# Stratified Sampling, part 2

Survey Sampling

Statistics 4234/5234

Fall 2018

September 25, 2018

Example: Chapter 3 Exercise 7

Consider the population of $N = 807$ college faculty members, stratified into $H = 4$ academic units. Let

$$y_{hj} = \text{publications by faculty member } j \text{ of academic unit } h$$

The goal is to estimate the total number of publications by the entire college faculty, and also the proportion of faculty with no publications

The data consist of a stratifed random sample, summarized here

| Stratum | $N_h$ | $n_h$ | $\bar{y}_h$ | $s_h$ | 0's |
|---|---|---|---|---|---|
| Biological Sciences | 102 | 7 | 3.14 | 2.61 | 1 |
| Physical Sciences | 310 | 19 | 2.11 | 2.87 | 10 |
| Social Sciences | 217 | 13 | 1.23 | 2.09 | 9 |
| Humanities | 178 | 11 | 0.45 | 0.93 | 8 |

We estimate $t = t_1 + t_2 + t_3 + t_4$ by

$$\begin{aligned}\widehat{t}_{\text{strat}} &= \sum_{h=1}^{H} \widehat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h \\ &= 102(3.14) + 310(2.11) + 217(1.23) + 178(0.45) \\ &= 1321.2\end{aligned}$$

Thus

$$\bar{y}_{\text{strat}} = \frac{\widehat{t}_{\text{strat}}}{N} = \frac{1321.2}{807} = 1.64$$

For standard errors we find

$$\begin{aligned}
\widehat{V}\left(\hat{t}_{\text{strat}}\right) &= \sum_{h=1}^{H} N_h^2 \widehat{V}\left(\hat{t}_h\right) = \sum_{h=1}^{H} N_h^2 \frac{s_h^2}{n_h}\left(1 - \frac{n_h}{N_h}\right) \\
&= 102^2 \frac{2.61^2}{7}\left(1 - \frac{7}{102}\right) + \cdots + 178^2 \frac{0.45^2}{11}\left(1 - \frac{11}{178}\right) \\
&= 65{,}611
\end{aligned}$$

and thus

$$\text{SE}\left(\hat{t}_{\text{strat}}\right) = \sqrt{\widehat{V}\left(\hat{t}_{\text{strat}}\right)} = 256.15$$

Also

$$\text{SE}\left(\bar{y}_{\text{strat}}\right) = \text{SE}\left(\frac{\hat{t}_{\text{strat}}}{N}\right) = \frac{1}{N}\text{SE}\left(\hat{t}_{\text{strat}}\right) = \frac{256.15}{807} = 0.32$$

We estimate that this college faculty produced a total of 1.321 published works, the standard error of this estimate is 256 publications.

Equivalently, we estimate that the average publications per faculty member at this college was 1.64; the standard error of our estimate is 0.32.

Treating the data as an SRS would have given us $\bar{y} = 1.66$ and $\mathrm{SE}(\bar{y}) = 0.33$.

We now take up the estimation of the proportion of faculty members who had no publications.

$$\begin{aligned}
\widehat{p}_{\text{strat}} &= \sum_{h=1}^{H} \frac{N_h}{N} \widehat{p}_h \\
&= \frac{1}{807}\left[102\left(\frac{1}{7}\right) + 310\left(\frac{10}{19}\right) + 217\left(\frac{9}{13}\right) + 178\left(\frac{8}{11}\right)\right] \\
&= 0.57
\end{aligned}$$

For standard error we obtain

$$
\begin{aligned}
\widehat{V}\left(\widehat{p}_{\text{strat}}\right) &= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) \\
&= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2 \frac{p_h(1-p_h)}{n_h-1}\left(1-\frac{n_h}{N_h}\right) \\
&= 0.0658^2
\end{aligned}
$$

We estimate that 57% of the faculty had no publications; the standard error of this estimate is 6.6%.

Treating the data as a SRS, we'd have gotten an estimate of 56%, with a standard error of 6.9%.

## Sampling weights (sections 2.4 and 3.3)

First consider the population $\{y_1, y_2, \ldots, y_N\}$.

Recall the *inclusion probability* for the $i$th unit is

$$\pi_i = P(\text{unit } i \text{ included in sample})$$

Define the **sampling weight** of unit $i$, for a particular sampling plan, by

$$w_i = \frac{1}{\pi_i}$$

The sampling weight $w_i$ can be interpreted as the number of population units represented by unit $i$ (if unit $i$ is included in the sample).

1. Special case: Simple random sampling (SRS)

Under SRS,

$$\pi_i = \frac{n}{N} \quad \text{and} \quad w_i = \frac{N}{n}$$

Each unit in the sample represents itself plus $N/n - 1$ of the unsampled units.

Also, for SRS,

$$\sum_{i \in \mathcal{S}} w_i = \sum_{i \in \mathcal{S}} \frac{N}{n} = n \left( \frac{N}{n} \right) = N$$

and thus

$$\hat{t}_{\mathsf{SRS}} = N\bar{y} = \frac{N}{n}\sum_{i\in\mathcal{S}} y_i = \sum_{i\in\mathcal{S}} w_i y_i$$

and

$$\bar{y} = \frac{\hat{t}}{N} = \frac{\displaystyle\sum_{i\in\mathcal{S}} w_i y_i}{\displaystyle\sum_{i\in\mathcal{S}} w_i}$$

Definition: A sampling plan in which every unit has the same sampling weight is called a **self-weighting** sample.

Proposition: SRS is self-weighting.

2. Special case: stratified random sampling (section 3.3)

Now the population is

$$\left\{y_{hj} : j = 1, \ldots, N_h; h = 1, \ldots, H\right\}$$

Under stratified random sampling the inclusion probabilities are

$$\pi_{hj} = \frac{n_h}{N_h}$$

and the sampling weight for unit $j$ of stratum $h$ is

$$w_{hj} = \frac{1}{\pi_{hj}} = \frac{N_h}{n_h}$$

Again, the sum of the sampling weights of sampled units, for any set of samples $\mathcal{S}_1, \ldots, \mathcal{S}_H$, gives the number of units in the population

$$\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} \frac{N_h}{n_h} = \sum_{h=1}^{H} N_h = N$$

And again, we find that the estimators of the population total and population mean satisfy

$$\widehat{t}_{\mathsf{strat}} = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j \in \mathcal{S}_h} y_{hj} = \sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}$$

and

$$\bar{y}_{\mathsf{strat}} = \frac{\widehat{t}_{\mathsf{strat}}}{N} = \frac{\displaystyle\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj} y_{hj}}{\displaystyle\sum_{h=1}^{H} \sum_{j \in \mathcal{S}_h} w_{hj}}$$

Example: In an SRS of $n = 50$ from a population of size $N = 807$, each sampled unit represents

$$\frac{807}{50} = 16.14 \text{ units}$$

In the stratified random sample we find

| Stratum | $N_h$ | $n_h$ | $w_{hj}$ |
|---|---|---|---|
| Biological Sciences | 102 | 7 | 14.6 |
| Physical Sciences | 310 | 19 | 16.3 |
| Social Sciences | 217 | 13 | 16.7 |
| Humanities | 178 | 11 | 16.2 |

Thus each sampled social science professor represents 16.7 professors, whereas each sampled biology professor represents only 14.6.