

# Analyzing Stratified Samples in R: Selected Example

Wayne Enanoria, PhD, MPH

26 November 2006

## 1 Introduction

The purpose of this article is to illustrate how to analyze stratified random samples using R.

## 2 Stratified Sampling

### 2.1 Example 1

#### 2.1.1 Data

The following data was given by Sharon Lohr (Reference: Lohr SL. Sampling: Design and Analysis. Duxbury Press:London, 1999). Here is the description of the data as it appears in the text.

'Sniff and Skoog (1964) used stratified random sampling to estimate the size of the Nelchina herd of Alaskan caribou in February 1962. ...The biologists used preliminary estimates of caribou densities to divide the area of interest into six strata; each stratum was then divided into a grid of 4-square-mile sampling units. Stratum A, for example, contained  $N_1 = 400$  sampling units;  $n_1 = 98$  of these were randomly selected to be in the survey.'

#### 2.1.2 Notation

Let's first start out by giving some notation. See Table 1 on Page 2.

So we can set up our stratified sample in a table or spreadsheet in order to illustrate how the calculations work. Refer to Table 2 on Page 7.

#### 2.1.3 Implementation in R

Now we will illustrate how to do these calculations in R.

```
> strat.id <- c("A", "B", "C", "D", "E", "F")
> strat.pop <- c(400, 30, 61, 18, 70, 120)
> strat.smpl <- c(98, 10, 37, 6, 39, 21)
> strat.sum.y <- c(2362, 256, 9901, 1074, 11454, 697)
> strat.var.y <- c(5575, 4064, 347556, 22798, 123578, 9795)
```

Table 1: Notation

Notation	Definition
$N_h$	= Number of sampling units in the $h$ th stratum.
$n_h$	= Number of observations in the $h$ th stratum.
$y_h = \sum_j y_{hj}$	= Sum of characteristic $y$ in $h$ th stratum.
$\bar{y}_h = \frac{\sum_j y_{hj}}{n_h}$	= Mean of characteristic $y$ in $h$ th stratum.
$s_h^2 = \sum_j \frac{(y_{hj} - \bar{y}_h)^2}{n_h - 1}$	= Sample variance in stratum $h$
$w_h = \frac{1}{n_h/N_h}$	= Weight for the $i$ th stratum
$\hat{t}_h = w_h \sum_j y_{hj}$	= Estimated population sum of characteristic

This is the only data we need. We can calculate the rest of the columns from the three vectors of numbers for each of the six strata.

First, we want to calculate the sample mean of  $y$  for each stratum.

```
> ybar <- strat.sum.y/strat.smpl
> ybar
[1] 24.10204 25.60000 267.59459 179.00000 293.69231 33.19048
```

Next, we can calculate the weight for each stratum. The weight is given by:

$$w_h = 1/(n_h/N_h). \quad (1)$$

```
> wt.strat <- 1/(strat.smpl/strat.pop)
> wt.strat
[1] 4.081633 3.000000 1.648649 3.000000 1.794872 5.714286
```

We can then estimate the population sum of the characteristic by taking a weighting the stratum-specific sums.

```
> estim.strat.tot <- wt.strat * strat.sum.y
> estim.strat.tot
[1] 9640.816 768.000 16323.270 3222.000 20558.462 3982.857
```

In order to calculate the overall estimated total for the population, we can sum the `estim.strat.tot` vector.

```
> estim.pop.tot <- sum(estim.strat.tot)
> estim.pop.tot
```

```
[1] 54495.4
```

The variance for the estimated population total is given by:

$$\hat{V}(\hat{t}_{str}) = \sum_h \left(1 - \frac{n_h}{N_h}\right) N_h^2 \left(\frac{s_h^2}{n_h}\right) \quad (2)$$

We can calculate the variance of the estimated population total with the following commands:

```
> var.tot.vec <- (1 - (strat.smpl/strat.pop)) * strat.pop^2 * (strat.var.y/strat.smpl)
> var.tot <- sum(var.tot.vec)
> var.tot
```

```
[1] 34105732
```

with standard error equal to

```
> se.tot <- sqrt(var.tot)
> se.tot
```

```
[1] 5840.011
```

This result can then be used to calculate the 95% Confidence Intervals for the estimated population total.

```
> lci <- estim.pop.tot - 1.96 * se.tot
> uci <- estim.pop.tot + 1.96 * se.tot
> result1 <- list(point.estim = estim.pop.tot, lower.CI = lci,
+               upper.CI = uci)
> result1
```

```
$point.estim
[1] 54495.4
```

```
$lower.CI
[1] 43048.98
```

```
$upper.CI
[1] 65941.83
```

Table 2: Stratified Random Sample

Stratum	$N_h$	$n_h$	$\sum_j y_{hj}$	$\bar{y}_h$	$s_h^2$	$w_h$	$\hat{t}_h$
(A)	(B)	(C)	(D)	(E)	(F)	(G)	(H)
A	400	98	2362	24.1	5575	$400/98 = 4.08$	9641
B	30	10	256	25.6	4064	$30/10 = 3.00$	768
C	61	37	9901	267.6	347556	$61/37 = 1.65$	16323
D	18	6	1074	179.0	22798	$18/6 = 3.00$	3222
E	70	39	11454	293.7	123578	$70/39 = 1.79$	20558
F	120	21	697	33.2	9795	$120/21 = 5.71$	3983