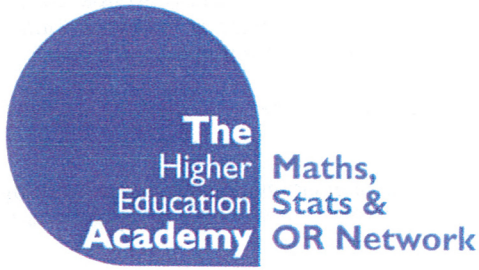
Guide to Statistical Information

Sampling for Surveys

By

Vic Barnett



Resources to support the learning of mathematics, statistics and OR in higher education.
www.mathstore.ac.uk

Sampling for Surveys
A Short Guide

By Vic Barnett

1 Introduction

Sample Survey methods play a vital role in helping us to understand different aspects of the world we live in. They are designed as a statistical basis for collecting and interpreting data from **finite populations** and underlie much of the work on opinion polls, market research and surveys of social, medical, environmental and other issues.

We need to use statistically sound methods of data analysis and inference designed for finite populations and also methods for handling non-statistical difficulties in their application (e.g. non-sampling errors, non-response, biases produced by sensitivity issues, question biases etc).

The finite population on which we conduct a survey might be very large (all voters in the UK) and the survey sample size quite small (perhaps 1000 or so voters), or the population may be small (the 140 members of a Social Club) with a much higher ***sampling fraction*** (e.g. a sample of 40 members).

Target population This is the total finite population of interest e.g all voters in the UK.

Study population This is the population we will actually study e.g. all voters in a chosen set of constituencies in which we will take observations (hoping it 'represents' the target population).

Population variable and population parameter What we will measure on each population member (voting intention: what party?) and its associated characteristic of interest such as the population parameter (e.g. the *proportion* who intend to vote for Party A or the implied *total* number of votes for that Party). Note that voting intention and actual votes cast may be quite different in the event; this is one of the non-statistical problems we face.

Sampling units and sampling frame The entities we will actually sample and the complete set of them e.g. shoppers in selected high streets at different times or eligible occupants at electoral roll addresses in selected wards. Choices have to be made in all these respects.

Why take a sample? Clearly a full population study (a *census*) is seldom feasible in terms of accessibility, time or cost and these three factors control the ***sampling imperative*** to obtain ***sufficiently statistically-reliable*** and ***affordable*** information about the target population. See Barnett (2002).

How should we sample? We must draw a sample which is representative of the population and for which we can assess its statistical properties. In particular, ***accessibility sampling*** ('take what's to hand') or ***judgmental sampling*** (deliberate

subjective choice of sample members) will inevitably lead to bias and will not have measurable statistical properties.

Suppose we are interested in a quantity (a ***variable***) *Y* measured on the *N* members of the population; so that the population can be represented: $Y_1, Y_2, \ldots, Y_N$. We will be interested in a ***characteristic*** of the population:

$$\text{Population mean} \quad \bar{Y} = (\sum_1^N Y_i)/N$$

$$\text{Population total} \quad Y_T = \sum_1^N Y_i$$

The population proportion, P, of population members falling into some category with respect to the variable Y, e.g. the voters who say they will vote for party A.
Of course, $\bar{Y}$, Y_T and P will not be known and the aim of sample survey methods is to construct statistically-sound methods to make inferences about the population values from a sample of $n < N$ values $y_1, y_2, \ldots, y_n$ drawn from the population. (Note that not all Y_i, nor all y_i, necessarily take different values; in the voting example there will only be a few possible values which can be taken.)

2 Random Sampling

In order to assess the statistical properties of inferences drawn from a sample, we need to draw the sample $y_1, y_2, \ldots, y_n$ according to a *probability sampling scheme*.

Simple Random Sampling

The simplest form is *simple random (sr) sampling* where observations are drawn successively with replacement from the population $Y_1, Y_2, \ldots, Y_N$ in such a way that each population member is equally likely to be drawn.

Sampling fraction* or *finite population correction is the ratio $f = n/N$.
We will want to estimate some population characteristic θ (e.g. Y_T) by some function $\tilde{\theta}(S)$ of the sample S. The properties of the ***estimator*** *(or* ***statistic****)* $\tilde{\theta}$ will be assessed from its ***sampling distribution*** i.e. the probability distribution of the different values $\tilde{\theta}$ may take as a result of employing the probability sampling scheme (e.g. sr sampling).

Thus if $E(\tilde{\theta}) = \theta$ we say $\tilde{\theta}$ is ***unbiased***, a property we normally require in sample survey work (we seldom are prepared to 'trade bias for precision'.)

For unbiased estimators we have $Var(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2$ as the variance of $\tilde{\theta}$ which provides an inverse measure of precision: the lower $Var(\tilde{\theta})$ the more precise the estimator $\tilde{\theta}$.

The broad aim is to choose a probability sampling scheme which is easy to use and yields unbiased estimators which effectively minimize the effects of sampling fluctuations (i.e. is as precise as possible).

With sr sampling, all samples y_1 , y_2 , ... , y_n are equally likely to arise and we estimate the population mean, $\bar{Y}$, by the *sample mean* $\bar{y} = (\sum_1^n y_i)/n$ which is easily seen to be unbiased ($E(\bar{y}) = \bar{Y}$), to have variance $Var(\bar{y}) = (1-f)S^2/n$ where $S^2 = \sum_i^N (Y_i - \bar{Y})^2/(N-1)$ is defined as the *population variance* and to be the best (minimum variance) linear estimator based on a sr sample.

We also need to estimate S^2 and use the unbiased estimator $s^2 = \sum_1^n (y_i - \bar{y})^2/(n-1)$, the ***sample variance*** which helps to:

1) assess the precision of $\bar{y}$;
2) compare $\bar{y}$ with other estimators;
3) determine sample size n needed to achieve desired precision.

Thus $s^2(\bar{y}) = (1-f)s^2/n$ is unbiased for $Var(\bar{y})$ and for large enough n we can assume that $\bar{y}$ is approximately normally distributed written $\bar{y} \sim N(\bar{Y}, (1-f)S^2/n)$. This yields an approximate $100(1-\alpha)\%$ symmetric two-sided confidence interval for $\bar{Y}$ as $\bar{y} - z_\alpha s\sqrt{(1-f)/n} \le \bar{Y} \le \bar{y} + z_\alpha s\sqrt{(1-f)/n}$ where z_α is the double-tailed α -point for $N(0,1)$. To choose a sample size n to yield required precision e.g. with $P(|\bar{Y} - \bar{y}| > d) \le \alpha$ for prescribed values of d and α we need $n \ge N/(1 + N(d/(z_\alpha S))^2)$ or specifying $Var(\bar{y}) \le (d/z_\alpha)^2 = V$ say, this becomes $n \ge (S^2/V)[1 + S^2/(NV)]^{-1} \approx S^2/V$ if $S^2/(NV)$ is small. Typically we do not know S^2 and need to estimate it, sometimes rather informally, from ***pilot studies***, ***previous surveys*** or a ***preliminary sample***.

Systematic Sampling With a complete *list* of population members, a simple sampling method is to choose sample members at regular intervals throughout the list to obtain the required sample size, n . This is not strictly sr sampling (nor a probability sampling scheme) but can be effective if there is no relationship between population value and order on the list.

Estimating Y_T An immediate estimate of Y_T is given by $y_T = N\bar{y}$ which is unbiased with $Var(y_T) = N^2(1-f)S^2/n$ and all properties (unbiasedness, minimum variance, confidence intervals, required sample size etc) transfer immediately from those of $\bar{y}$.

With sr sampling we sample with equal probabilities. Non-equal probability (***non-epsem***) schemes are also important – e.g. sampling with ***probability proportional to size*** (***pps***) with the ***Hansen-Hurwitz*** and ***Horvitz-Thompson estimators***.

Broader access to finite population sampling methods (survey sampling, opinion polls etc) is provided by the following brief bibliography.

7 Bibliography

Barnett, V. (2002) *Sample Survey Principles and Methods* 3rd Ed. Arnold, London

Barnett, V. (2004) *Environmental Statistics,* Wiley, Chichester

Buckland, S.T., Anderson, D.R., Burnham, K.P. and Laake, J.L. (1993) *Distance Sampling: Estimating Abundance in Biological Populations,*Chapman & Hall

Cochran, W.G. (1977) *Sampling Techniques*, 3rd Ed. Wiley, New York

Dillman, D.A.(2000) *Mail and Internet Surveys* 2nd Ed. Wiley, New York

Groves, R.M. (1989) *Survey Errors and Survey Costs.* Wiley, New York

Lohr, S (2010). *Sampling: Design and Analysis.* 2nd edition. Boston: Brooks/Cole.

Peterson, R.A. Constructing Effective Questionnaires *Sage, Thousand Oaks CA.*

Thompson, S.K. (1992) *Sampling* Wiley, New York

$\bar{y}_{cl}$ follow from the results for basic sr sampling. Thus $\bar{y}_{cl}$ is unbiased for $\bar{Y}$ with $Var(\bar{y}_{cl}) = \frac{1-f}{m}\sum_1^M (\bar{Y}_i - \bar{Y})^2/(M-1)$ since $\bar{y}_{cl}$ can be expressed as $\sum_1^m \bar{y}_i / m$ where the $\bar{y}_i$ are cluster sample means for a sr sample of m of the M clusters.

Consider the alternative of a sr sample-type mean of size $n = mL$ drawn from the whole population. We find that $Var(\bar{y}) - Var(\bar{y}_{cl})$ is a positive multiple of $\bar{S}^2 - S^2$ where $\bar{S}^2 = \sum_1^m S_i^{\,2}/M$ is the average within-cluster variance and so we conclude that $\bar{y}_{cl}$ will be more efficient than $\bar{y}$ if the average within-cluster variance is *larger* than the overall population variance. This is essentially the opposite of what we found for stratified sampling. But it is the administrative convenience of cluster sampling which is its main appeal. We will of course have to estimate $Var(\bar{y}_{cl})$ by the obvious sample analogue to make use of the above results.

What if the clusters are not all the same size? This is a more realistic scenario and there are three types of estimator that are used: **the *cluster sample ratio***, or if the total population size is known the ***cluster sample total*** or (as a useful quick estimate) the ***unweighted average of the chosen cluster sample means***. Alternatively, and usefully, we can replace sr sampling with some scheme with varying chances of choosing a sample member: in particular, by using *sampling with probability proportional to size.* (See Barnett, 2002, sections 5.2, 5.3 for further details.)

Cluster sampling in practice is usually employed in more complex ***multi-stage sampling schemes*** where the selected clusters in the primary cluster sample are themselves sub-sampled perhaps on various different criteria but these more complicated sampling methods take us beyond our brief in this introductory review.

6 Concluding remarks

In this short guide, we have discussed many basic concepts in finite population sampling: considering the defining issues of inference for finite populations, the distinctive fundamental sampling schemes and many of the practical results for taking samples and inferring properties of the population. More detailed study would take us into a range of other considerations of which the following are examples.

- Practical aspects of carrying out a survey: sources of error, pilot studies, interviews and questionnaires, handling non-response.
- Rare and sensitive events: how to access information in such cases, snowball sampling, randomized response, composite sampling, ranked set sampling etc.
- Environmental and wild-life sampling: problems of sampling plants and animals, transect sampling, mark- and capture- recapture methods

Estimating a proportion P Let P be the proportion of population members with some quality A. For each population member define X_i =1 if Y_i has quality A and X_i =0 otherwise. Then clearly $P = \sum_1^N X_i / N = \bar{X}$ and we are *again* concerned with estimating a population mean (now for the derived X-variable). The only difference now is that the population variance depends on its mean, as $S_X^2 = NP(1-P)/(N-1)$ and the previous inferences have to be modified to reflect this. Thus we estimate P by the *sample proportion* $p = \bar{x} = \sum_1^n x_i / n$ which is unbiased with minimum variance $Var(p) = (N-n)P(1-P)/((N-1)n)$ with unbiased estimator $s^2(p) = (1-f)p(1-p)/(n-1)$. An approximate 100(1-α)% two-sided confidence interval for P is now given as the region between the two roots of a quadratic equation which for large n simplifies to $p \pm z_\alpha \sqrt{(1-f)p(1-p)/(n-1)}$. Choice of sample size is now more complex depending on whether we want *absolute* or *relative* accuracy represented as $P(|p-P|>d) \le \alpha$ or $P(|p-P|>\xi P) \le \alpha$, respectively. The first (absolute) form requires $n \ge N\left[1 + \frac{(N-1)}{P(1-P)}\left(\frac{d}{z_\alpha}\right)^2\right]^{-1} = \frac{P(1-P)}{V}\left[1 + \frac{1}{N}\left(\frac{P(1-P)}{V} - 1\right)\right]^{-1}$ if we put $V = (d/z_\alpha)^2$. So as first approximation we have $n_0 = P(1-P)/V$ or more accurately $n = n_0\left[1 + (n_0 - 1)/N\right]^{-1}$ Corresponding results are readily obtained for the ***relative*** case.

3 Ratio and Regression Estimators

We may be interested in many variables $X, Y, \ldots$ in a sample survey. For example, in a household expenditure survey Y may be annual household expenditure and X may be household size. Ratios of these variables may be relevant and interesting. For example ***per capita expenditure*** is represented by the *population ratio* $R = Y_T / X_T = \bar{Y}/\bar{X}$. There are various possible estimators of R based on a simple random sample $(y_1, x_1), (y_2, x_2) \ldots (y_n, x_n)$; in particular there is the *sample average ratio* $r_1 = \frac{1}{n}\sum_1^n (y_i / x_i)$ and the *ratio of the sample averages* $r_2 = \bar{y}/\bar{x} = y_T / x_T$.

How do ratio estimators compare? In spite of its intuitive appeal, r_1 is not widely used to estimate R: it is biased and can have large mean square error (mse) compared with r_2. Consider the population of values $R_i = Y_i / X_i$ with mean $\bar{R}$ and variance S_R^2 then r_1 must have mean $\bar{R}$ and variance $(1-f)S_R^2/n$. But $\bar{R} \ne R$, so the bias of r_1 *is* $\bar{R} - R = -\sum_1^N R_i(X_i - \bar{X})/X_T$. An unbiased estimator of the bias is obtained as $-(N-1)n(\bar{y} - r_1\bar{x})/\left[(n-1)X_T\right]$ and if X_T is known, which is not uncommon, we can correct the bias with an unbiased estimator called the ***Hartley-Ross estimator,*** $r_1' = r_1 + (N-1)n(\bar{y} - r_1\bar{x})/\left[(n-1)X_T\right]$. The mse is readily estimated. Another way to eliminate the bias is to sample with probability proportional to the X_i values rather than using sr sampling.

The estimator r_2 tends to be used more widely than r_1, because even though it is still biased, it is likely to be less so than r_1 and with lower mse. The bias becomes negligible in large samples and the sampling distribution tends to normality.

Asymptotically, $E(r_2) = \bar{Y}/\bar{X} = Y_T/X_T = R$ and

$$Var(r_2) = \frac{(1-f)}{n\bar{X}^2}\sum_1^N \frac{(Y_i - RX_i)^2}{N-1},$$

which can be estimated by

$$s^2(r_2) = \frac{(1-f)}{n\bar{x}^2}\sum_1^n \frac{(y_i - r_2 x_i)^2}{n-1}$$

so that an approximate $100(1-\alpha)\%$ symmetric two-sided confidence interval is given by

$$r_2 - z_\alpha s(r_2) \le R \le r_2 + z_\alpha s(r_2).$$

Frequently two population variables (Y, X) will be correlated and we can exploit the relationship between them to obtain improved estimates of a population mean $\bar{Y}$ or total Y_T using what is called a ***ratio estimator*** or a ***regression estimator.***

Ratio Estimators Suppose we want to estimate the total expenditure, Y_T , of all local authorities on community services from a simple random sample (y_i, x_i) for $i = 1,...,n$ where y_i is authority spend on community services and we also have sampling authority population sizes x_i. We would expect Y to be positively correlated with X and might hope to be able to exploit this relationship. Instead of using the sr sample estimator y_T we might assume that $Y_i \approx RX_i$ so that $Y_T = RX_T$ and if X_T, the total population size, is known (which is not unreasonable) we can estimate Y_T by $y_{TR} = rX_T$ where r is an estimator of the ratio R, as discussed above. We use $r_2 = \bar{y}/\bar{x} = y_T/x_T$ for r. Then $y_{TR} = rX_T = (y_T/x_T)X_T$ is known as the ***sr sample ratio estimator of the population total.*** This provides a natural compensation: if x_T happens to be larger, or smaller, than X_T then the estimate of Y_T is reduced, or increased, accordingly. Of course, the corresponding ***ratio estimator of the population mean*** is just $\bar{y}_R = r\bar{X} = (\bar{X}/\bar{x})\bar{y}$. The properties of these ratio estimators are immediately found from what we discussed about $r = r_2$ above. We see that $\bar{y}_R$ is asymptotically unbiased, sometimes exactly unbiased, and

$$Var(\bar{y}_R) \approx \frac{(1-f)}{n}\sum_1^N \frac{(Y_i - RX_i)^2}{N-1} = \frac{(1-f)}{n}(S_Y{}^2 - 2R\rho_{YX}\, S_Y S_X + R^2 S_X{}^2)$$

where $\rho_{YX} = S_{YX}/(S_Y S_X)$ is the population correlation coefficient. The larger the (positive) correlation, the smaller will be $Var(\bar{y}_R)$ which can be estimated using the results discussed above for $r_2 = r$. Approximate confidence intervals are correspondingly obtained. Properties for the population total using y_{TR} are similarly obtained.

commercial surveys such as government, politics, commerce, opinion polls and so forth. The practical difficulties of conducting such sampling can lead to some lack of representation or randomness of the resulting samples. Non-response from some selected individuals also further complicates the sampling scheme and use of conventional results for stratified sr sampling schemes (single- or multi-factor) may at best be approximations to the actual, but often unassessible, statistical properties of the employed quota sampling method.

5 Cluster sampling

The final broad class of survey sampling schemes we will consider are known as *cluster sampling* methods. Again we are concerned with a population which is naturally divided into sub-populations or strata, possibly on many different criteria, as with stratification. A simple example is where we are sampling individuals from a list of addresses at which they reside. So the sampling unit is an address at which there may be several individuals. In stratified sampling, we take a sr sample *from each stratum*. But for ease of access it is more natural to sample addresses and then use information *on all individuals at that address*. However we will not sample all addresses, that is, all strata in the earlier example – to do so is prohibitive and would imply observing the whole population. Instead we take a sr sample of addresses, which are now known as *clusters* rather than *strata* to distinguish the fact that the sub-populations are typically much smaller than 'strata' and that we observe all individuals in each sampled unit. A one-stage cluster sample is a sr sample of the clusters where we observe all individuals in the sampled clusters.

If we were to take ***samples*** of the individuals in selected clusters we have what is called ***sub-sampling*** or ***two-stage cluster sampling.*** This extends to ***multi-stage cluster sampling*** where we choose from a set of primary units, then from secondary units within the chosen primary units and so on. For example, primary units may be educational authorities, secondary units the schools within them, tertiary units the classes in the schools etc.

In ***one-stage cluster sampling***, the population consists of a set of clusters and we take a sr sample of them. Suppose we have M clusters of sizes $N_1 N_2,...,N_M$ and the cluster means are $\bar{Y}_i$ and within-cluster variances are $S_i{}^2 (i = 1,2,...,M)$. The population mean and variance are again $\bar{Y}, S^2$. We will take a sr sample of m clusters and observe all members of the chosen clusters to obtain a sample of size $n \ge m$.

The simplest case is where all clusters are of the same size L, say. Thus $N = ML$ and so the sampling fraction is $f = n/N = m/M$.
The cluster sample mean $\bar{y}_{cl}$ is the sample average of all the observations and is thus $\bar{y}_{cl} = \sum_1^m \bar{y}_i / m$ for equal-sized clusters. We are effectively just taking a sr sample of m of the M cluster means $\bar{y}_i (i = 1,2,...,M)$ and properties of the estimator

overall cost of taking the stratified sample is $C = c_0 + \sum_1^k c_i n_i$ and then choose the n_i to minimize $Var(\bar{y}_{st})$ for a prescribed fixed overall cost C. Appropriate constrained minimization yields expressions for the stratum sample sizes n_i and overall sample size n for given $c_i (i = 1,2,...,k)$. For the special case, where each observation costs the same amount, c, in each stratum we obtain the allocation $n_i = nW_iS_i / \sum_1^k W_iS_i$ with overall sample size $n = (C - c_0)/c$. This is known as ***Neyman Allocation***. Alternatively, we can prescribe the value V we need for $Var(\bar{y}_{st})$ and choose the allocation to minimise the overall cost. For constant fixed sampling cost (c per observation) we again obtain the Neyman allocation above with overall sample size $n = (\sum W_iS_i)^2 / (V + \sum W_iS_i^2 / N)$.

We can also express our need for precision in terms of a sample size needed to yield a specified margin of error, d, and maximum probability of error, α, in the form $\Pr(|\bar{y}_{st} - \bar{Y}| \geq d) \leq \alpha$. If we assume that $\bar{y}_{st}$ is approximately normally distributed this reverts to the case just discussed with $V = (d / z_\alpha)^2$. We obtain

$$n = \sum (W_i^2 S_i^2 / w_i) / (V + \sum W_i S_i^2 / N)$$

giving as a first approximation to the required sample size $n_0 = \sum (W_i^2 S_i^2 / w_i) / V$ or more accurately $n = n_0 (1 + \sum W_i S_i^2 / (NV))^{-1}$. For the special cases of ***proportional allocation*** and ***Neyman allocation*** we get, respectively,

$$n_0 = \sum W_i S_i^2 / V \,,\; n = n_0 (1 + n_0 / N)^{-1}$$

and

$$n_0 = (\sum W_i S_i)^2 / V \,,\; n = n_0 (1 + \sum W_i S_i^2 / (NV))^{-1} .$$

Is optimal allocation always noticeably more efficient than the convenient proportional allocation? This, of course, does not require stratum variances or relative sampling costs. The simple answer is that the advantage of optimal allocation (specifically Neyman allocation) ***is greater the more the variability of the stratum variances.***

We must recognize that much of the above discussion of stratified sampling implicitly assumes that ***we know the stratum sizes and the stratum variances***. Often this is not so, particularly for the stratum variances. If we have to estimate these from the survey data or assign 'reasonable values' say from previous experience the above results may not be reliable and far more complex methods will need to be employed. These are not pursued in this brief review.

Quota Sampling Often we will want to exploit many (crossed) factors of stratification e.g. age ranges, locations, types of individual etc. and complex methods of sampling for multi-factor stratification must be used. One form of such stratified sampling is called *quota sampling* in which proportional allocation is used with respect to the various crossed factors and samplers seek to fill the 'quotas' implied for the various allocations. This is the method used predominantly in

Under what circumstances are $\bar{y}_R$ and y_{TR} more efficient (have smaller sampling variance) than the sr sample estimators $\bar{y}$ and y_T? It can be shown that this will happen if $\rho_{YX} \geq C_X / (2C_Y)$ where $C_X = S_X / \bar{X}; C_Y = S_Y / \bar{Y}$ are the *coefficients of variation*. Any efficiency gain clearly requires $C_X \leq 2C_Y$, but efficiency gains can be quite high if the correlation between Y and X is highly positive.

Regression Estimators Ratio estimators are especially beneficial when there is a degree of proportionality between the two variables Y and X; the more so the higher the correlation. When there is rough linearity between the principal variable Y and the auxiliary variable X, but this is not through the origin (i.e. there is no 'proportionality'), the link between Y and X can be exploited to improve sr sample estimators by using so-called ***regression estimators***.

The ***linear regression estimator*** of $\bar{Y}$ is $\bar{y}_L = \bar{y} + b(\bar{X} - \bar{x})$ for a suitable choice of b reflecting any (even a rough) linear regression relationship between Y and X. It is readily confirmed that this produces an appropriate compensation depending on the sign of b. Of course, Y_T can be estimated by $N\bar{y}_L$. We might ***pre-assign*** a value of b or ***estimate*** it. In the former case $\bar{y}_L$ is clearly unbiased (as is $N\bar{y}_L$) and its variance is

$$Var(\bar{y}_L) = \frac{1-f}{n}(S_Y^2 - 2bS_YS_X + b^2S_X^2)$$

with corresponding unbiased sample estimate

$$s^2(\bar{y}_L) = \frac{1-f}{n}(s_Y^2 - 2bs_Ys_X + b^2s_X^2) .$$

$Var(\bar{y}_L)$ will take a minimum value $MinVar(\bar{y}_L) = \frac{1-f}{n} S_Y^2 (1 - \rho^2_{YX})$ if b is chosen as $b_0 = \rho_{YX}(S_Y / S_X)$ so that irrespective of any relationship between Y and X in the population, $\bar{y} + \rho_{YX} \frac{S_Y}{S_X}(\bar{X} - \bar{x})$ is the most efficient estimator of $\bar{Y}$ in the form of $\bar{y}_L$. However, $b_0 = \rho_{YX}(S_Y / S_X)$ will not be known. So if there is no basis for an *a priori* assignment of a value for b_0 we will need to estimate b; usually we would use the sample analogue $\tilde{b} = \frac{s_{YX}}{s_X^2} = \frac{\sum_1^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_1^n (x_i - \bar{x})^2}$. So the linear regression estimator of $\bar{Y}$ is now $\bar{y}_L = \bar{y} + \tilde{b}(\bar{X} - \bar{x})$. Its distributional properties are difficult to determine but it is found to be asymptotically unbiased with approximate variance $\frac{1-f}{n} S_Y^2 (1 - \rho^2_{YX})$ (estimated by $s^2(\bar{y}_L) = \frac{1-f}{n}(s_Y^2 - \tilde{b}s_{YX})$) so that having to estimate b in large samples is no disadvantage.

Clearly $\bar{y}_L$ can be no less efficient than $\bar{y}$ and since $Var(\bar{y}_R) - Var(\bar{y}_L) \approx \frac{1-f}{n}(RS_X - \rho_{YX}S_Y)^2$ it must be at least as efficient (asymptotically) as the ratio estimator with equality of variance only if $R = \rho_{YX}\frac{S_Y}{S_X}$.

Summary of the use of ratio and regression estimators They are useful in estimating $\bar{Y}$ (or Y_T) when there is an auxiliary variable X (with known population mean $\bar{X}$) also sampled. If Y and X bear some reasonable degree of linear relationship then we obtain a useful increase in efficiency over $\bar{y}$ (or y_T) by using $\bar{y}_L$ (or y_{TL}) – if the relationship is one of rough proportionality we expect similar benefits but for somewhat less computational effort from ratio estimators.

4 Stratified sampling

Sometimes a finite population is divided naturally into non-overlapping exhaustive sub-populations or *strata* e.g. in a school survey, different local education authorities make up distinct strata. There can be an administrative advantage in taking separate sr samples of prescribed size from each stratum (a *stratified sr sample*) rather than taking an overall sr sample from the whole population. If we have k strata of sizes $N_i; i = 1,2,...,k$ we can estimate the population mean $\bar{Y}$ by the *stratified sr sample mean* $\bar{y}_{st} = \sum_1^k W_i \bar{y}_i$ where $W_i = N_i / N$ and $\bar{y}_i$ is the sample mean of the sr sample of size n_i chosen from the i^{th} stratum. The overall sample size is $n = \sum_1^k n_i$. We note that $\bar{Y} = \sum_1^k W_i \bar{Y}_i$ and that $S^2 = \frac{1}{N-1}\left[\sum_1^k (N_i - 1)S_i^2 + \sum_1^k N_i(\bar{Y}_i - \bar{Y})^2\right]$. It is easily confirmed that $\bar{y}_{st}$ is unbiased for $\bar{Y}$ with $Var(\bar{y}_{st}) = \sum_1^k W_i^2(1-f_i)S_i^2/n_i$ with $f_i = (n_i / N_i)$. If the $f_i = f$ (constant sampling fractions), we have what is called *proportional allocation* in which case $Var(\bar{y}_{st}) = \frac{1-f}{n}\sum_1^k W_i S_i^2$. If the sampling fractions are negligible, we have $Var(\bar{y}_{st}) = \sum_1^k W_i^2 S_i^2 / n_i$. We estimate $Var(\bar{y}_{st})$ using sample analogues $s_i^2 = \frac{1}{n_i - 1}\sum_1^{n_i}(y_{ij} - \bar{y}_i)^2 (i = 1,2,...k)$ for the typically unknown S_i^2. Analogous results hold for estimating the population total Y_T.

For proportions we have corresponding results. With a derived variable X which is 0 or 1 depending on whether or not the population member possesses the attribute of interest, then the population mean $\bar{X} = P$ is the proportion of population

members with the attribute. So the stratified sr sample mean $\bar{x}_{st} = \sum_1^k W_i \bar{x}_i$ provides an unbiased estimator of P in the form $p_{st} = \sum_1^k W_i p_i$ where p_i is the sampled proportion in the i^{th} stratum with population proportions $P_i (i = 1,2,...,k)$. (ignoring terms in $1/n_i$) with unbiased estimate $s^2(p_{st}) = \sum_1^k W_i^2 (1-f_i) p_i (1-p_i)/(n_i - 1)$.

For proportional allocation $Var(p_{st}) = \frac{1-f}{n}\sum_1^k W_i P_i (1-P_i)$.

Some key questions

- is the stratified sr sample mean $\bar{x}_{st}$ more efficient than the sr sample mean $\bar{x}$?
- how should we choose the stratum sample sizes, $n_i (i = 1,2,...,k)$?
- what do we do if the stratum sizes N_i and stratum variances S_i^2 are unknown (a special problem if determining the allocation of stratum sample sizes)?

We can compare the efficiencies of $\bar{y}_{st}$ and $\bar{y}$ by examining, firstly for proportional allocation, $Var(\bar{y}) - Var(\bar{y}_{st}) \approx \frac{1-f}{n}\sum_1^k W_i (\bar{Y}_i - \bar{Y})^2$ (if the stratum sizes N_i are large enough). But this is always non-negative, so that $\bar{y}_{st}$ must always be at least as efficient as $\bar{y}$.

More detailed investigation tempers this simple conclusion. We find that the stratified sr sample mean $\bar{y}_{st}$ is more efficient than the sr sample mean $\bar{y}$ provided

$$\sum_1^k N_i(\bar{Y}_i - \bar{Y})^2 > \frac{1}{N}\sum_1^k (N - N_i)S_i^2$$

i.e. *if the variation between the stratum means is sufficiently large compared with the within-strata variation* so that the higher the variability in stratum means and the lower the accumulated within-stratum variability the greater the advantage in using the stratified sr sample mean (or corresponding estimators of population total or proportion).

How do we allocate the stratum sample sizes $n_i (i = 1,2,...,k)$? There is a clear practical advantage in stratified sr sampling. With naturally defined strata it will usually be more economical and more convenient to sample separately from each stratum. We have now seen that it can also lead to more efficient estimators than those obtained from overall simple random sampling. As far as allocation of stratum sample sizes is concerned, proportional allocation has intuitive appeal, is easy to operate and can lead to efficiency gains. But for more effort we might be able to do better by choosing the $n_i (i = 1,2,...,k)$ optimally, that is, to minimise $Var(\bar{y}_{st})$ for given overall sample size or cost. Specifically, we can assume that the