LECTURE 15: **11.1 Probability Proportional to Size (PPS) Sampling**

We will study cluster sampling in which clusters are sampled with probabilities proportional to cluster size (PPS). The second-stage sampling (within sampled clusters) can be anything.

Example Suppose there are three population clusters: hospital 1 with 3,000 outpatient surgical procedures, hospital 2 with 4,000, and hospital 3 with 10,000. The measurement of interest is the number of unnecessary outpatient surgical procedures. Assume that $X_1 = 35$, $X_2 = 38$, and $X_3 = 100$. Suppose we draw a simple one-stage cluster sample with $m = 2$. We would obtain the following sample characteristics.

Hospitals in Sample	n	x	$x'_{clu} = (M/m)\,x$
1,2	7,000	73	$(3/2)(73) = 109.5$
1,3	13,000	135	$(3/2)(135) = 202.5$
2,3	14,000	138	$(3/2)(138) = 207$

$\mathbb{E}(x'_{clu}) = (109.5 + 202.5 + 207)/3 = 173$ (unbiased since 173=35+38+100),
$SE(x'_{clu}) = \Big[\,\big[(109.5-173)^2 + (202.5-173)^2 + (207-173)^2\big]/2\,\Big]^{1/2} = 55.04.$

Note that hospital 3 has more outpatients surgical procedures, and thus more unnecessary procedures than hospitals 1 and 2. Hence, if hospital 3 is not in the sample, the population total is grossly underestimated ($109.5 < 173$), whereas if hospital 3 IS in the sample, the population total is grossly overestimated ($202.5 > 173,\ 207 > 173$). This is a clear disadvantage of the simple one-stage cluster sample where each cluster (hospital) has equal chance 1/3 of being chosen for the sample.

A better sampling method is to sample clusters with probability proportional to cluster sizes. Let N_i be the size of cluster i (in our example, the number of outpatient surgical procedures in hospital i), and N be the population size. Then the first cluster is chosen for the sample with probability

$$P_i = N_i/N\,.$$

Clusters are chosen without replacement, so the probability that clusters i and j are chosen is (typo in the book)

$$\pi_{ij} = \mathbb{P}(\textit{cluster } i \textit{ and cluster } j) = \mathbb{P}(\textit{cluster } i \mid \textit{cluster } j)\mathbb{P}(\textit{cluster } j)$$

$$+\mathbb{P}(\textit{cluster } j \mid \textit{cluster } i)\mathbb{P}(\textit{cluster } i) = \frac{N_j}{N}\Big(\frac{N_i}{N-N_j}\Big) + \frac{N_i}{N}\Big(\frac{N_j}{N-N_i}\Big)$$

$$= \frac{N_i\,N_j}{N}\Big(\frac{1}{N-N_i} + \frac{1}{N-N_j}\Big)\,.$$

In our example, the selection probabilities are

$$\pi_{12} = 0.10472,\ \pi_{13} = 0.37815,\ \pi_{23} = 0.51713\,.$$

The probability that cluster i appears in the sample is

$$\pi_i = \sum_{j=1, j\neq i}^{M} \pi_{ij}\,.$$

In our example,

$$\pi_1 = 0.48287,\ \pi_2 = 0.62185,\ \pi_3 = 0.89528\,.$$

The population total X is estimated by the *Horvitz-Thompson estimator* (1952)

$$x'_{hte} = \sum_{i=1}^{m} \frac{x_i}{\pi_i}\,.$$

In our example, if hospitals 1 and 2 are sampled, $x'_{hte} = \frac{35}{0.48287} + \frac{38}{0.62185} = 133.59$. If hospitals 1 and 3 are in the sample, $x'_{hte} = \frac{35}{0.48287} + \frac{100}{0.89528} = 184.18$. If hospitals 2 and 3, $x'_{hte} = \frac{38}{0.62185} + \frac{100}{0.89528} = 172.8$.

Note that $\mathbb{E}(x'_{hte}) = 173$, that is the Hotvitz-Thompson estimator is unbiased. Also,

$$SE(x'_{hte}) = \sqrt{\sum_{all\ samples} (x'_{hte} - \mathbb{E}(x'_{hte}))^2 \pi_{ij}} = 14.49$$

which is much less than $SE(x'_{clu}) = 55.04$.