Wiley StatsRef:
Statistics Reference Online

Two-Phase Sampling

By *K. J. Keen*

1 Historical Background

The* theory of two-phase sampling was created by Jerzy Neyman[8] in response to a problem posed at a conference on sampling human populations in April, 1937. Neyman[8] introduces his solution "by describing the problem in much the same form as it was stated to me, without using any mathematical symbols": Simply put, a field survey is to be undertaken to determine the average value of some character of a population; for example, the amount of money families spend on food. As the collection of data requires long interviews by specially trained enumerators, the cost per family is quite high. The cost of the survey is constrained within a specified amount but the sample does not appear to yield an estimate of desired precision because of the great variability of the character. Nevertheless, the character is correlated with a second character that can be determined at a lower cost per family so that a precise estimate of the distribution of this second character is readily obtained. Hence, a more precise estimate of the original character can be found by first estimating the distribution of the second character alone from a large random sample, then dividing this sample, as in stratified sampling, into classes or strata according to the value of the second character and then drawing* at random from each of the strata a small sample for the costly procedure of measuring the first character.

Neyman[8] called this method *double sampling*, and this term remains in use among statisticians working in the area of quality control and assurance. Survey statisticians, however, tend instead to use the term *two-phase sampling* so that this method is distinguished from *two-stage sampling*. Two-phase sample designs differ from two-stage sample designs in that the stratification occurs after the first sample is collected (i.e. *post hoc*) in the case of two-phase designs rather than before the first sample is collect (i.e. *ante hoc*) in the case of two-stage designs. It is understandably regrettable and a source of confusion that the biometrics literature refers to a two-phase survey sampling design as a two-stage design, as noted by Whittemore[13].

In the case of genetic epidemiology, the interest lies primarily with the estimation of means associated with Bernoulli distributed random variables, such as disease prevalence and allele frequencies (Gene), rather than economic variables. Nevertheless, two-phase sampling designs are applicable to quantitative phenotypes.

University of Northern British Columbia, Prince George, British Columbia, Canada

This article was originally published online in 2005 in Encyclopedia of Biostatistics, © John Wiley & Sons, Ltd and republished in Wiley StatsRef: Statistics Reference Online, 2014.
* This article was published on 29th September 2014. A revision was made at this location on 6th April 2016.

This article is © 2016 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118445112.stat05440

2 Formulation and Allocation

Assume a finite population size N of which n' are to be selected by a simple random sample without replacement at the first phase of the design. Suppose that there are determined to be K strata with (possibly unknown) population size N_h for the hth stratum. Suppose n'_h sample units are observed to be in the hth stratum in the first phase sample with a sample mean of $\bar{y}_h$. Let $\overline{Y}_h$ denote the population mean and S_h^2 the population variance for the hth stratum. Let s_h^2 denote the usual unbiased estimator of S_h^2 based on the n'_h sample units in the first phase. Let n_h denote the sample size for the hth stratum at the second phase. For convenience, let $W_h = N_h/N$, $w_h = n'_h/n'$, and $\nu_h = n_h/n'_h$ with $0 < \nu_h \leq 1$ for all h. Assuming that $\Pr(n'_h = 0) = 0$ for all h, Rao[9] showed that an unbiased estimator of the population mean $\overline{Y} = \Sigma W_h \overline{Y}_h$ is $\bar{y} = \Sigma w_h \bar{y}_h$ with variance

$$\text{var}(\bar{y}) = \left(\frac{1}{n'} - \frac{1}{N}\right) S^2 + \sum_{h=1}^{K} \frac{W_h S_h^2}{n'} \left(\frac{1}{\nu_h} - 1\right),$$

where S^2 is the population variance. These results are available elsewhere in the literature, but Rao[9] obtained them under the assumption that the second-phase sample sizes $\{n_h\}$ for the strata are random variables, unlike Cochran[1] who assumed that they are fixed values. Rao[9] further showed that a nonnegative unbiased estimator of $\text{var}(\bar{y})$ is

$$v(\bar{y}) = \frac{1}{Nn'} \left[\left(\frac{N-1}{n'-1}\right) \sum_{h=1}^{K} n'_h s_h^2 \left(\frac{1}{\nu_h} - 1\right)^* + \left(\frac{N-n'}{n'-1}\right) \left(\sum_{h=1}^{K} \frac{1}{\nu_h} \sum_{j=1}^{n_h} y_{h_j}^2 - n' \bar{y}^2\right)\right]^*,$$

provided n' is sufficiently large so that $\Pr(n_h \geq 2) = 1^*$ for all h. Särndal & Swensson[10] showed that the result for $v(\bar{y})$ continues to be valid if K is a random variable or if there is random nonresponse at the second phase described by a Bernoulli distribution with a fixed but unknown probability of inclusion within each stratum with the possibility that the probability of inclusion varies among strata.

With respect to the optimal allocation of the first-phase sample size n' and the second-sample sampling fractions $\{\nu_h\}$, the cost function is taken as

$$C = n' c' + \sum_{h=1}^{K} n_h c_h,$$

where c' is usually much smaller than c_h. Since C is a random variable, we take

$$C^* = \text{E}(C) = n' \left(c' + \sum_{h=1}^{K} W_h c_h \nu_h\right).$$

From the Cauchy inequality, the optimal ν_h for the hth stratum for given C^* and $\text{var}(\bar{y})$ is

$$\nu_h = S_h \left[\frac{c'}{c_h(S^2 - \sum W_h S_h^2)}\right]^{1/2}.$$

* This article was published on 29th September 2014. A revision was made at this location on 6th April 2016.

Wiley StatsRef: Statistics Reference Online, © 2014–2016 John Wiley & Sons, Ltd.
This article is © 2016 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118445112.stat05440

As noted by Singh & Singh[12], it is important to realize that the upper limit on the second-phase sample size is n'_h if randomly sampled without replacement from the first-phase sample. As suggested by Rao[9] in this case for which $\nu_h > 1$, set the corresponding $\nu_h = 1$ and repeat the procedure until all the $\nu_h \leq 1$.

By Rao[9], if the strata weights $\{W_h\}$ are not known, then the subsampling fraction $\nu_h = n_h/n'_h$ varies as a function of the observed value of n'_h. Nevertheless, in this case, replace W_h by its estimate w_h.

3 Bayesian Approaches

Draper & Guttman[3] assumed that the observations $\{y_{hj}\}$ are normally distributed with independent improper prior distributions for the mean μ_h and variance σ_h^2 of the hth stratum given by

$$p(\mu_h)\mathrm{d}\mu_h \propto \mathrm{d}\mu_h, \qquad p(\sigma_h^2)\ \mathrm{d}\sigma_h^2 \propto \frac{\mathrm{d}\sigma_h^2}{\sigma_h^2}.$$

Draper & Guttman[3] further assumed that C, K, and $\{n'_h\}$ are fixed with $\Sigma n_h c_h < C$ but with the prior information concerning the means and variances of the strata available before the first phase and showed that the posterior distribution of

$$T_h = \frac{(n'_h + n_h - 1)\ (\mu_h - \tilde{\mu}_h)}{\tilde{\sigma}_h}$$

is Student's* t with $n'_h + n_h - 1$ degrees of freedom where, if $\{x_{hj}\}$ denotes the observations from the first phase,

$$\tilde{\mu}_h = \frac{n'_h\ \overline{x}_h + n_h\ \overline{y}_h}{n'_h + n_h},$$

and

$$\tilde{\sigma}_h^2 = \frac{1}{n'_h + n_h}\ \left[\left(\frac{n'_h\ n_h}{n'_h + n_h}\right)\ (\overline{x}_h - \overline{y}_h)^2 + (n'_h - 1)\ s_h^2 + (n_h - 1)\ t_h^2\right]$$

with $\overline{x}_h$ and s_h^2 the usual unbiased estimators of the mean and variance of the h stratum from the first phase and $\overline{y}$ and t_h^2, respectively, from the second phase. Furthermore, Draper & Guttman[3] showed that the posterior distribution of $(n_h - 1)\ s_h^2/\sigma_h^2$ is ${\chi^2_{n'_h-1}}^*$. From these results concerning the posterior distributions, the posterior exception of $\mu = \Sigma_h W_h \mu_h$ is

$$\sum_{h=1}^{K} W_h^2\ \frac{(n'_h - 1)\ s_i^2}{(n'_h + n_h)\ (n'_h - 3)}.$$

Choosing the sample size n_h for the h stratum at the second phase subject to $n_h \geq 0$ for all h leads to

$$n_h = \frac{C}{c_h}\ q_h - n'_h,$$

where

$$q_h = \frac{\left(\frac{n_h - 1}{n_h - 3}\right)^{1/2}\ W_h\ s_h\ \sqrt{c_h}}{\sum_h\ \left(\frac{n_h - 1}{n_h - 3}\right)^{1/2}\ W_h\ s_h\ \sqrt{c_h}}.$$

* This article was published on 29th September 2014. A revision was made at this location on 6th April 2016.

This article is © 2016 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118445112.stat05440

There is the possibility that this allocation rule will lead to negative n_h for some strata. This merely indicates that the hth statum has been oversampled. Draper & Guttman[3] discussed an algorithmic adjustment to the optimal allocation rule to compensate for this.

If, on the other hand, the posterior after the first phase is used to provide the prior for the second phase, then by Draper & Guttman[3], the optimal allocation rule becomes instead $n_h \propto Cq_h/c_h$. Compare this with the optimal allocation rule

$$C \frac{W_h S_h}{\sum_h W_h S_h}$$

of Neyman[8] assuming $\{W_h\}$ are known with n' and $\sum_h n_h$ fixed. This so-called *Neyman allocation* can also be obtained from the expression derived by Rao[9] for var$(\bar{y})$ using the Lagrange multiplier.

Draper & Guttman[3] also considered the situation when the strata weights $\{W_h\}$ are no longer known but rather follow a Dirichlet prior distribution with parameter ν_h corresponding to the hth stratum. In this case, the answer is the same as the case for $\{W_h\}$ known except that the unbiased estimator

$$\tilde{w}_h = \frac{n'_h + \nu_h}{\sum_h (n'_h + \nu_h)}$$

replaces W_h everywhere. Note that Jeffrey's prior coincides with the uniform prior (all $\nu_h = 1$).

For a multivariate normal extension to this approach, see Draper & Guttman[4]. Although the approach of Draper & Guttman[3,4] is suitable for quantitative phenotypes with a normally distributed likelihood, it is not suitable for estimation of disease prevalence or allele frequencies for which a solution is given by Zacks[15] assuming a hypergeometric likelihood and a discrete uniform prior distribution for the number of successes out of the number of trials for the hth stratum. For a heterogeneous situation in which the prevalence varies among strata, see the optimal allocation rules of Newbold[7] which assumes the invariant Jeffreys' prior distribution

$$p(P_h) \propto P_h^{-1/2} (1 - P_h)^{-1/2}$$

and a binomial likelihood for the parallel cases comparable to those of Draper & Guttman[3].

4 Prevalence Estimation and Practical Considerations

For a Bayesian solution to the problem of estimation of prevalence in a two-phase sampling design using a Markov chain Monte Carlo method for a Dirichlet conjugate prior distribution for sensitivity, specificity, and prevalence jointly with a beta posterior distribution, see Erkanli et al.[5]. On the other hand, the results of Neyman[8] and Rao[9] do not assume a distributional form for the likelihood and thus apply to the problem of estimation of disease prevalence. While these results do assume a finite population, the hypothetical case of an infinite population is easily derived as a limiting case.*

As discussed by Deming[2], a two-phase design is not necessarily more efficient than a one-phase design, nor is Neyman allocation necessarily more efficient than *proportional allocation*: $n_h \propto w_h$ for all h.

Calculations in Deming[2], suggest that, as a rule-of-thumb, it is only when the ratio of interview cost per sampling unit at the second phase compared with screening cost per sampling unit at the first phase exceeds 6:1 that two-phase sampling will be more advantageous. Note that the ratio is likely to be high when the screening and stratification is done on the basis of records, typically on the order of 40:1 or 100:1 according to Deming[2].

* This article was published on 29th September 2014. A revision was made at this location on 6th April 2016.

Wiley StatsRef: Statistics Reference Online, © 2014–2016 John Wiley & Sons, Ltd.
This article is © 2016 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118445112.stat05440

A sample design using Neyman allocation that incorporates an estimate of the proportion of false negatives that is wide of the mark may well yield an estimate of the prevalence with greater variance than the estimate by proportional sampling. Deming[2] also noted that it is easy "to fall into the trap in the planning stages by putting unwarranted credence into an advance estimate" of the proportion of false positives when in fact a large sample or a long history of usage of the exact plan of screening is required. The example of the heavy workload encountered by a psychiatrist interviewing 30 subjects for a pilot study is cited despite the fact that the estimate of a small proportion of false positives in such a small sample is subject to a wide standard error. Whereas, a fairly large preliminary sample will often reveal problems that one would not otherwise foresee, for example, a set of admission records intended to contain individuals only aged 21 to 60 but actually including admissions of age 20 and under.

For a discussion of two-phase sampling designs in the context of prevalence for a rare disease for which all those screened positive in the first phase must ethically be included in the second phase, see Shrout & Newman[11].

For a discussion concerning the estimation of disease prevalence with nonresponse at the second phase, see Särndal & Swensson[10], and Gao et al.[6] for a representation incorporating a logistic model for nonresponse that is not completely random.

A maximum likelihood approach for the multinomial distribution with discussion of options involving the EM algorithm and thebootstrap method are given in Zhou et al.[16] together with a likelihood ratio test for the null hypothesis of completely random nonresponse.

5 Multistage Sampling in Genetic Epidemiology

One of the greatest challenges to successfully concluding a disease–marker association study is heterogeneity in the distribution of alleles among races, ethnic and regional groups (Bias in Case-Control Studies). For example, cystic fibrosis (CF) can be caused by many different mutations. The most common mutation in the North American non-Ashkenazi population is Δ508. The proportion of CF genes that are Δ508 varies widely among different countries, within a country, and among different ethnic and racial groups. But in the case of CF, these observations were noted after the gene was successfully cloned.

A case–control study using a two-phase design is discussed by Whittemore & Halpern[14] in which men were asked whether they were diagnosed with prostate cancer and whether they had a first degree male relative with prostate cancer at the first phase. The subjects are stratified according to diagnosis and family history in preparation for second-phase sampling. The parameters of interest in this study were prostate cancer hazard rates in carriers and noncarriers and the probability that an arbitrary allele contains a deleterious mutation. The study budget could accommodate a second-phase sample of size 570 from the first-phase sample of size 1500. Calculations showed that Neyman allocation of 570 sampling units resulted in little loss of efficiency compared with a complete one-phase sample of 3000 with respect to the variances of the three parameter estimators.

With respect to the theoretical discussion in Whittemore & Halpern[14], the use of the Horvitz–Thompson estimating equation for multiphase sampling is treated in greater detail in Whittemore[13] where it is noted that although it can yield estimates less efficient than the maximum likelihood estimates, substantial efficiency loss appears to occur chiefly when multiphase sampling is unnecessary.

References

[1] Cochran, W. G. (1963). *Sampling Techniques*, 2nd Ed. Wiley, New York.

[2] Deming, W. E. (1977). An essay on screening, or on a two-phase sampling, applied to surveys of a community, *International Statistical Review* **45**, 29–37.

This article is © 2016 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118445112.stat05440

[3] Draper, N. R. & Guttman, I. (1968). Some Bayesian stratified two-phase sampling results, *Biometrika* **55**, 131–139.
[4] Draper, N. R. & Guttman, I. (1968). Bayesian stratified two-phase sampling results: *k* characteristics, *Biometrika* **55**, 587–589.
[5] Erkanli, A., Soyer, R. & Stangl, D. (1997). Bayesian inference in two-phase prevalence studies, *Statistics in Medicine* **16**, 1121–1133.
[6] Gao, S., Hui, S. L., Hall, K. S. & Hendrie, H. C. (2000). Estimating disease prevalence from two-phase surveys with non-response at the second phase, *Statistics in Medicine* **19**, 2101–2114.
[7] Newbold, P. (1971). Optimum allocation in stratified two-phase sampling for proportions, *Biometrika* **58**, 587–589.
[8] Neyman, J. (1938). Contributions to the theory of sampling human populations, *Journal of the American Statistical Association* **33**, 101–116.
[9] Rao, J. N. K. (1973). On double sampling for stratification and analytical surveys, *Biometrika* **60**, 125–133.
[10] Särndal, C. -E. & Swensson, B. (1987). A general view of estimation for two phases of selection with application to two-phase sampling and nonresponse, *International Statistical Review* **55**, 279–294.
[11] Shrout, P. E. & Newman, S. C. (1989). Design of two-phase prevalence surveys of rare disorders, *Biometrics* **45**, 549–555.
[12] Singh, B. D. & Singh, D. (1965). Some remarks on double sampling for stratification, *Biometrika* **52**, 587–590.
[13] Whittemore, A. S. (1997). Multistage sampling designs and estimating equations, *Journal of the Royal Statistical Society, Series B* **59**, 589–602.
[14] Whittemore, A. S. & Halpern, J. (1997). Multi-stage sampling in genetic epidemiology, *Statistics in Medicine* **16**, 153–167.
[15] Zacks, S. (1970). Bayesian design of single and double stratified sampling for estimating proportion in finite population, *Technometrics* **12**, 119–130.
[16] Zhou, X. -H., Castelluccio, P., Hui, S. L. & Rodenberg, C. A. (1999). Comparing two prevalence rates in a two-phase design study, *Statistics in Medicine* **18**, 1171–1182.

Wiley StatsRef: Statistics Reference Online, © 2014–2016 John Wiley & Sons, Ltd.
This article is © 2016 John Wiley & Sons, Ltd.
DOI: 10.1002/9781118445112.stat05440