the total field cost for a given value of $\text{Var}(\hat{\mu}_{st})$.

- (b) Using this optimum allocation, find the sample size required to make $\text{Var}(\hat{\mu}_{st}) = 1$. Ignore the fpc.
- (c) How much will the total field cost be?

After this sample is obtained the sampler finds that his/her field costs were actually \$2 per unit in stratum 1 and \$12 per unit in stratum 2.

- (d) By how much does the actual field cost exceed the anticipated field cost?
- (e) If he/she had known the correct costs in advance, could he/she have achieved $\text{Var}(\hat{\mu}_{st}) = 1$ for the original estimated field cost?
3. Dairy farms in a certain geographic region are divided into four categories, depending on their total acreage and on whether or not they concentrate exclusively on dairy products. The numbers of farms in the four categories are 72, 37, 50 and 11. In a survey to estimate the total number of milk-producing cows in the region, a stratified sample of 28 farms is chosen with (roughly) proportional allocation. The number of cows in the selected farms are:

category		no. of cows
1		61 47 44 70 28 39 51 62 101 49 54 71
2		160 148 89 139 142 93
3		26 19 21 34 28 15 20 24
4		17 11

Estimate the number of cows on dairy farms in the region, and the standard error of the estimator.

4. If the cost function is of the form $C = c_0 + \sum_{r=1}^L c_r \sqrt{n_r}$, for c_0, c_1, \dots, c_L known, show that in order to minimise $\text{Var}(\hat{\mu}_{st})$ for a fixed total cost, n_r must be proportional to

$$(N_r^2 S_r^2 / c_r)^{2/3}.$$

Find the n_r for a sample of size $n = 1000$, given the following information:

r	N_r/N	S_r	c_r
1	0.5	4	1
2	0.3	5	2
3	0.2	6	4

5. The following data show the stratification of all the farms in a county by farm size and the average acres of corn (maize) per farm in each stratum.

Farm Size (acres)	Number of Farms N_h	Average Corn Acres μ_Y	Standard Deviation S_h
0–40	394	5.4	8.3
41–80	461	16.3	13.3
81–120	391	24.3	15.1
121–160	334	34.5	19.8
161–200	169	42.1	24.5
202–240	113	50.1	26.0
241–	148	63.8	35.2

For a sample of 100 farms, compute the sample sizes in each stratum under (a) proportional allocation, (b) optimum (Neyman) allocation. Compare the precision of these methods with that of simple random sampling.

6. Two dentists, A and B, make a survey of the state of teeth of 200 children in a small town. Dr A selects a SRS of 20 children and counts the number of decayed teeth for each child with the following results

No. decayed teeth/child:	0	1	2	3	4	5	6	7	8	9	10
Frequency:	8	4	2	2	1	1	0	0	0	1	1

Dr B, using the same dental technique, examines all 200 children, recording only those who have no decayed teeth. He finds 60 children with no decayed teeth.

Estimate the mean number of decayed teeth for the town's children

- (a) using A's results only;
- (b) using A's and B's results.

Are the estimates unbiased? Which estimate do you expect to be more precise?

7. A company intends to interview a simple random sample of employees who have been with it for more than five years. The company has \$1000 to spend and each interview costs \$10. There is no separate list of employees with more than five years service, but a list can be compiled from the files at a cost of \$200. The company can either

- (a) compile the list and then interview a simple random sample of size 80, drawn from the eligible employees, or
- (b) select employees from the whole company at random but interview them only if they are eligible, repeating this until 100 employees have been interviewed. (We are assuming that the cost of selecting and then not interviewing an employee is negligible.)

Show that to estimate the total of some variable of interest X , over the population of eligible employees, the first plan gives a smaller variance than the second plan if $CV < 2\sqrt{q}$, where CV is the coefficient of variation of X among eligible employees and q is the proportion of non-eligible employees in the company.

Ignore the fpc in your analysis.

5 Cluster Sampling

The population is divided in non-overlapping groups called **clusters**. A sample of clusters are selected.

The clusters are the *primary units* of sampling. The members of the clusters are the *secondary units*. If all the members of each selected cluster are included in our sample (of secondary units), the method is called a *one-stage cluster sampling*. If we take a random sample of each selected cluster, the method is called a *two-stage cluster sampling*.

The secondary units may themselves be groups of *tertiary units*, and we carry on into sub-sampling tertiary units from the selected secondary unit, etc. This is called a *multi-stage cluster sampling* scheme. For example, in surveying the performance of school children, the country may be divided into areas (which form the primary units), schools within the areas form the secondary units, the classes within the schools form the tertiary units and the children within the classes form the main objects of the study population.

Clustering is typically forced upon us, rather than something we wish to happen. For example, it may be very expensive to compile a list of all the members of the population we want to interview, but relatively cheap to form a list of all the clusters, e.g. all the schools. Alternatively, the cost of sampling may increase as the distance between the units increases, making it cost effective to cluster together units which are close to each other.

We will restrict ourselves to one-stage cluster sampling.

5.1 Notation:

Overall:

- Y = variable of interest
- M = total number of elements in population
- μ_Y = mean per element of population
- τ_Y = total over the whole population ($= M\mu_Y$)

Clusters:

- N = number of clusters in the population
- m_h = size of cluster h (so $M = \sum_{h=1}^N m_h$)
- Y_{hj} = Y -value of j -th element of cluster h
- μ_h = mean per element of cluster h
- τ_h = total over cluster h ($= m_h\mu_h$)
- σ_h^2 = variance per element within cluster h

Sample of clusters:

- n = number of clusters in sample
- $k(i)$ = index of i -th sample cluster, $i = 1, \dots, n$.

5.2 One-stage cluster sampling with unequal-sized clusters

In this case, with each cluster can be associated two measurements: the cluster total $\tau_h = m_h\mu_h$ and the cluster size m_h . The population mean per element μ_Y is, at cluster level, a ratio of the two totals and so we estimate it by a *ratio estimator*:

$$\mu_Y = \frac{\sum_{h=1}^N \tau_h}{\sum_{h=1}^N m_h} \quad \left(\neq \sum_{h=1}^N \frac{\tau_h}{m_h} \right)$$

$$\hat{\mu}_{clr} = \frac{\bar{\tau}}{\bar{m}}$$

where $\bar{\tau} = (\sum_{i=1}^n \tau_{k(i)})/n$ and $\bar{m} = (\sum_{i=1}^n m_{k(i)})/n$ are the sample means of cluster total and cluster size respectively.

Alternatively, we can express $\hat{\mu}_{clr}$ as a weighted sum of cluster means:

$$\hat{\mu}_{clr} = \frac{\sum_{i=1}^n \tau_{k(i)}}{\sum_{i=1}^n m_{k(i)}} = \sum_{i=1}^n \frac{m_{k(i)}}{\sum_{j=1}^n m_{k(j)}} \mu_{k(i)}.$$

Unfortunately, because $E(1/\bar{m}) \neq 1/E(\bar{m})$, $\hat{\mu}_{clr}$ is biased in general. Approximating \bar{m} by the population analogue $(\sum_{h=1}^N m_h)/N = M/N$, we have

$$\hat{\mu}_{clr} - \mu = \frac{\bar{\tau} - \mu\bar{m}}{\bar{m}} \approx \frac{\bar{\tau} - \mu\bar{m}}{M/N} = \frac{1}{M/N} \bar{z}$$

where $\bar{z} = \sum_{i=1}^n z_{k(i)}$ and $z_{k(i)} = \tau_{k(i)} - \mu m_{k(i)}$. The $z_{k(i)}$ are just a simple random sample, with $\mu_Z = 0$, so

$$\text{Var}(\hat{\mu}_{clr}) = \text{Var}(\hat{\mu}_{clr} - \mu) \approx \frac{1}{(M/N)^2} \text{Var} \bar{z} = \frac{1}{(M/N)^2} \frac{1-f}{n} \frac{1}{N-1} \sum_{h=1}^N (\tau_h - \mu m_h)^2$$

To estimate this we use

$$\hat{\text{Var}}(\hat{\mu}_{clr}) = \frac{1}{(M/N)^2} \frac{1-f}{n} \frac{1}{n-1} \sum_{i=1}^n (\tau_{k(i)} - \hat{\mu}_{clr} m_{k(i)})^2.$$

If M is not known then we use \bar{m} instead of M/N .

Remark: we approximate $\hat{\mu}_{clr} - \mu$ rather than $\hat{\mu}_{clr}$ directly because the numerator of the former is smaller, and thus the error introduced by substituting M/N for \bar{m} is reduced.

5.3 One-stage cluster sampling with equal-sized clusters

If we assume

$$m_1 = m_2 = \dots = m_N = m,$$

then cluster sampling simplifies somewhat. We have $M = mN$ and

$$\mu_Y = \frac{1}{mN} \sum_{h=1}^N \sum_{j=1}^m Y_{hj} = \frac{1}{N} \sum_{h=1}^N \left(\frac{1}{m} \sum_{j=1}^m Y_{hj} \right) = \frac{1}{N} \sum_{h=1}^N \mu_h.$$

That is, the population mean is the mean of all the cluster means. Hence we can treat the clusters as the sampling units and the cluster means as the measurement per unit of interest. The population of cluster means are $\{\mu_1, \mu_2, \dots, \mu_N\}$, which has mean μ_Y and variance (the between cluster variance)

$$\sigma_b^2 = \frac{1}{N} \sum_{h=1}^N (\mu_h - \mu_Y)^2.$$

We have a SRS of clusters, which have means $\{\mu_{k(1)}, \mu_{k(2)}, \dots, \mu_{k(n)}\}$. Using the results we have on SRS,

$$\begin{aligned} \hat{\mu}_{cl} &= \frac{1}{n} \sum_{i=1}^n \mu_{k(i)} \quad (\text{mean of the sampled cluster means}) \\ \text{Var}(\hat{\mu}_{cl}) &= \frac{S_b^2}{n} \left(1 - \frac{n}{N}\right) \quad \text{where} \quad S_b^2 = \frac{N\sigma_b^2}{N-1} \\ \hat{\text{Var}}(\hat{\mu}_{cl}) &= \frac{s_b^2}{n} \left(1 - \frac{n}{N}\right) \quad \text{where} \quad s_b^2 = \frac{1}{n-1} \sum_{i=1}^n (\mu_{k(i)} - \hat{\mu}_{cl})^2 \quad (\text{sample variance of the cluster means}) \end{aligned}$$

5.3.1 Example (Scheaffer et al, Ex 8.5)

The circulation manager of a newspaper wishes to estimate the average numbers of purchased newspapers per household in a given community. Travel costs from household to household are substantial. Therefore the 4000 households in the community are listed in 400 geographical clusters of 10 households each, and a SRS of 4 clusters is selected. Estimate the average number of newspapers per household for the community, and place a bound on the error of estimation.

Cluster	Numbers of newspapers										Total	Mean
1	1	2	1	3	3	2	1	4	1	1	19	1.9
2	1	3	2	2	3	1	4	1	1	2	20	2.0
3	2	1	1	1	1	3	2	1	3	1	16	1.6
4	1	1	3	2	1	5	1	2	3	1	20	2.0

Solution: We have

$$\begin{aligned}\hat{\mu}_{cl} &= \frac{1.9 + 2.0 + 1.6 + 2.0}{4} = \frac{7.5}{4} = 1.875 \\ s_b^2 &= \frac{1}{3}(1.9^2 + 2.0^2 + 1.6^2 + 2.0^2 - (7.5)^2/4) = 0.0358 \\ \widehat{\text{Var}}(\hat{\mu}_{cl}) &= \frac{0.0358}{4}\left(1 - \frac{4}{400}\right) = 0.0086\end{aligned}$$

5.4 Comparison of cluster sampling with SRS

We consider one-stage cluster sampling with equal-sized clusters.

The total number of elements in n clusters is $n^* = mn$. So, we are comparing a SRS of mn elements with a cluster sample of n clusters of equal size m .

For SRS,

$$\text{Var}(\hat{\mu}_{srs}) = \frac{S^2}{n^*}\left(1 - \frac{n^*}{M}\right) = \frac{S^2}{n^*}\left(1 - \frac{n}{N}\right).$$

For cluster sampling,

$$\text{Var}(\hat{\mu}_{cl}) = \frac{S_b^2}{n}\left(1 - \frac{n}{N}\right).$$

To compare them, we again recall that

$$\text{total SS} = \text{between SS} + \text{within SS}.$$

For the population clusters we have

$$\begin{aligned}\text{total SS} &= (M - 1)S^2 = (mN - 1)S^2 \\ \text{between SS} &= m \times (N - 1)S_b^2 \\ \text{within SS} &= \sum_{h=1}^N (m - 1)S_h^2 = N(m - 1)\bar{S}^2\end{aligned}$$

where $\bar{S}^2 = (\sum_{h=1}^N S_h^2)/N$ is the average within-cluster ‘variance’. This gives

$$m(N - 1)S_b^2 = (mN - 1)S^2 - N(m - 1)\bar{S}^2$$

whence

$$\begin{aligned}\text{Var}(\hat{\mu}_{srs}) - \text{Var}(\hat{\mu}_{cl}) &= \frac{1 - n/N}{mn(N - 1)}[(N - 1)S^2 - m(N - 1)S_b^2] \\ &= \frac{1 - n/N}{mn(N - 1)}[(N - 1)S^2 - (mN - 1)S^2 + N(m - 1)\bar{S}^2] \\ &= \frac{(N - n)(m - 1)}{mn(N - 1)}[\bar{S}^2 - S^2]\end{aligned}$$

This shows that, ignoring cost and other convenience factors, a cluster sample mean is better than a SRS sample mean, if the average within-cluster variance \bar{S}^2 is larger than the overall population variance S^2 , and vice versa.

Compare this to what was found in stratified random sampling, which is better than SRS if the within-strata variation is *low*. The difference between the two schemes is that cluster sampling takes whole groups, while stratified sampling takes a sub-sample from each group

5.5 Estimation of population total

How we estimate τ depends on whether we know M , the total number of elements in the population. It is not a problem when all clusters are of equal size, which implies $M = mN$.

5.5.1 When M is known:

We use

$$\hat{\tau}_Y = M\hat{\mu}, \quad \text{and} \quad \text{Var}(\hat{\tau}_Y) = M^2\text{Var}(\hat{\mu})$$

where $\hat{\mu}$ may be $\hat{\mu}_{cl}$ or $\hat{\mu}_{clr}$.

5.5.2 When M is unknown:

Here, we consider the cluster total as a measurement of interest in the population of clusters. The population total is the total of all cluster totals and we have a SRS of clusters. Using our results on SRS, we use $\bar{\tau} = (\sum_{i=1}^n \tau_{k(i)})/n$ as an estimate of the population mean of cluster totals, and

$$\hat{\tau}_Y = N\bar{\tau}.$$

Its variance and variance estimator are

$$\begin{aligned} \text{Var}(\hat{\tau}_Y) &= N^2\text{Var}(\bar{\tau}) = N^2\frac{S_\tau^2}{n}\left(1 - \frac{n}{N}\right) \\ \hat{\text{Var}}(\hat{\tau}_Y) &= N^2\frac{s_\tau^2}{n}\left(1 - \frac{n}{N}\right) \end{aligned}$$

where

$$S_\tau^2 = \frac{1}{N-1} \sum_{h=1}^N (\tau_h - \tau/N)^2 \quad \text{and} \quad s_\tau^2 = \frac{1}{n-1} \sum_{i=1}^n (\tau_{k(i)} - \bar{\tau})^2.$$

5.6 Exercises

1. For the sample of 30 households shown below, the data refer to visits to the dentist in the last year.

Number of Persons	Dentist seen		Number of Persons	Dentist seen	
	Yes	No		Yes	No
5	1	4	5	1	4
6	0	6	4	4	0
3	1	2	4	1	3
3	2	1	3	1	2
2	0	2	3	0	3
3	0	3	4	1	3
3	1	2	3	0	3
3	1	2	3	1	2
4	1	3	1	0	1
4	0	4	2	0	2
3	1	2	4	0	4
2	0	2	3	1	2
7	2	5	4	1	3
4	1	3	2	0	2
3	0	3	4	0	4

Estimate the variance of the proportion of persons who saw a dentist, and compare this with the estimate of the variance obtained if you ignore the cluster structure.

2. The population values are:

1	3	2	3	4	1	3	2	1	2	2	3
4	5	7	6	7	4	7	6	4	5	4	7
9	10	8	9	11	9	9	8	10	8	9	12

A sample of size 6 is to be selected from the above population. Compare the accuracy of

- (a) simple random sampling ($n = 6$),
 - (b) cluster sampling with columns as clusters and $m = 3$ ($n = 2$),
 - (c) cluster sampling with pairs of columns as clusters and $m = 6$ ($n = 1$),
 - (d) stratified sampling with rows as strata and $n_1 = n_2 = n_3 = 2$.
3. (a) Obtain, from the formulae for estimating a mean μ for cluster sampling with clusters of equal sizes, the formulae for estimating a proportion P and the s.e. of the estimate.
- (b) To emphasise safety, a taxicab company wants to estimate the proportion of unsafe tyres on their 175 cabs. (Ignore the spare tyres.) Selecting a simple random sample of tyres is impractical, so cluster sampling is used, with each cab as a cluster. A random sample of 25 cabs gives the following number of unsafe tyres per cab:

2	4	0	1	2	0	4	1	3	1	2	0	1
1	2	2	4	1	0	0	3	1	2	2	1	

Estimate the proportion of unsafe tyres being used on the company's cabs, and give the s.e. of the estimate.

6 Ratio and regression estimators

We now return to SRS and consider ways of estimating μ_Y when we have auxiliary information. Suppose associated with each unit in the population is a measurement X which is highly correlated with Y , the variable of interest. We assume that μ_X is known. Can we use this information to estimate μ_Y more accurately?

6.1 Ratio estimator

Since $\mu_Y = \frac{\mu_Y}{\mu_X} \mu_X = R\mu_X$, this suggests the estimator

$$\hat{\mu}_{ratio} = r\mu_X$$

where $r = \bar{y}/\bar{x}$ is the estimate of the ratio $R = \mu_Y/\mu_X$.

$\hat{\mu}_{ratio}$ is biased in general, because $E(1/\bar{x}) \neq 1/\mu_X$. None-the-less it is often preferred to $\hat{\mu}_{srs} = \bar{y}$ because, depending on the correlation between X and Y , it can have a lower variance.

To estimate the variance of $\hat{\mu}_{ratio}$ we note

$$\hat{\mu}_{ratio} - \mu_Y = (r - R)\mu_X = \frac{\bar{y} - R\bar{x}}{\bar{x}}\mu_X \approx \bar{y} - R\bar{x}.$$

Let $Z_i = Y_i - R X_i$ then $\mu_Z = 0$ and we have

$$\text{Var}(\hat{\mu}_{ratio}) = \text{Var}(\hat{\mu}_{ratio} - \mu_Y) \approx \text{Var}(\bar{z}) = \frac{1}{n} S_Z^2 (1 - f)$$

where

$$\begin{aligned} S_Z^2 &= \frac{1}{N-1} \sum_{i=1}^N (Z_i - \mu_Z)^2 = \frac{1}{N-1} \sum_{i=1}^N (Y_i - R X_i)^2 \\ &= S_Y^2 - 2RS_{XY} + R^2 S_X^2 \end{aligned}$$

where $S_{XY} = \frac{1}{N-1} \sum_{\ell=1}^N (Y_\ell - \mu_Y)(X_\ell - \mu_X)$. We estimate S_Z^2 by

$$\begin{aligned} s_z^2 &= \frac{1}{n-1} \sum_{i=1}^n (y_i - r x_i)^2 \\ &= s_y^2 - 2r s_{xy} + r^2 s_x^2 \\ &= s_y^2 - 2r \hat{\rho} s_x s_y + r^2 s_x^2 \end{aligned}$$

where $\hat{\rho} = s_{xy}/(s_x s_y)$ is the correlation coefficient of the sample data on X and Y .

6.1.1 Theorem

In SRS, the ratio estimator $r\mu_X$ is more precise than the mean per unit estimator \bar{y} , if and only if

$$R > 0 \text{ and } \rho_{XY} > \frac{1}{2} \frac{CV_X}{CV_Y} \quad \text{or} \quad R < 0 \text{ and } \rho_{XY} < \frac{1}{2} \frac{CV_X}{CV_Y}$$

Proof: Suppose $R > 0$, then

$$\begin{aligned} \text{Var}(\bar{y}) - \text{Var}(\hat{\mu}_{ratio}) &= \frac{1-f}{n} (2RS_{XY} - R^2 S_X^2) > 0 \\ &\Leftrightarrow 2RS_{XY} - R^2 S_X^2 > 0 \end{aligned}$$

$$\begin{aligned}
&\Leftrightarrow S_{XY} > \frac{1}{2}RS_X^2 = \frac{1}{2}\frac{\mu_Y}{\mu_X}S_X^2 \\
&\Leftrightarrow \frac{S_{XY}}{S_XS_Y} > \frac{1}{2}\frac{\mu_Y}{\mu_X}\frac{S_X}{S_Y} = \frac{1}{2}\frac{S_X/\mu_X}{S_Y/\mu_Y} \\
&\Leftrightarrow \rho_{XY} > \frac{1}{2}\frac{CV_X}{CV_Y}
\end{aligned}$$

If $R < 0$ then the same argument shows $\text{Var}(\bar{y}) - \text{Var}(\hat{\mu}_{ratio}) > 0$ iff $\rho_{XY} < CV_X/(2CV_Y)$. If $R = 0$ then $\text{Var}(\bar{y}) - \text{Var}(\hat{\mu}_{ratio}) = 0$.

Remark: We see that it is not guaranteed that a ratio estimator will be more efficient than the mean per unit estimator, \bar{y} : we need the population correlation coefficient to be sufficiently large. Even then, it may not be sufficient to make a ratio estimator more efficient. For example, if $CV_X > 2CV_Y$, then $CV_X/(2CV_Y) > 1$ and ρ_{XY} can never be larger than 1.

However, there are many practical situations where a ratio estimator is much better than \bar{y} . We need the following conditions to hold:

- We need to be able to observe two variables X and Y that are roughly proportional, and μ_X must be known;
- The CV of X must not be much larger than that of Y .

6.2 Regression estimator

If we find that the relation between X and Y is approximately linear, but the regression line does not pass through the origin, then a *linear regression estimator* is more appropriate than a ratio estimator. Again we suppose that μ_X is known.

The regression estimator of μ_Y is

$$\hat{\mu}_{lr} = \bar{y} + b(\mu_X - \bar{x})$$

where b may be pre-assigned or estimated from the sample.

The rationale behind $\hat{\mu}_{lr}$ is as follows. Consider the points (X_ℓ, Y_ℓ) of the whole population. If they are related by

$$Y_\ell = a + bX_\ell + \epsilon_\ell, \quad \text{where } E(\epsilon_\ell) = 0,$$

then,

$$\mu_Y = a + b\mu_X.$$

It is reasonable to expect the sample means to satisfy this relation approximately

$$\bar{y} \approx a + b\bar{x} \quad \Rightarrow \quad a \approx \bar{y} - b\bar{x}$$

whence,

$$\mu_Y \approx \bar{y} - b\bar{x} + b\mu_X = \bar{y} + b(\mu_X - \bar{x}).$$

6.2.1 When b is pre-assigned

Suppose the pre-assigned value is b_0 . Then the linear regression estimator of μ_Y is

$$\hat{\mu}_{lr} = \bar{y} + b_0(\mu_X - \bar{x})$$

It has the following properties:

-

$$E(\hat{\mu}_{lr}) = E(\bar{y}) + b_0[\mu_X - E(\bar{x})] = \mu_Y + b_0(\mu_X - \mu_X) = \mu_Y \text{ (unbiased).}$$

•

$$\begin{aligned}
\text{Var}(\hat{\mu}_{lr}) &= \text{Var}[\bar{y} + b_0(\mu_X - \bar{x})] \\
&= \text{Var}(\bar{y} - b_0\bar{x}) \\
&= \frac{1-f}{n} S_{Y-b_0X}^2 \\
&= \frac{1-f}{n} (S_Y^2 - 2b_0 S_{XY} + b_0^2 S_X^2).
\end{aligned}$$

- Unbiased estimates of S_Y^2 , S_{XY} and S_X^2 are s_y^2 , s_{xy} and s_x^2 respectively, so we put

$$\hat{\text{Var}}(\hat{\mu}_{lr}) = \frac{1-f}{n} (s_y^2 - 2b_0 s_{xy} + b_0^2 s_x^2).$$

6.2.2 When b is estimated from the sample

Consider $b_0 = b$ as a variable and ask what value to give it to minimise $\text{Var}(\hat{\mu}_{lr})$. Putting $\partial \text{Var}(\hat{\mu}_{lr}) / \partial b = 0$ and treating S_Y^2 , S_{XY} and S_X^2 as constants, gives

$$-2S_{XY} + 2bS_X^2 = 0 \quad \Rightarrow \quad b = \frac{S_{XY}}{S_X^2}.$$

Denote this value of b by

$$\beta := \frac{S_{XY}}{S_X^2}.$$

Thus the minimum value of $\text{Var}(\hat{\mu}_{lr})$ is

$$\min_b [\text{Var}(\hat{\mu}_{lr})] = \frac{1-f}{n} \left(S_Y^2 - \frac{S_{XY}^2}{S_X^2} \right) = \frac{1-f}{n} S_Y^2 (1 - \rho^2).$$

The obvious estimate of β to use is

$$\hat{\beta} = \frac{s_{XY}}{s_X^2}.$$

Using this, the linear regression estimate of μ_Y is

$$\hat{\mu}_{lr}(\hat{\beta}) = \bar{y} + \hat{\beta}(\mu_X - \bar{x}).$$

Its variance is

$$\text{Var}[\hat{\mu}_{lr}(\hat{\beta})] \approx \text{Var}[\hat{\mu}_{lr}(\beta)] = \frac{1-f}{n} \left(S_Y^2 - \frac{S_{XY}^2}{S_X^2} \right)$$

The variance estimator is

$$\begin{aligned}
\hat{\text{Var}}[\hat{\mu}_{lr}(\hat{\beta})] &= \frac{1-f}{n} \frac{n-1}{n-2} \left(s_y^2 - \frac{s_{xy}^2}{s_x^2} \right) \\
&= \frac{1-f}{n} \frac{n-1}{n-2} s_y^2 (1 - \hat{\rho}^2) \quad \text{where} \quad \hat{\rho} = \frac{s_{xy}}{s_x s_y}
\end{aligned}$$

Note that $((n-1)/(n-2))(s_y^2 - s_{xy}^2/s_x^2)$ is a less biased estimator of $(S_Y^2 - S_{XY}^2/S_X^2)$ than $(s_y^2 - s_{xy}^2/s_x^2)$. Also, when using $\hat{\text{Var}}[\hat{\mu}_{lr}(\hat{\beta})]$ to form a confidence interval for μ , we should use percentage points from the t_{n-2} distribution.

6.3 Comparison of $\hat{\mu}_{lr}(\hat{\beta})$, $\hat{\mu}_{ratio}$ and \bar{y} :

We use the following asymptotic variances:

$$\begin{aligned}\text{Var}[\hat{\mu}_{lr}(\hat{\beta})] &\approx \frac{1-f}{n} \left(S_Y^2 - \frac{S_{XY}^2}{S_X^2} \right) = \frac{1-f}{n} S_Y^2 (1 - \rho^2) \\ \text{Var}(\hat{\mu}_{ratio}) &\approx \frac{1-f}{n} (S_Y^2 - 2RS_{XY} + R^2 S_X^2) \\ \text{Var}(\bar{y}) &= \frac{1-f}{n} S_Y^2.\end{aligned}$$

We have already compared $\text{Var}(\hat{\mu}_{ratio})$ and $\text{Var}(\bar{y})$ and know that the former is more smaller if $\rho_{XY} > CV_X / (2CV_Y)$.

To compare $\hat{\mu}_{lr}(\hat{\beta})$ with \bar{y} :

$$\text{Var}(\bar{y}) - \text{Var}[\hat{\mu}_{lr}(\hat{\beta})] = \frac{1-f}{n} S_Y^2 \rho^2 \geq 0$$

The difference is 0 only when $\rho = 0$. Hence, $\hat{\mu}_{lr}(\hat{\beta})$ is always better than \bar{y} except when X and Y are uncorrelated, in which case the variances are the same.

To compare $\hat{\mu}_{lr}(\hat{\beta})$ with $\hat{\mu}_{ratio}$:

$$\begin{aligned}\text{Var}(\hat{\mu}_{ratio}) - \text{Var}[\hat{\mu}_{lr}(\hat{\beta})] &= \frac{1-f}{n} (-2RS_{XY} + R^2 S_X^2 + \rho^2 S_Y^2) \\ &= \frac{1-f}{n} (\rho^2 S_Y^2 - 2R\rho S_X S_Y + R^2 S_X^2) \\ &= \frac{1-f}{n} (\rho S_Y - R S_X)^2 \geq 0\end{aligned}$$

So, $\hat{\mu}_{lr}(\hat{\beta})$ is always as good as $\hat{\mu}_{ratio}$, with equal variances when $\rho S_Y = R S_X$. In that case we have

$$\frac{S_{XY}}{S_X S_Y} S_Y = R S_X \quad \Leftrightarrow \quad R = \frac{S_{XY}}{S_X^2} = \beta$$

Recall our assumption that $\mu_Y = a + b\mu_X$. If $a = 0$ then $b = R$, but β is just the least squares fit of b , so we see that the two estimators are the equally efficient when the population scatter is linear and through the origin. Otherwise, the regression estimator is better.

6.4 Exercises

1. An experienced estimator makes eye estimates X for each of $N = 2,000$ elements of a population. He obtains $\tau_X = 66,000$.

A random sample of $n = 10$ elements is obtained and the true value Y obtained

true value y	31	42	27	25	34	37	25	32	26	35
estimated value x	31	45	25	26	35	39	29	34	27	33

- (a) Find a ratio estimate of τ_Y and an estimate of its standard error.
 - (b) Find a linear regression estimate of τ_Y and an estimate of its standard error.
2. An experienced farmer makes an eye estimate of the weight of peaches X on each tree in an orchard of $N = 200$ trees. He finds a total of $\tau_X = 5,262$ kg. The peaches from a simple random sample of 10 trees are picked and weighed, with the following results

actual weight y	28	19	23	26	30	20	18	26	32	24
estimated weight x	27	21	24	27	30	22	20	26	34	26

As an estimate of the total actual weight τ_Y we take

$$\hat{\tau}_Y = N[\mu_X + (\bar{y} - \bar{x})].$$

Compute $\hat{\tau}_Y$ and estimate its standard error.

Does it appear that the linear regression estimate, with the sample least squares b , would give a more precise estimate?

3. A pilot survey of 21 households gave the following data for the numbers of members (x), children (y_1), cars (y_2), and TV sets (y_3).

x	5	2	4	4	6	3	5	2	3	2	6	3	4	5	6	4	3	2	4	3	4
y_1	3	0	1	2	4	1	3	0	1	0	4	1	2	3	3	2	2	1	0	2	1
y_2	1	1	2	1	1	1	1	0	1	2	2	0	1	1	2	1	0	2	1	1	1
y_3	3	1	0	1	1	2	1	1	1	0	1	0	1	1	0	1	1	1	1	1	1

For these data, the mean vector and variance-covariance matrix are

$$(3.8095, \quad 1.7143, \quad 1.0952, \quad 0.95238)'$$

and

$$\begin{pmatrix} 1.76190 & 1.44286 & 0.21905 & 0.09048 \\ 1.44286 & 1.61429 & 0.12857 & 0.18571 \\ 0.21905 & 0.12857 & 0.39048 & -0.09524 \\ 0.09048 & 0.18571 & -0.09524 & 0.44762 \end{pmatrix}$$

Suppose it is known that the population has 1,000 households with total $\tau_X = 4,000$. It is desired to estimate the total number of children, cars and TV sets in the community. Determine, for each case, whether it is better to use a ratio estimate or mean per unit estimate and find your chosen estimates.

4. The values of Y and X are measured for each unit in a simple random sample from a population. If μ_X , the population mean of X , is known, which of the following procedures do you recommend for estimating μ_Y/μ_X ?
- Always use \bar{y}/\bar{x} .
 - Sometimes use \bar{y}/μ_X and sometimes use \bar{y}/\bar{x} .
 - Always \bar{y}/μ_X .

Give reasons for your answer.

5. From a list of 468 secondary schools, a simple random sample of 100 schools is selected. The sample contained 54 Government schools and 46 private schools.

Data for the number of students Y and the number of teachers X at each school are given below.

	n	$\sum Y_i$	$\sum X_i$	$\sum Y_i^2$	$\sum X_i Y_i$	$\sum X_i^2$
government	54	31,281	2,024	29,881,219	1,729,349	111,090
private	46	13,707	1,075	6,366,785	431,041	33,119

- For each type of school, estimate the ratio (number of students)/(number of teachers), and give the standard error of the estimates.

- (b) Specify 90% confidence limits for the student/teacher ratio in the population of government schools.
- (c) Test at the 5% level whether the student/teacher ratio is different in the two types of schools.
- (d) Estimate the total number of teachers in the government schools
 - (i) given that the total number of government schools is 251;
 - (ii) without knowing this figure.

In each case specify the standard error of your estimate.