PracTools: Computations for Design of Finite Population Samples

by Richard Valliant, Jill A. Dever, and Frauke Kreuter

Abstract PracTools is an R package with functions that compute sample sizes for various types of finite population sampling designs when totals or means are estimated. One-, two-, and three-stage designs are covered as well as allocations for stratified sampling and probability proportional to size sampling. Sample allocations can be computed that minimize the variance of an estimator subject to a budget constraint or that minimize cost subject to a precision constraint. The package also contains some specialized functions for estimating variance components and design effects. Several finite populations are included that are useful for classroom instruction.

Introduction

Samples from finite populations are one of the mainstays of research in demographics, economics, and public health. In the U.S., for example, the Consumer Price Index is based on samples of business establishments and households (Bureau of Labor Statistics, 2013, chap. 17); the unemployment rate is estimated from the Current Population Survey, which is a sample of households (Bureau of Labor Statistics, 2013, chap. 1); and various health characteristics of the population are estimated from the National Health Interview Survey (Center for Disease Control and Prevention, 2013b) and the National Health and Nutrition Examination Survey (Center for Disease Control and Prevention, 2013a) both of which are household surveys. Smaller scale academic and marketing research surveys are also typically done using finite population samples.

Standard techniques used in sample design are stratification, clustering, and selection with varying probabilities. Depending on the units to be surveyed (e.g., persons, schools, businesses, institutions) and the method of data collection (e.g., telephone, personal interview, mail survey), the samples may be selected in one or several stages. There are several packages in R that can select samples and analyze survey data. Among the packages for sample selection are **pps** (Gambino, 2012), **sampling** (Tillé and Matei, 2013), **samplingbook** (Manitz, 2013), and **simFrame** (Alfons et al., 2010). The **survey** package (Lumley, 2010, 2014) has an extensive set of features for creating weights, generating descriptive statistics, and fitting models to survey data.

A basic issue in sample design is how many units should be selected at each stage in order to efficiently estimate population values. If strata are used, the number of units to allocate to each stratum must be determined. In this article, we review some basic techniques for sample size determination in complex samples and the package **PracTools** (Valliant et al., 2015) that contains specialized routines to facilitate the calculations, most of which are not found in the packages noted above. We briefly summarize some of selection methods and associated formulas used in designing samples and describe the capabilities of **PracTools**. The penultimate section presents a few examples using the **PracTools** functions and the final section is a conclusion.

Designing survey samples

Complex samples can involve any or all of stratification, clustering, multistage sampling, and sampling with varying probabilities. This section discusses these techniques, why they are used, and formulas that are needed for determining sample allocations. Many texts cover these topics, including Cochran (1977), Lohr (1999), Särndal et al. (1992), and Valliant et al. (2013).

Simple random sampling

Simple random sampling without replacement (*srswor*) is a method of probability sampling in which all samples of a given size n have the same probability of selection. The function `sample` in the **base** package in R can be used to select simple random samples either with or without replacement. One way of determining an *srswor* sample size is to specify that a population value θ be estimated with a certain coefficient of variation (CV) which is defined as the ratio of the standard error of the estimator, $\hat{\theta}$, to the value of the parameter: $CV(\hat{\theta}) = \sqrt{Var(\hat{\theta})}/\theta$. For example, suppose that y_k is a value associated with element k, U denotes the set of all elements in the universe, N is the number of elements in the population, and the population parameter to be estimated is the mean, $\bar{y}_U = \sum_{k \in U} y_k / N$. With a simple random sample, this can be estimated by the sample mean, $\bar{y}_s = \sum_{k \in s} y_k / n$, where s is the set

of sample elements and n is the sample size. Setting the required CV of $\bar{y}_s$ to some desired value CV_0 in an *srswor* leads to a sample size of

$$n = \frac{\frac{S_U^2}{\bar{y}_U^2}}{CV_0^2 + \frac{S_U^2}{N\bar{y}_U^2}}, \tag{1}$$

where S_U^2 is the population variance of the y_k's. The term $S_U^2/\bar{y}_U^2$ is referred to as the *unit relvariance*. If y_k is a 0/1 variable identifying whether an element has a characteristic or not, then $S_U^2 = N(N-1)^{-1}p_U(1-p_U)$ where p_U is the proportion in the population with the characteristic. The function `nCont` in **PracTools** will make the computation in (1). In a real application, the population values in (1) must be estimated from a sample or guessed based on prior knowledge.

Another way of determining a sample size is to set a tolerance for how close the estimate should be to the population value. If the tolerance (sometimes called the *margin of error*) is e_0 and the goal is to be within e_0 of the population mean with probability $1-\alpha$, this translates to requiring $Pr\left(|\bar{y}_s - \bar{y}_U| \leq e_0\right) = 1-\alpha$. This is equivalent to setting the half-width of a $100(1-\alpha)\%$ normal approximation, two-sided confidence interval (CI) to $e_0 = z_{1-\alpha/2}\sqrt{V\left(\bar{y}_s\right)}$. The notation z_ϵ denotes the 100ϵ percentile of the standard normal distribution. The sample size required to accomplish this is

$$n = \frac{z_{1-\alpha/2}^2 S_U^2}{e_0^2 + z_{1-\alpha/2}^2 S_U^2/N}\ . \tag{2}$$

One could also require that the relative absolute error, $|(\bar{y}_s - \bar{y}_U)/\bar{y}_U|$, be less than e_0 with a specified probability. In that case, (2) is modified by replacing S_U^2 with the unit relvariance, $S_U^2/\bar{y}_U^2$. Both calculations can be made using the function `nContMoe` in **PracTools**. When estimating a proportion, there are options other than a normal approximation confidence interval on p_U for setting a margin of error. Two are to work with the log-odds, $p_U/(1-p_U)$, or to use the method due to Wilson (1927), which are both available in **PracTools**.

Another estimand in a survey might be the difference in means or proportions. The difference could be between two disjoint groups or between the estimates for the same group at two different time periods. The standard approach in such a case would be to find a sample size that will yield a specified power for detecting a particular size of the difference. The functions `power.t.test` and `power.prop.test` in the R package **stats** will do this for independent simple random samples.

The case of partially overlapping samples can also be handled (e.g., see Woodward 1992). For example, persons may be surveyed at some baseline date and then followed-up at a later time. An estimate of the difference in population means may be desired, but the samples do not overlap completely because of dropouts, planned sample rotation, or nonresponse. Such non-overlap would be common in panel surveys. Suppose that s_1 and s_2 are the sets of sample units with data collected only at times 1 and 2, and that s_{12} denotes the overlap. Thus, the full samples at times 1 and 2 are $s_1 \cup s_{12}$ and $s_2 \cup s_{12}$. Also, suppose that the samples at the two time periods are simple random samples. Assume that the samples at times 1 and 2 are not necessarily the same size, so that $n_1 = rn_2$ for some positive number r. The samples might be of different sizes because of other survey goals or because the budget for data collection is different for the two times. A case that is covered by the analysis below is one where an initial sample of size n_1 is selected, a portion of these respond at time 2, and additional units are selected to obtain a total sample of size n_2 for time 2. Taking the case of simple random sampling, the difference in means at the first and second time points can be written as

$$\hat{\bar{d}} = \hat{\bar{x}} - \hat{\bar{y}} = \frac{1}{n_1}\sum_{s_1} x_i - \frac{1}{n_2}\sum_{s_2} y_i + \sum_{s_{12}}\left(\frac{x_i}{n_1} - \frac{y_i}{n_2}\right)\ .$$

The variance can be expressed as

$$Var\left(\hat{\bar{d}}\right) = \frac{\sigma_x^2}{n_1} + \frac{\sigma_y^2}{n_2} - 2\sigma_{xy}\frac{n_{12}}{n_1 n_2}, \tag{3}$$

where σ_x^2 and σ_y^2 are the population variances at the two time periods, σ_{xy} is the element-level covariance, and n_{12} is the number of units in s_{12}. Writing $n_{12} = \gamma n_1$ and $r = n_1/n_2$, the variance becomes $Var\left(\hat{\bar{d}}\right) = \frac{1}{n_1}\left[\sigma_x^2 + r\sigma_y^2 - 2\gamma r\sigma_{xy}\right]$. For a one-sided test of $H_0 : \mu_D = 0$ versus $H_A : \mu_D = \delta$ to be done with power β, the required sample size n_1 is

$$n_1 = \frac{1}{\delta^2}\left[\sigma_x^2 + r\sigma_y^2 - 2\gamma r\rho\sigma_x\sigma_y\right]\left(z_{1-\alpha} - z_\beta\right)^2\ . \tag{4}$$

The value of n_2 is then determined from $r = n_1/n_2$. The function `nDep2sam` in **PracTools** will perform

this calculation. `nProp2sam` will do a similar calculation for testing the difference in proportions with overlapping samples.

Probability proportional to size sampling

Probability proportional to size (*pps*) sampling can be very efficient if the measure of size (*mos*) used for sampling is correlated with the quantities measured in a survey. For example, enrollment in an elementary school may be related to the number of children receiving a government assistance program. In an establishment survey, the total number of employees is often correlated with other employment counts in the establishment, like the number who participate in a retirement plan. In a hospital survey, the number of inpatient beds is usually related to numbers of patients discharged in a month's time. Household samples are often selected using several stages, the first of which is a sample of geographic areas. An effective *mos* is typically the number of persons or housing units in each geographic area.

If a variable follows a certain regression structure in the population, then an optimal measure of size can be estimated. The key finding is due to Godambe and Joshi (1965). Isaki and Fuller (1982) extended this to a linear model M where $E_M(y_i) = \mathbf{x}_i^T\beta$ and $Var_M(y_i) = v_i$ with $\mathbf{x}_i$ defined as a vector of x's (auxiliary variables), and β is a vector of regression slopes of the same dimension as $\mathbf{x}_i$. Assume that a population total is estimated and a regression estimator is used that is approximately unbiased when averaging over the model and a probability sampling design. In that case, $\sqrt{v_i}$ is the best *mos* for *pps* sampling.

A model that may fit some establishment or institutional populations reasonably well has a variance with the form, $Var_M(y_i) = \sigma^2 x_i^\gamma$, where x_i is a *mos* and γ is a power. Typical values of γ are in the interval $[0, 2]$. The function `gammaFit` in **PracTools** returns an estimate of γ using an iterative algorithm. The algorithm is based on initially running an ordinary least squares (OLS) regression of y_i on $\mathbf{x}_i$. The OLS residuals, e_i, are then used to regress $\log(e_i^2)$ on $\log(\mathbf{x}_i)$ with an intercept. The slope in this regression is an estimate of γ. This procedure iterates by using the latest estimate of γ in a weighted least squares regression of y_i on $\mathbf{x}_i$. The parameter γ is then re-estimated in the logarithmic regression. The algorithm proceeds until some user-controllable convergence criteria are met.

The variance formulas for *pps* without replacement sampling are difficult or impossible to use for sample size determination because they involve joint selection probabilities of units and the sample size is not readily accessible. One practical approach is to use a variance formula appropriate for *pps* with replacement (*ppswr*) sampling. The simplest estimator of the mean that is usually studied with *ppswr* sampling is called "p-expanded with replacement" (*pwr*) (Särndal et al., 1992, chap. 2) and is defined as

$$\hat{\bar{y}}_{pwr} = \frac{1}{Nn}\sum_s \frac{y_i}{p_i}\,, \tag{5}$$

where p_i is the probability that element i would be selected in a sample of size 1. A unit is included in the sum as many times as it is sampled. The variance of $\hat{\bar{y}}_{pwr}$ in *ppswr* sampling is

$$Var\left(\hat{\bar{y}}_{pwr}\right) = \frac{1}{N^2 n}\sum_U p_i\left(\frac{y_i}{p_i} - t_U\right)^2 \equiv \frac{V_1}{N^2 n}, \tag{6}$$

where t_U is the population total of y. If the desired coefficient of variation is CV_0, Equation (6) can be solved to give the sample size as

$$n = \frac{V_1}{N^2}\frac{1}{\bar{y}_U^2 CV_0^2}\,. \tag{7}$$

We later give an example of how (7) may be evaluated using **PracTools**.

Stratified sampling

Stratified sampling is a useful way of restricting the dispersion of a sample across groups in a population. It can also lead to improvements in precision of overall estimates if an efficient allocation to the strata is used. For example, establishments can be stratified by type of business (retail, wholesale, manufacturing, etc.). Other methods of creating strata are provided by the R package **stratification** (Baillargeon and Rivest, 2014). Given that strata have been created, there are various ways of efficiently allocating a sample to strata: (i) minimize the variance of an estimator given a fixed total sample size (Neyman allocation), (ii) minimize the variance of an estimator for a fixed total budget, (iii) minimize the total cost for a target CV or variance of an estimator, or (iv) allocate the sample subject to several CV or cost criteria subject to a set of constraints on stratum sample sizes or other desiderata. The last is referred to as *multicriteria optimization*.

The standard texts noted earlier give closed form solutions for (i) and (ii). For example, suppose a mean is estimated, an *srswor* is to be selected within each stratum ($h = 1, ..., H$), and that the total cost can be written as $C = C_0 + \sum_{h=1}^{H} c_h n_h$ where C_0 denotes fixed costs that do not vary with the sample size, c_h is the cost per-element in stratum h, and n_h is the number of elements sampled from stratum h. The allocation to strata that minimizes the variance of the estimated mean subject to a fixed total budget C is

$$n_h = (C - C_0) \frac{W_h S_h / \sqrt{c_h}}{\sum_{h=1}^{H} \left(W_h S_h \sqrt{c_h}\right)}, \tag{8}$$

where W_h is the proportion in the population in stratum h and S_h is the population standard deviation in stratum h of the variable whose mean is estimated. This and the allocations for (i) and (iii) above can be found using `strAlloc` in **PracTools**.

Multicriteria optimization and allocations with constraints are more realistic for multipurpose surveys. In some cases, solutions to particular allocation problems are available as in Gabler et al. (2012). More generally, the **alabama** (Varadhan, 2015) and **Rsolnp** (Ghalanos and Theussl, 2014) packages will solve nonlinear optimization problems with constraints and can be very useful for complicated sample allocations. Among the constraints that are used in practical work are ones on minimum and maximum stratum sample sizes and relvariances of overall and individual stratum estimates. Theussl and Borchers (2015) present a CRAN Task View on optimization and mathematical programming in R. An R package that will form strata for multipurpose samples is **SamplingStrata** (Barcaroli, 2014). Also, the `Solver` add-on to Microsoft *Excel*® (Fylstra et al., 1998) will handle allocation problems that are quite complex and is easy to use. Use of these tools for sample allocation is covered in some detail in Valliant et al. (2013, chap. 5).

Two- and three-stage sampling

Two- and three-stage sampling is commonplace in household surveys but can be also used in other situations. For example, the U.S. National Compensation Survey selects a three-stage sample—geographic areas, establishments, and occupations—to collect compensation data (Bureau of Labor Statistics, 2013, chap. 8). Allocating the sample efficiently requires estimates of the contribution to the variance of an estimate by each stage of sampling.

As an example, consider a two-stage design in which the primary sampling units (PSUs) are selected with varying probabilities and with replacement and elements are selected at the second-stage by *srswor*. As noted earlier, determining sample sizes as if the PSUs are selected with-replacement is a standard workaround in applied sampling to deal with the fact that without-replacement variance formulas for *pps* samples are too complex to use for finding allocations. (Other selection methods along with three-stage designs are reviewed in Valliant et al. (2013, chap.9).) Let m be the number of sample PSUs, N_i be the number of elements in the population for PSU i, and suppose that the same number of elements, $\bar{n}$, is selected from each PSU. The *pwr*-estimator of a total is

$$\hat{t}_{pwr} = \frac{1}{m} \sum_{i \in s} \frac{\hat{t}_i}{p_i},$$

where $\hat{t}_i = \frac{N_i}{\bar{n}} \sum_{k \in s_i} y_{ik}$ is the estimated total for PSU i from a simple random sample and p_i is the 1-draw selection probability of PSU i, i.e., the probability in a sample of size one. The variance of $\hat{t}_{pwr}$ is

$$V\left(\hat{t}_{pwr}\right) = \frac{S^2_{U1(pwr)}}{m} + \frac{1}{m\bar{n}} \sum_{i \in U} \left(1 - \frac{\bar{n}}{N_i}\right) \frac{N_i^2 S^2_{U2i}}{p_i},$$

where U is the universe of PSUs, $S^2_{U1(pwr)} = \sum_{i \in U} p_i \left(\frac{t_i}{p_i} - t_U\right)^2$, t_i is the total of y for PSU i, and S^2_{U2i} is the population variance of y within PSU i. Dividing this by t_U^2 and assuming that the within-PSU sampling fraction, $\bar{n}/N_i$, is negligible, we obtain the relative variance (relvariance) of $\hat{t}_{pwr}$ as, approximately,

$$\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} \doteq \frac{B^2}{m} + \frac{W^2}{m\bar{n}} = \frac{\tilde{V}}{m\bar{n}} k \left[1 + \delta\left(\bar{n} - 1\right)\right], \tag{9}$$

with $\tilde{V} = S^2_U / \bar{y}^2_U$, $\bar{y}_U$ is the population mean per element, $k = (B^2 + W^2)/\tilde{V}$, $B^2 = S^2_{U1(pwr)}/t^2_U$, $W^2 = t_U^{-2} \sum_{i \in U} N_i^2 S^2_{U2i}/p_i$, and $\delta = B^2/(B^2 + W^2)$.

A simple cost function for two-stage sampling assumes that there is a cost per sample PSU and a cost per sample element of collecting and processing data. We model the total cost as

$$C = C_0 + C_1 m + C_2 m\bar{n},$$

where

C_0 = costs that do not depend on the number of sample PSUs and elements;

C_1 = cost per sample PSU; and

C_2 = cost per element within PSU.

The optimal number of units to select per PSU, i.e., the number that minimizes the approximate relvariance, is

$$\bar{n}_{opt} = \sqrt{\frac{C_1}{C_2}\frac{1-\delta}{\delta}}. \tag{10}$$

Only the ratio of the unit costs needs to be known in order to compute $\bar{n}_{opt}$. To find the optimal m for a fixed total cost, we substitute $\bar{n}_{opt}$ into the cost function to obtain

$$m_{opt} = \frac{C - C_0}{C_1 + C_2\bar{n}_{opt}}. \tag{11}$$

Alternatively, to find the optimal m for a fixed relvariance, CV_0^2, $\bar{n}_{opt}$ is substituted into the relvariance formula (9). `clusOpt2` in **PracTools** will do either of these calculations.

For three-stage sampling, suppose that m PSUs are selected with varying probabilities and with-replacement, $\bar{n}$ secondary sampling units (SSUs) are selected within each PSU by *srswor*, and $\bar{\bar{q}}$ elements are sampled by *srswor* within each sample SSU. This design is referred to as *ppswr/srs/srs* below. The relvariance of the *pwr*-estimator of a total in such a three-stage sample (with a negligible sampling fraction in the second and third stages) can be written as, e.g., see Hansen et al. (1953) and Valliant et al. (2013, chap.9):

$$\begin{aligned}\frac{V\left(\hat{t}_{pwr}\right)}{t_U^2} &\doteq \frac{B^2}{m} + \frac{W_2^2}{m\bar{n}} + \frac{W_3^2}{m\bar{n}\bar{\bar{q}}} \\ &= \frac{\tilde{V}}{m\bar{n}\bar{\bar{q}}}\left\{k_1\delta_1\bar{n}\bar{\bar{q}} + k_2\left[1+\delta_2\left(\bar{\bar{q}}-1\right)\right]\right\},\end{aligned} \tag{12}$$

where $B^2 = M^2S_{U1}^2/t_U^2$, $W_2^2 = M\sum_{i\in U}N_i^2S_{U2i}^2/t_U^2$, and $W_3^2 = M\sum_{i\in U}N_i\sum_{j\in U_i}Q_{ij}^2S_{U3ij}^2/t_U^2$. The variance components and other terms in Equation (12) are defined as:

$\tilde{V} = \frac{1}{Q-1}\sum_{i\in U}\sum_{j\in U_i}\sum_{k\in U_{ij}}\left(y_k-\bar{y}_U\right)^2\Big/\bar{y}_U^2$ is the unit relvariance of y in the population with Q being the total number of elements;

$S_{U1}^2 = \frac{\sum_{i\in U}\left(t_i-\bar{t}_U\right)^2}{M-1}$, the variance among the M PSU totals;

$S_{U2i}^2 = \frac{1}{N_i-1}\sum_{j\in U_i}\left(t_{ij}-\bar{t}_{Ui}\right)^2$ is the unit variance of the N_i SSU totals in PSU i with $t_{ij} = \sum_{k\in U_{ij}}y_k$ being the population total for PSU/SSU ij, $\bar{t}_{Ui} = \sum_{j\in U_i}t_{ij}\Big/N_i$ is the average total per SSU in PSU i;

$S_{U3ij}^2 = \frac{1}{Q_{ij}-1}\sum_{k\in U_{ij}}\left(y_k-\bar{y}_{Uij}\right)^2$ is the unit variance among the Q_{ij} elements in PSU/SSU ij with $\bar{y}_{Uij} = \sum_{k\in U_{ij}}y_k\Big/Q_{ij}$.

$k_1 = \left(B^2+W^2\right)/\tilde{V}$;

$W^2 = \frac{1}{t_U^2}\sum_{i\in U}Q_i^2S_{U3i}^2/p_i$ with Q_i being the number of elements in PSU i,

$S_{U3i}^2 = \frac{1}{Q_i-1}\sum_{j\in U_i}\sum_{k\in U_{ij}}\left(y_k-\bar{y}_{Ui}\right)^2$ and $\bar{y}_{Ui} = \sum_{j\in U_i}\sum_{k\in U_{ij}}y_k/Q_i$; i.e., S_{U3i}^2 is the element-level variance among all elements in PSU i; and

$k_2 = \left(W_2^2+W_3^2\right)/\tilde{V}$;

$\delta_1 = B^2/\left(B^2+W^2\right)$;

$\delta_2 = W_2^2/\left(W_2^2+W_3^2\right)$.

The terms δ_1 and δ_2 are referred to as *measures of homogeneity*, as is δ for two-stage sampling. Equation (12) is useful for sample allocation because the measures of homogeneity are in $[0,1]$, k_1 and k_2 are usually near 1, and $\tilde{V}$ can usually be estimated.

To arrive at an optimal allocation, costs need to be considered. A cost function for three-stage sampling, analogous to the one for two-stage sampling, is

$$C = C_0 + C_1m + C_2m\bar{n} + C_3m\bar{n}\bar{\bar{q}}. \tag{13}$$

The term C_0 is again costs that do not depend on the sample sizes at different stages; C_1 is the cost per PSU; C_2 is the cost per SSU; and C_3 is the cost per element within each SSU. Minimizing the *ppswr/srs/srs* relvariance in Equation (12) subject to a fixed total cost gives the following optima:

$$\bar{\bar{q}}_{opt} = \sqrt{\frac{1-\delta_2}{\delta_2}\frac{C_2}{C_3}}, \tag{14}$$

$$\bar{n}_{opt} = \frac{1}{\bar{\bar{q}}}\sqrt{\frac{1-\delta_2}{\delta_1}\frac{C_1}{C_3}\frac{k_2}{k_1}}, \tag{15}$$

$$m_{opt} = \frac{C-C_0}{C_1 + C_2\bar{n} + C_3\bar{n}\bar{\bar{q}}}. \tag{16}$$

If a target relvariance is set at CV_0^2, then the equations for finding the optima $\bar{\bar{q}}_{opt}$ and $\bar{n}_{opt}$ are the same. The optimum number of PSUs is found by substituting $\bar{n}_{opt}$ and $\bar{\bar{q}}_{opt}$ into the relvariance in Equation (12). `clusOpt3` will do these computations for three-stage samples.

Two-phase sampling

In finite population sampling, a distinction is drawn between multi*stage* sampling and multi*phase* sampling. In a multiphase sample, an initial sample is selected, some characteristics of the units are observed, and a decision is made about how to select a subsample from the initial sample based on what has been observed. In multistage sampling, the same design is used in later stages regardless of what was found in the first-stage units. There is a more technical definition of the difference between multistage and multiphase, but it is unimportant for this discussion. An example of two-phase sampling is to select a subsample of nonrespondents to the initial phase to attempt to get them to cooperate. This is known as a *nonresponse follow-up study* (*NRFU*).

Another type of two-phase design is *double sampling for stratification*. In this design, information is collected in the first phase which is then used to stratify elements for second phase sampling. For example, researchers working to develop a case definition for undiagnosed medical symptoms in U.S. personnel serving in the 1991 Persian Gulf War surveyed a stratified simple random sample of Gulf War-era veterans (Iannacchione et al., 2011). Based on survey responses in the first phase, respondents were classified as likely having or not having a certain type of illness. Blood specimens were requested from randomly sampled phase-1 respondents within the illness strata and analyzed using expensive tests.

As an illustration, take the case of double sampling for stratification. Cochran (1977) and Neyman (1938) give the two-phase variance of an estimated mean or proportion when phase-1 is a simple random sample of $n_{(1)}$ elements, phase-2 is a stratified simple random sample (*stsrs*) of $n_{(2)}$ elements, and an optimal allocation to strata is used in the second phase. The sampling fractions at both stages are assumed to be negligible. The optimal proportion of the phase-2 sample to assign to stratum h for estimating the population mean is $n_{(2)h}/n_{(2)} = W_h S_h / \sum_h W_h S_h$. The formula for the variance of an estimated stratified mean with this allocation is

$$V_{opt} = \frac{\sum_h W_h\,(\bar{y}_{Uh} - \bar{y}_U)^2}{n_{(1)}} + \frac{(\sum_h W_h S_h)^2}{n_{(2)}} \equiv \frac{V_{(1)}}{n_{(1)}} + \frac{V_{(2)}}{n_{(2)}}, \tag{17}$$

where $\bar{y}_{Uh}$ is the stratum h population mean. The phase-2 subsampling rate from the phase-1 sample units that minimizes Equation (17) is

$$\frac{n_{(2)}}{n_{(1)}} = \sqrt{\frac{V_{(2)}}{V_{(1)}} \Big/ \frac{c_{(2)}}{c_{(1)}}},$$

where $c_{(1)}$ and $c_{(2)}$ are the per-unit costs in the first and second-phases, respectively. The formulas for the phase-1 and phase-2 sample sizes that minimize V_{opt} subject to a fixed total cost C are

$$n_{(1)} = \frac{C}{c_{(1)} + c_{(2)}\sqrt{K}}, \qquad\qquad n_{(2)} = n_{(1)}\sqrt{K},$$

where

$$K = \left(V_{(2)}/V_{(1)}\right) \Big/ \left(c_{(2)}/c_{(1)}\right).$$

The function `dub` in **PracTools** will calculate the optimal second-phase sampling fraction and the phase-1 and phase-2 sample sizes that will minimize the variance in Equation (17). The function also computes the size of an *srs* that would cost the same as the two-phase sample and the ratio of the

two-phase stratified variance to the *srs* variance.

Extension to nonlinear estimators and limitations

PracTools will calculate sample sizes for estimators whose variance can be written in one of the forms given in Section Designing survey samples. This directly covers linear estimators of means and totals. A nonlinear estimator that is a differentiable function of a vector of estimated totals is also covered. But, a user must do some work to linearize the estimator and determine the inputs that are required for a **PracTools** function. Suppose that the estimator is $\hat{\theta} = f\left(\hat{t}_1, ..., \hat{t}_p\right)$ where f is a differentiable function and $\hat{t}_j$ is a linear estimator of a population total, t_j $(1, 2, ..., p)$. The linear approximation to $\hat{\theta} - \theta$ is

$$\hat{\theta} - \theta \doteq \sum_{j=1}^{p} \frac{\partial f}{\partial t_j}\left(\hat{t}_j - t_j\right), \tag{18}$$

where $\theta = f\left(t_1, ..., t_p\right)$, and the partials are evaluated at the population values. The repeated sampling variance of this approximation is then the same as the variance of $\sum_{j=1}^{p} \frac{\partial f}{\partial t_j}\hat{t}_j$ since θ and t_j are treated as constants. Taking the case of two-stage sampling, suppose that the estimator of the total for variable j is $\hat{t}_j = \sum_{i\in s}\sum_{k\in s_i} w_{ik}y_{ik}(j)$ where w_{ik} is a weight for element k in PSU i, $y_{ik}(j)$ is its data value, s is the set of sample PSUs, and s_i is the sample of elements within SSU i. Substituting this into (18) and reversing the order of summation between PSUs and variables leads to the expression

$$\hat{z} = \sum_{i\in s} \underbrace{\sum_{j=1}^{p} \frac{\partial f}{\partial t_j}\hat{t}_i\,(j)}_{\hat{z}_i}. \tag{19}$$

where $\hat{t}_i(j) = \sum_{k\in s_i} w_{ik}y_{ik}(j)$. One then computes the design-variance of $\hat{z}$ based on the particular sample design used and the form of the derivatives. This general approach is known as the *linear substitute* method and is described in detail in Wolter (2007).

In a two-stage sample where m PSUs are selected with replacement and with varying probabilities, and $\bar{n}$ elements are selected by simple random sampling from each PSU, (9) applies. The weight is defined as $w_{ik} = (mp_i)^{-1}(N_i/\bar{n})$ where p_i is the 1-draw selection probability, as before. If the estimator is the mean computed as $\hat{\theta} = \hat{t}_y/\hat{M}$ with $\hat{M} = \sum_{i\in s}\sum_{k\in s_i} w_{ik}$, then $\hat{z}_i = (Mmp_i)^{-1}N_ie_i$ where $e_i = \bar{y}_i - \bar{y}_U$ with y_i being the sample mean of y in PSU i and $\bar{y}_U$ the population mean. Expression (19) becomes

$$\hat{z} = \frac{1}{Mm}\sum_{i\in s}\frac{\hat{t}_{zi}}{p_i},$$

with $\hat{t}_{zi} = \left(N_i/\bar{n}\sum_{k\in s_i} e_{ik}\right)$ and $e_{ik} = y_{ik} - \bar{y}_U$. Expression (9) would then be evaluated with e_{ik} replacing y_{ik}. With the linear substitute method, residuals typically appear in the linear approximation and are the basis for a variance estimator. Population quantities, like M and $\bar{y}_U$, are replaced by sample estimates in a variance estimator. The Section Examples gives an illustration of this method using one of the datasets in **PracTools**.

Nonetheless, there are types of estimators that our package does not cover. Quantile estimators require a special approximation and variance formula (Francisco and Fuller, 1991) that does not come from the standard linearization approach. The Gini coefficient, used as a measure of income inequality (e.g., Deaton, 1997), is another example of an estimator that is too complicated to linearize using the methods above.

The R package PracTools

PracTools is a collection of specialized functions written in R along with several example finite populations that can be used for teaching. The package is available from the Comprehensive R Archive Network (CRAN) at `http://CRAN.R-project.org/package=PracTools`. Because the function code is visible, the routines can be modified (and improved) by any user. A brief description of the functions is given in Table 1. The use of the functions, their input parameters, and the values they return are described in the help files. The package also contains nine example populations of different types. These are listed in Table 2.

Function	Purpose
`BW2stagePPS`, `BW2stageSRS`	Variance components in two-stage samples from a population frame
`BW2stagePPSe`	Estimated variance components in two-stage samples from a sample
`BW3stagePPS`	Variance components in three-stage samples from a population frame
`BW3stagePPSe`	Estimated variance components in three-stage samples from a sample
`clusOpt2`	Optimal allocation in a two-stage sample
`clusOpt2fixedPSU`	Optimal second-stage sample size in a two-stage sample when the PSU sample is fixed
`clusOpt3`	Optimal allocation in a three-stage sample
`clusOpt3fixedPSU`	Optimal second and third-stage sample sizes in a three-stage sample when the PSU sample is fixed
`CVcalc2, CVcalc3`	Compute the coefficient of variation of an estimated total in two- and three-stage designs
`deffH, deffK, deffS`	Henry, Kish, and Spencer design effects
`dub`	Allocation of double sample for stratification
`gammaFit`	Estimate variance power in a linear model
`nCont, nContMoe`	Sample size to meet CV, variance, or margin of error targets (continuous variable)
`nDep2sam`	Sample sizes for two-sample comparison of means with overlapping samples (continuous variable)
`nLogOdds`	Sample size calculation for a proportion using log-odds method
`nProp`	Sample size calculation for a proportion using target CV or variance
`nProp2sam`	Sample sizes for two-sample comparison of proportions with overlapping samples (continuous variable)
`nPropMOE`	Sample size calculation for a proportion using margin of error target
`NRFUopt`	Sample sizes for two-phase nonresponse follow-up study
`nWilson`	Sample size calculation for a proportion using Wilson method
`pclass`	Form nonresponse adjustment classes based on propensity scores
`strAlloc`	Sample allocation in stratified samples

Table 1: Functions in **PracTools**.

Population	Description
`HMT`	Generate population that follows the model in Hansen et al. (1983)
`hospital`	Population of 393 short-stay hospitals with fewer than 1000 beds
`labor`	Clustered population of 478 persons extracted from Sept. 1976 Current Population Survey
`MDarea.pop`	Artificial population of of 403,997 persons arrayed in census tracts and block groups
`nhis, nhispart`	Datasets of persons with demographic and socioeconomic variables
`nhis.large`	21,588 persons with 18 demographic and health-related variables
`smho.N874`	874 mental health organizations with 6 financial variables
`smho98`	875 mental health organizations with 8 financial and patient-count variables

Table 2: Finite populations in **PracTools**.

Examples

This section gives some examples for computing sample sizes for estimating proportions and differences of proportions, an allocation to strata, and the optimal numbers of PSUs, secondary units, and elements in a three-stage sample.

Proportions and differences in proportions

The function `nProp` will return the sample size required for estimating a proportion with a specified CV or variance. For a CV target, Equation (1) is used; the formula for a variance target is similar. The function takes the following parameters:

`CV0`	target value of coefficient of variation of the estimated proportion
`V0`	target value of variance of the estimated proportion
`pU`	population proportion
`N`	number of units in finite population; default is `Inf`

A single numeric value, the sample size, is returned. An advance guess is needed for the value of the population proportion, p_U. By default, the population is assumed to be very large ($N = \infty$), but specifying a finite value of N results in a finite population correction being used in calculating the sample size. To estimate a proportion anticipated to be $p_U = 0.1$ with a CV_0 of 0.05, the function call and resulting output is:

```
> nProp(CV0 = 0.05, N = Inf, pU = 0.1)
[1] 3600
```

If the population has only 500 elements, then the necessary sample size is much smaller:

```
> nProp(CV0 = 0.05, N = 500, pU = 0.1)
[1] 439.1315
```

In this function and others in the package, sample sizes are not rounded in case the exact value is of interest to a user. To obtain sample sizes for two overlapping groups, `nDep2sam` is the appropriate function, which uses Equation (4). The function takes these inputs:

`S2x`	unit variance of analysis variable x in sample 1
`S2y`	unit variance of analysis variable y in sample 2
`g`	proportion of sample 1 that is in the overlap with sample 2
`r`	ratio of the size of sample 1 to that of sample 2
`rho`	unit-level correlation between x and y
`alt`	should the test be 1-sided or 2-sided; allowed values are `"one.sided"` or `"two.sided"`
`del`	size of the difference between the means to be detected
`sig.level`	significance level of the hypothesis test
`pow`	desired power of the test

Among other things, the user must specify the unit (or population) standard deviations in the two populations from which the samples are selected, the proportion of the first sample that is in the second (i.e., a measure of overlap), and the unit-level correlation between the variables being measured in the two samples. If there is no overlap in the samples, it would be natural to set `rho = 0`. The size of the difference in the means that is to be detected and the power of the test on the difference in means must also be declared. The code below computes the sample size needed to detect a difference of 5 in the means with a power of 0.8 when the unit variances in both groups are 200, 75 percent of the first sample is in the second, the samples from the two groups are to be the same size, and the unit-level correlation is 0.9. This function and several others in the package use the class 'power.htest' as a convenient way of returning the output.

```
> nDep2sam(S2x = 200, S2y = 200, g = 0.75, r = 1, rho = 0.9,
+          alt = "one.sided", del = 5, sig.level = 0.05, pow = 0.80)

     Two-sample comparison of means
 Sample size calculation for overlapping samples

            n1 = 33
            n2 = 33
        S2x.S2y = 200, 200
          delta = 5
          gamma = 0.75
              r = 1
            rho = 0.9
            alt = one.sided
      sig.level = 0.05
          power = 0.8
```

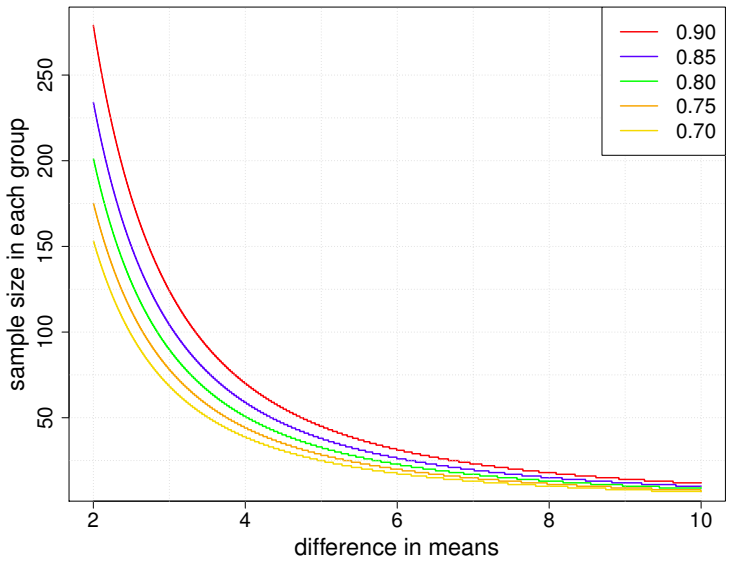

Figure 1: Sample size in each group for detecting a range of differences in means with several levels of power.

`nDep2sam` will also accept a vector of differences as input, e.g., `del <-seq(2,10,0.001)`. This makes it easy to generate a plot like in Figure 1, where the sample size in each group is plotted against the difference in means for several levels of power. This is a useful way to present options to users.

Probability proportional to size sampling

The function `gammaFit` will estimate the variance parameter in the model $E_M(y_i) = \mathbf{x}_i^T\beta$, $Var_M(y_i) = \sigma^2 x_i^\gamma$. The code below uses the `hospital` population bundled with **PracTools** to estimate the variance power in the model, $E_M(y) = \beta_1\sqrt{x} + \beta_2 x, V_M(y) = \sigma^2 x^\gamma$, returning $\hat{\gamma} = 1.612009$. Using the function `UPrandomsystematic` from the package **sampling**, a sample of size 30 is then selected with probability proportional to $\sqrt{x^{\hat{\gamma}}}$, which is the optimal measure of size for estimating the population total of y.

```
> data("hospital")
> x <- hospital$x
> y <- hospital$y
> X <- cbind(sqrt(x), x)
> (res <- gammaFit(X = X, x = x, y = y, maxiter = 100, tol = 0.001))
Convergence attained in  47 steps.
g.hat = 1.612009
$g.hat
       X
1.612009
> require(sampling)
> n <- 30
> pik <- n * sqrt(x^res$g.hat) / sum(sqrt(x^res$g.hat))
> sam <- UPrandomsystematic(pik)
> hosp.sam <- hospital[sam == 1, ]
```

To determine a sample size for *ppswr* sampling, the function `nCont` can be used. This function is designed to evaluate Equation (1) for simple random samples. However, Equation (7) has the same form if we equate $V_1/(N^2\bar{y}_U^2)$ in Equation (7) to $S^2/\bar{y}_U^2$ in Equation (1) and set $N = \infty$. If $V_1/(N^2\bar{y}_U^2) = 2$ and the CV target is 0.05, the call to `nCont` is

```
> nCont(CV0 = 0.05, N = Inf, CVpop = sqrt(2))
[1] 800
```

Allocations to strata

A sample can be allocated to strata using `strAlloc`. The function takes a number of parameters which are described in the help file. A standard problem in applied sampling is to find an allocation that will

minimize the variance of an estimated mean subject to a fixed total budget. To solve this problem, the stratum population sizes, standard deviations, stratum per-unit costs, total budget, and the type of allocation (`alloc = "totcost"`) are specified; partial output is:

```
> Nh <- c(215, 65, 252, 50, 149, 144)
> Sh <- c(267, 106, 69, 110, 98, 445)
> ch <- c(1400, 200, 300, 600, 450, 1000)
> strAlloc(Nh = Nh, Sh = Sh, cost = 100000, ch = ch, alloc = "totcost")

   allocation = totcost
           Nh = 215, 65, 252, 50, 149, 144
           Sh = 267, 106, 69, 110, 98, 445
           nh = 30.578027, 9.710196, 20.008418, 4.475183, 13.719233, 40.387433
         nh/n = 0.25722085, 0.08168169, 0.16830983, 0.03764502, 0.11540551, 0.33973710
```

The function returns a list with the type of allocation, population stratum counts and standard deviations, sample sizes for each stratum, the proportions of the sample allocated to each stratum, and the anticipated standard error of the mean. Other options for the allocation types are proportional to stratum population sizes (`"prop"`), Neyman (`"neyman"`), and minimization of the total cost subject to a specified variance or CV target (`"totvar"`). The `nh` component of the list can then be used in, e.g., the `strata` function in **sampling**, to select the sample. Either the `round` or `ceiling` function could be applied to `nh` to create integer sample sizes. (If non-integers are supplied to `strata`, they will be truncated to the integer floor.)

Allocations in two- and three-stage samples

When designing multistage samples, decisions must be made about how many units to select at each stage. To illustrate this, we consider a three-stage sample. A considerable amount of data is needed to estimate realistic ingredients required for `clusOpt3`. The function, `BW3stagePPSe`, will estimate B^2, W^2, W_2^2, W_3^2, δ_1, and δ_2 from a three-stage where the first-stage is selected *ppswr* and the last two stages are selected by *srswor*. `BW2stagePPSe` does similar calculations for two-stage sampling. Variance component estimation is, of course, a difficult area where a number of alternatives have been developed in the model-based literature. The forms used in `BW2stagePPSe` and `BW3stagePPSe` are fairly simple ANOVA-type estimates. These estimates have known defects, like occasionally being negative.

The following example computes a three-stage allocation that minimizes the variance of the *pwr*-estimator assuming that the budget for variable cost is 100,000; the PSU, SSU, and per-element costs are 500, 100, and 120, respectively; $\delta_1 = 0.01$, $\delta_2 = 0.10$; the unit relvariance is $\tilde{V} = 1$; and the ratios, k_1 and k_2, are both 1. `cal.sw = 1` specifies that the optima be found for a fixed total budget. The full description of the input parameters can be found in the help file for `clusOpt3`.

```
> clusOpt3(unit.cost = c(500, 100, 120), delta1 = 0.01, delta2 = 0.10, unit.rv = 1,
+          k1 = 1, k2 = 1, tot.cost = 100000, cal.sw = 1)

            C1 = 500
            C2 = 100
            C3 = 120
        delta1 = 0.01
        delta2 = 0.1
    unit relvar = 1
            k1 = 1
            k2 = 1
          cost = 1e+05
         m.opt = 28.3
         n.opt = 7.1
         q.opt = 2.7
            CV = 0.0499
```

Along with the inputs, the output includes the optimal sample size for each stage and the CV that is anticipated for the *pwr*-estimator given that design. The sample sizes above can be rounded and then used in the **sampling** package to select units at each of the stages. For example, suppose that 28 PSUs and 7 SSUs will be selected with probabilities proportional to a *mos*. Within each sample SSU, a sample of 3 elements will be selected via *srswor*. The PSUs can be selected using the function `cluster` and the data extracted. Then, a cluster sample of 7 SSUs can be selected from each of those 28 units in a loop, again using `cluster` and the data for those sample SSUs extracted. The sample of SSUs would then be treated as strata and the `strata` function used to select 3 elements from each SSU using *srswor*.

Double sampling for stratification

The function dub will compute the allocation to strata for a double sampling design in which phase-1 is used to assign units to strata. The function takes these input parameters:

c1	cost per unit in phase-1
c2	cost per unit in phase-2
Ctot	total variable cost
Nh	vector of stratum population counts or proportions
Sh	vector of stratum population standard deviations
Yh.bar	vector of stratum population means

The inputs, N_h, S_h, and $\bar{Y}_h$, will typically have to be estimated from the first-phase sample. The example below computes the allocation to four strata assuming a total cost of 20,000 and unit costs of $c_{(1)} = 10$ and $c_{(2)} = 50$. A proportion is being estimated.

```
> Wh <- rep(0.25, 4)
> Ph <- c(0.02, 0.12, 0.37, 0.54)
> Sh <- sqrt(Ph * (1 - Ph))
> c1 <- 10; c2 <- 50; Ctot <- 20000
> dub(c1, c2, Ctot, Nh = Wh, Sh, Yh.bar = Ph)

            V1 = 0.04191875
            V2 = 0.1307118
            n1 = 404.1584
            n2 = 319.1683
         n2/n1 = 0.789711
     ney.alloc = 30.89801, 71.71903, 106.55494, 109.99634
          Vopt = 0.0005132573
          nsrs = 400
          Vsrs = 0.0004839844
        Vratio = 1.06
```

The function also computes the size of an *srs*, nsrs, that would cost the same total amount, assuming that the per-unit cost is $c_{(2)}$; the anticipated variances with the optimal two-phase allocation and the *srs* of size nsrs; and the ratio of the two variances. Often, the two-phase design has very little gain, and sometimes a loss as in this example, compared to simple random sampling. However, double sampling for stratification is usually undertaken to control the sample sizes in the strata whose members are not known in advance.

Sample size for a nonlinear estimator

To illustrate a calculation for a nonlinear estimator, consider the proportion of Hispanics with insurance coverage in the MDarea.pop, which is part of the package. Define y_{2k} to be 1 if a person is Hispanic and 0 if not; $\alpha_{1k} = 1$ if a person has insurance coverage. Then, $y_{1k} = \alpha_{1k} y_{2k}$ is 1 if person k has insurance and is Hispanic and is zero otherwise. The linear substitute is $z_k = y_{1k} - \theta y_{2k}$ where θ is the proportion of Hispanics with insurance coverage. In this case, z_k can take only three values: $-\theta$, 0, and $1 - \theta$. If a simple random sample of clusters and persons within clusters is selected, BW2stageSRS can be used to compute B^2, W^2, and δ using the linear substitutes as inputs. Assuming that the full population is available, the R code is the following. We do the calculation for clusters defined as tracts (a small geographic area with about 4000 persons defined for census-taking).

```
> # recode Hispanic to be 1 = Hispanic, 0 if not
> y2 <- abs(MDarea.pop$Hispanic - 2)
> y1 <- y2 * MDarea.pop$ins.cov
> #  proportion of Hispanics with insurance
> p <- sum(y1) / sum(y2)
> # linear sub
> z <- y1 - p * y2
> BW2stageSRS(z, psuID = MDarea.pop$TRACT)
```

The result is $\delta = 0.00088$. Thus, the effect of clustering on this estimated proportion is inconsequential—a two-stage sample will estimate the proportion almost as precisely as an *srs* would. In contrast, if the estimate is the total number of Hispanics with insurance, then we call BW2stageSRS this way:

```
> BW2stageSRS(y1, psuID = MDarea.pop$TRACT)
```

which returns $\delta = 0.02251$.

Summary

Finite population sampling is one of the more important areas in statistics since many key economic and social measures are derived from surveys. R through its packages is gradually accumulating capabilities for selecting and analyzing samples from finite populations. Pieces that have been missing are sample size computations for the kinds of complex designs that are used in practice. **PracTools** contributes to filling that gap by providing a suite of sample size calculation routines for one-, two-, and three-stage samples. We also include features for stratified allocations, for probability proportional to size sampling, and for incorporating costs into the computations. Several realistic example populations, that should be useful for classroom instruction, are also part of the package.

The package is limited in the sense that it covers only some of the sample selection schemes that we have found are most useful and prevalent in the practice of survey sampling. There are many other selection algorithms that have their own, specialized variance formulas. Tillé (2006) covers many of these. **PracTools** also does not select samples, but there are a number of other R packages, mentioned in this paper that do. One of the great advantages of R is that users can readily access different packages for specialized tasks like sample size calculation and sample selection.

Bibliography

A. Alfons, M. Templ, and P. Filzmoser. An object-oriented framework for statistical simulation: The R package simFrame. *Journal of Statistical Software*, 37(3):1–36, 2010. URL http://www.jstatsoft.org/v37/i03/. [p163]

S. Baillargeon and L.-P. Rivest. *stratification: Univariate Stratification of Survey Populations*, 2014. URL https://CRAN.R-project.org/package=stratification. R package version 2.2-5. [p165]

G. Barcaroli. SamplingStrata: An R package for the optimization of stratified sampling. *Journal of Statistical Software*, 61(4):1–24, 2014. URL http://www.jstatsoft.org/v61/i04/. [p166]

Bureau of Labor Statistics. BLS handbook of methods, 2013. URL http://www.bls.gov/opub/hom/. [p163, 166]

Center for Disease Control and Prevention. National health and nutrition examination survey, 2013a. URL http://www.cdc.gov/nchs/nhanes.htm. [p163]

Center for Disease Control and Prevention. National health interview survey, 2013b. URL http://www.cdc.gov/nchs/nhis.htm. [p163]

W. Cochran. *Sampling Techniques*. John Wiley & Sons, Inc., New York, 1977. [p163, 168]

A. Deaton. *Analysis of Household Surveys*. Johns Hopkins University Press, Baltimore, 1997. [p169]

C. Francisco and W. A. Fuller. Quantile estimation with a complex survey design. *The Annals of Statistics*, 19:454–469, 1991. [p169]

D. Fylstra, L. Lasdon, J. Watson, and A. Waren. Design and use of the Microsoft Excel solver. *INFORMS Interfaces*, 28:29–55, 1998. [p166]

S. Gabler, M. Ganninger, and R. Münnich. Optimal allocation of the sample size to strata under box constraints. *Metrika*, 75:151–161, 2012. [p166]

J. G. Gambino. *pps: Functions for PPS Sampling*, 2012. URL https://CRAN.R-project.org/package=pps. R package version 0.94. [p163]

A. Ghalanos and S. Theussl. *Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method*, 2014. URL https://CRAN.R-project.org/package=Rsolnp. R package version 1.15. [p166]

V. P. Godambe and V. M. Joshi. Admissibility and Bayes estimation in sampling finite populations – I. *The Annals of Mathematical Statistics*, 36:1707–1723, 1965. [p165]

M. H. Hansen, W. H. Hurwitz, and W. G. Madow. *Sample Survey Methods and Theory*, volume I. John Wiley & Sons, Inc., New York, 1953. [p167]

M. H. Hansen, W. G. Madow, and B. J. Tepping. An evaluation of model-dependent and probability sampling inferences in sample surveys. *Journal of the American Statistical Association*, 78:776–793, 1983. [p170]

V. G. Iannacchione, J. A. Dever, C. M. Bann, K. A. Considine, D. Creel, C. P. Carson, H. L. Best, and R. W. Haley. Validation of a research case definition of Gulf War illness in the 1991 U.S. military population. *Neuroepidemiology*, 37(2):129–140, 2011. [p168]

C. T. Isaki and W. A. Fuller. Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96, 1982. [p165]

S. L. Lohr. *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove CA, 1999. [p163]

T. Lumley. *Complex Surveys*. John Wiley & Sons, Inc., New York, 2010. [p163]

T. Lumley. *survey: Analysis of Complex Survey Samples*, 2014. URL https://CRAN.R-project.org/package=survey. R package version 3.30-3. [p163]

J. Manitz. *samplingbook: Survey Sampling Procedures*, 2013. URL https://CRAN.R-project.org/package=samplingbook. R package version 1.2.0, with contributions by M. Hempelmann, G. Kauermann, H. Kuechenhoff, S. Shao, C. Oberhauser, N. Westerheide, M. Wiesenfarth. [p163]

J. Neyman. Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33(201):101–116, 1938. [p168]

C. Särndal, B. Swensson, and J. Wretman. *Model Assisted Survey Sampling*. Springer-Verlag, New York, 1992. [p163, 165]

S. Theussl and H. Borchers. CRAN task view: Optimization and mathematical programming, 2015. URL https://CRAN.R-project.org/view=Optimization. [p166]

Y. Tillé. *Sampling Algorithms*. Springer-Verlag, New York, 2006. [p175]

Y. Tillé and A. Matei. *sampling: Survey Sampling*, 2013. URL https://CRAN.R-project.org/package=sampling. R package version 2.6. [p163]

R. Valliant, J. Dever, and F. Kreuter. *Practical Tools for Designing and Weighting Survey Samples*. Springer-Verlag, New York, 2013. [p163, 166, 167]

R. Valliant, J. Dever, and F. Kreuter. *PracTools: Tools for Designing and Weighting Survey Samples*, 2015. URL https://CRAN.R-project.org/package=PracTools. R package version 0.2. [p163]

R. Varadhan. *alabama: Constrained Nonlinear Optimization*, 2015. URL https://CRAN.R-project.org/package=alabama. R package version 2015.3-1, with contributions from G. Grothendieck. [p166]

E. B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22:209–212, 1927. [p164]

K. M. Wolter. *Introduction to Variance Estimation*. Springer-Verlag, New York, 2nd edition, 2007. [p169]

M. Woodward. Formulas for sample size, power, and minimum detectable relative risk in medical studies. *The Statistician*, 41:185–196, 1992. [p164]

Richard Valliant
Joint Program in Survey Methodology
Universities of Michigan and Maryland
1218 Lefrak Hall, College Park MD 20742 USA
rvallian@umd.edu

Jill A. Dever
RTI International
701 13th Street NW, Suite 750, Washington, DC 20005-3967 USA
jdever@rti.org

Frauke Kreuter
Joint Program in Survey Methodology, University of Maryland
1218 Lefrak Hall, College Park MD 20742 USA
fkreuter@umd.edu