Sampling Terminology

- **Observational unit.** An object on which a measurement is taken.
- **Target population.** The complete collection of observations under study.
- **Sample.** A subset of the population.
- **Sampled population.** The population from which the sample has been taken.
- **Sampling unit.** The unit we actually sample.
- **Sampling frame.** The list of sampling units.

Stratified Sampling

In this type of sampling, the **sampling frame** is partitioned into mutually exclusive and exhaustive groups or strata.

- If a simple random sample is then used to sample units within strata, this is called a **statified random sample**.
- The idea here is that observations within strata tend to be more homogeneous than observations across strata.

For the flights data, what variable(s) might it make sense to stratify on?

Estimates from Stratified Samples

- Let $\mathbf{S}_h$ denote the set of observational units within stratum $h = 1, \ldots, H$.
- Let N_h denote the number observational units within stratum $\mathbf{S}_h$ (stratum size).
- Let n_h denote the number of sampled units within stratum $\mathbf{S}_h$ (stratum sample size).
- Let y_{jh} denote the value of some quantitative characteristic for the jth sampled unit within $\mathbf{S}_h$.
- Let $\overline{Y}_h$ and s_h^2 denote the sample mean and variance, respectively for $\mathbf{S}_h$.

Estimates of Means

1. Within stratum $\mathbf{S}_h$:

$$\begin{aligned} \overline{Y}_h &= \frac{1}{n_h} \sum_{j \in \mathbf{S}_h} y_{jh} \\ \widehat{Var}(\overline{Y}_h) &= \left(1 - \frac{n_h}{N_h}\right) \frac{s_h^2}{n_h} \end{aligned}$$

2. Across all strata:

$$\begin{aligned} \overline{Y} &= \sum_{h=1}^{H} \overline{Y}_h \\ \widehat{Var}(\overline{Y}) &= \frac{1}{N^2} \sum_{h=1}^{H} N_h (N_h - n_h) \frac{s_h^2}{n_h} \end{aligned}$$

Estimates of Totals

1. Within $\mathbf{S}_h$:

$$\begin{aligned} \hat{t}_h &= \frac{N_h}{n_h} \sum_{j \in \mathbf{S}_h} y_{jh} \\ \widehat{Var}(\hat{t}_h) &= N_h^2 \widehat{Var}(\overline{Y}_h) \end{aligned}$$

2. Across all strata:

$$\begin{aligned} \hat{t} &= \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \overline{Y}_h \\ \widehat{Var}(\hat{t}) &= \sum_{h=1}^{H} \widehat{Var}(\hat{t}_h) \end{aligned}$$

Estimates of Proportions

Now suppose that y_{jh} is a Bernoulli observation.

1. Within $\mathbf{S}_h$:

$$\begin{aligned} \hat{p}_h &= \frac{\sum_{j \in \mathbf{S}_h} y_{jh}}{n_h} \\ \widehat{Var}(\hat{p}_h) &= \left(1 - \frac{n_h}{N_h}\right)\left(\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}\right) \end{aligned}$$

2. Across strata:

$$\begin{aligned} \hat{p} &= \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h \\ \widehat{Var}(\hat{p}) &= \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right)\left(\frac{N_h}{N}\right)^2 \left(\frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}\right) \end{aligned}$$

Properties

- The stratified estimators of a population mean or a population proportion (i.e., the weighted averages) are unbiased.
- The variances of stratified estimators are sometimes smaller and sometimes larger than those of estimators based on simple random samples.
- The issue is homogeneity within strata and heterogeneity across strata.