Lecture 20-21.
Neyman Allocation vs Proportional Allocation
and
Stratified Random Sampling vs Simple Random Sampling

March 8-13, 2013
Agenda

- Neyman Allocation and its properties
- Variance of the optimal stratified estimate $\bar{X}_{n,\text{opt}}$
- Drawbacks of Neyman Allocation
- Proportional Allocation
- Neyman vs Proportional
- Stratified vs Simple
- Summary
Neyman allocation

In Lecture 19, we described the optimal allocation scheme for stratified random sampling, called Neyman allocation. Neyman allocation scheme minimizes variance $\mathbb{V}[\bar{X}_n^*]$ subject to $\sum_{k=1}^{N} n_k = n$.

**Theorem**

The sample sizes $n_1, \ldots, n_L$ that solve the optimization problem

\[
\mathbb{V}[\bar{X}_n^*] = \sum_{k=1}^{L} \omega_k^2 \frac{\sigma_k^2}{n_k} \rightarrow \min \quad \text{s.t.} \quad \sum_{k=1}^{L} n_k = n
\]

are given by

\[
\hat{n}_k = n \frac{\omega_k \sigma_k}{\sum_{j=1}^{L} \omega_j \sigma_j} \quad k = 1, \ldots, L
\]  

(1)

The theorem says that if $\omega_k \sigma_k$ is large, then the corresponding stratum should be sampled heavily. This is very natural since

- if $\omega_k$ is large, then the stratum contains a large portion of the population
- if $\sigma_k$ is large, then the population values in the stratum are quite variable and, therefore, to estimate $\mu_k$ accurately a relatively large sample size must be used
Variance of the optimal stratified estimate

In stratified random sampling, an (unbiased) estimate of $\mu$ is

$$\overline{X}_n^* = \sum_{k=1}^{L} \omega_k \overline{X}_{n_k}^{(k)}$$

If Neyman (i.e. optimal) allocation is used ($n_k = \hat{n}_k$), then the optimal stratified estimate of $\mu$, denoted by $\overline{X}_{n, opt}^*$, is

$$\overline{X}_{n, opt}^* = \sum_{k=1}^{L} \omega_k \overline{X}_{\hat{n}_k}^{(k)}$$

**Theorem**

The variance of the optimal stratified estimate is

$$\text{Var}[\overline{X}_{n, opt}^*] = \frac{1}{n} \left( \sum_{k=1}^{L} \omega_k \sigma_k \right)^2$$
Proportional Allocation

There are two main disadvantages of Neyman allocation:

1. Optimal allocations $\hat{n}_k$ depends on $\sigma_k$ which generally will not be known
2. If a survey measures several values for each population member, then it is usually impossible to find an allocation that is simultaneously optimal for all values

A simple and popular alternative method of allocation is proportional allocation: to choose $n_1, \ldots, n_L$ such that

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_L}{N_L}$$

This holds if

$$\tilde{n}_k = n \frac{N_k}{N} = n\omega_k \quad k = 1, \ldots, L \quad (2)$$
Proportional Allocation

If proportional allocation is used \((n_k = \tilde{n}_k = n\omega_k)\), then the corresponding stratified estimate of \(\mu\), denoted by \(\bar{X}_{n,p}^*\), is

\[
\bar{X}_{n,p}^* = \sum_{k=1}^{L} \omega_k \bar{X}_k^{(k)} = \sum_{k=1}^{L} \omega_k \frac{1}{\tilde{n}_k} \sum_{i=1}^{\tilde{n}_k} X_i^{(k)} = \frac{1}{n} \sum_{k=1}^{L} \sum_{i=1}^{\tilde{n}_k} X_i^{(k)}
\]

Thus, \(\bar{X}_{n,p}^*\) is simply the unweighted mean of the sample values.

**Theorem**

The variance of \(\bar{X}_{n,p}^*\) is given by

\[
\text{Var}[\bar{X}_{n,p}^*] = \frac{1}{n} \sum_{k=1}^{L} \omega_k \sigma_k^2
\]
Neyman vs Proportional

By definition, Neyman allocation is always better than proportional allocation (since Neyman allocation is optimal).

Question: When is it substantially better?

Proposition

\[ \text{Proposition} \]

\[
\begin{align*}
\text{Var}[\bar{X}_{n,p}^*] - \text{Var}[\bar{X}_{n,opt}^*] &= \frac{1}{n} \sum_{k=1}^{L} \omega_k (\sigma_k - \bar{\sigma})^2, \\
\bar{\sigma} &= \sum_{k=1}^{L} \omega_k \sigma_k
\end{align*}
\]

Therefore,

- if the variances \( \sigma_k \) of the strata are all the same, then proportional allocation is as efficient as Neyman allocation, \( \text{Var}[\bar{X}_{n,p}^*] = \text{Var}[\bar{X}_{n,opt}^*] \)
- the more variable \( \sigma_k \), the more efficient the Neyman allocation scheme
Stratified vs Simple

Let us now compare simple random sampling and stratified random sampling with proportional allocation.

**Question:** What is more efficient? (more efficient = has smaller variance)

**Proposition**

\[
\mathbb{V}[\bar{X}_n] - \mathbb{V}[\bar{X}_{n,p}] = \frac{1}{n} \sum_{k=1}^{L} \omega_k (\mu_k - \mu)^2
\]

Thus, stratified random sampling with proportional allocation always gives a smaller variance than simple random sampling does (providing that the finite population correction is ignored, \( (n - 1)/(N - 1) \approx 0 \)).
Summary

- The variance of the optimal stratified estimate (Neyman allocation) of $\mu$ is

$$\mathbb{V}[\overline{X}_{n,\text{opt}}^*] = \frac{1}{n} \left( \sum_{k=1}^{L} \omega_k \sigma_k \right)^2$$

- Neyman allocation is difficult to implement in practice

- Proportional allocation: $\tilde{n}_k = n \frac{N_k}{N} = n \omega_k$

- The variance of the stratified estimate under proportional allocation:

$$\mathbb{V}[\overline{X}_{n,p}^*] = \frac{1}{n} \sum_{k=1}^{L} \omega_k \sigma_k^2$$

- By definition, Neyman allocation is better than proportional allocation, but if the variances $\sigma_k$ of the strata are all the same, then proportional allocation is as efficient as Neyman allocation

- Stratified random sampling with proportional allocation is always more efficient than simple random sampling.