Stat472/572 Sampling: Theory and Practice

Instructor: Yan Lu

University of New Mexico

Chapter 3: Stratified Sampling

Example: SRS vs. Stratified Sampling

Consider a population with 1000 males and 100 females.

- If we take a Simple Random Sample (SRS) of size 55, it is possible to end up with a sample containing no females.
  - Such a sample might not be representative of the population, especially if males and females respond differently to the item of interest.
- Stratified Sampling:
  - Divide the population into strata based on gender (e.g., male and female).
  - Take a proportional sample from each stratum, e.g., 50 males from the male pool and 5 females from the female pool.
  - Prevents the possibility of selecting a sample with no or very few females.
  - Improves the precision of estimators by ensuring better representation of both strata.

Stratified Sampling

- Divide the population into H distinct subpopulations, called **strata**, such that:
  - The strata are mutually exclusive (no overlap).
  - The strata collectively cover the entire population.
- Each sampling unit belongs to exactly one stratum.
- Draw an independent probability sample from each stratum.
- Combine the information from all strata to calculate overall population estimates.

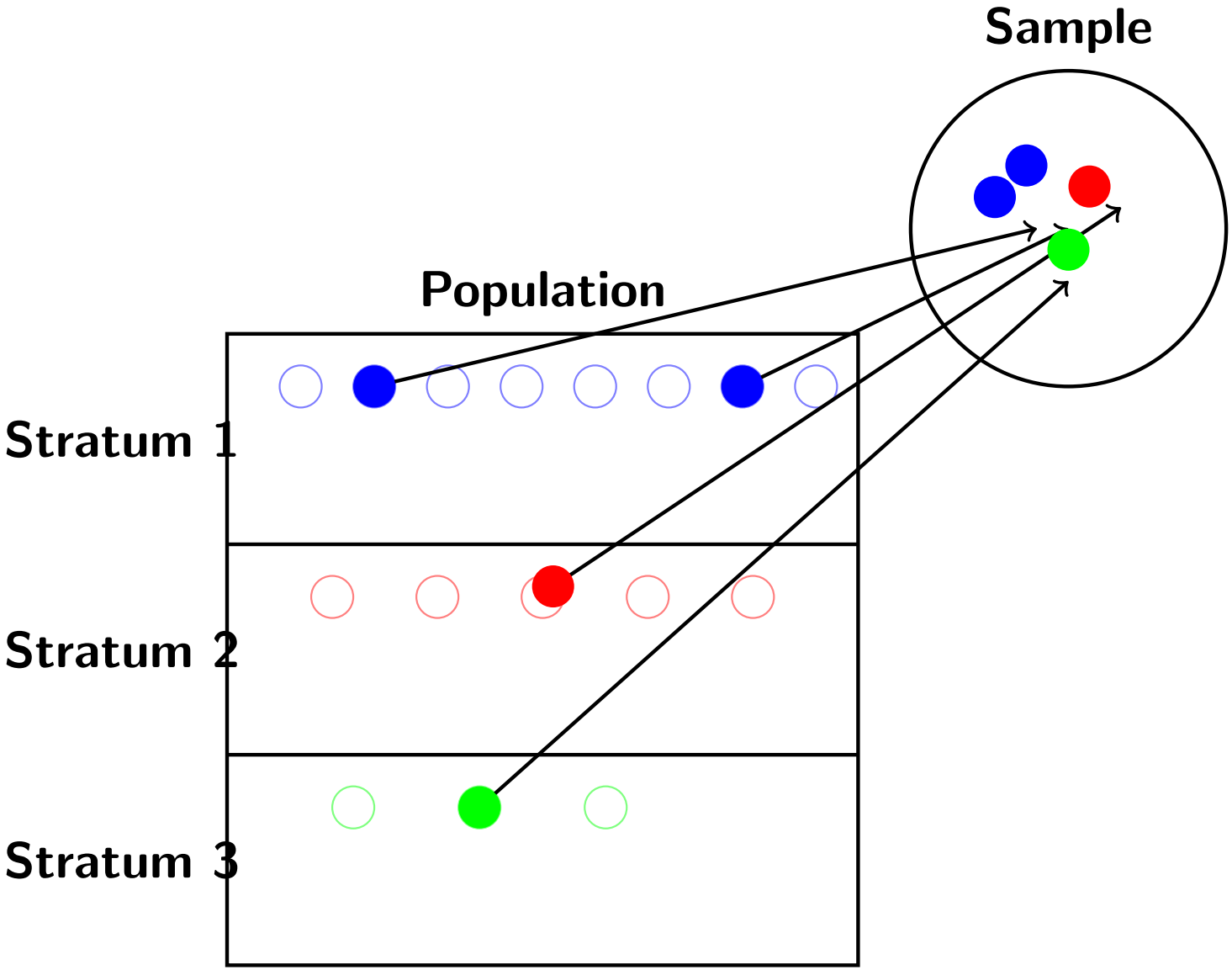

Figure 1: Illustration of Stratified Sampling

Example 3.2

- In Example 2.6, a random sample led to overrepresentation of some areas and no representation of others.
- Larger counties (common in the western US) tend to have larger y values, contributing to high variability.
- Stratified sampling provides balance in the sample based on the stratifying variable.
- Four census regions (Northeast, North Central, South, West) are used as strata. About 10% of counties in each stratum are sampled.

Visualization of the Four Regions

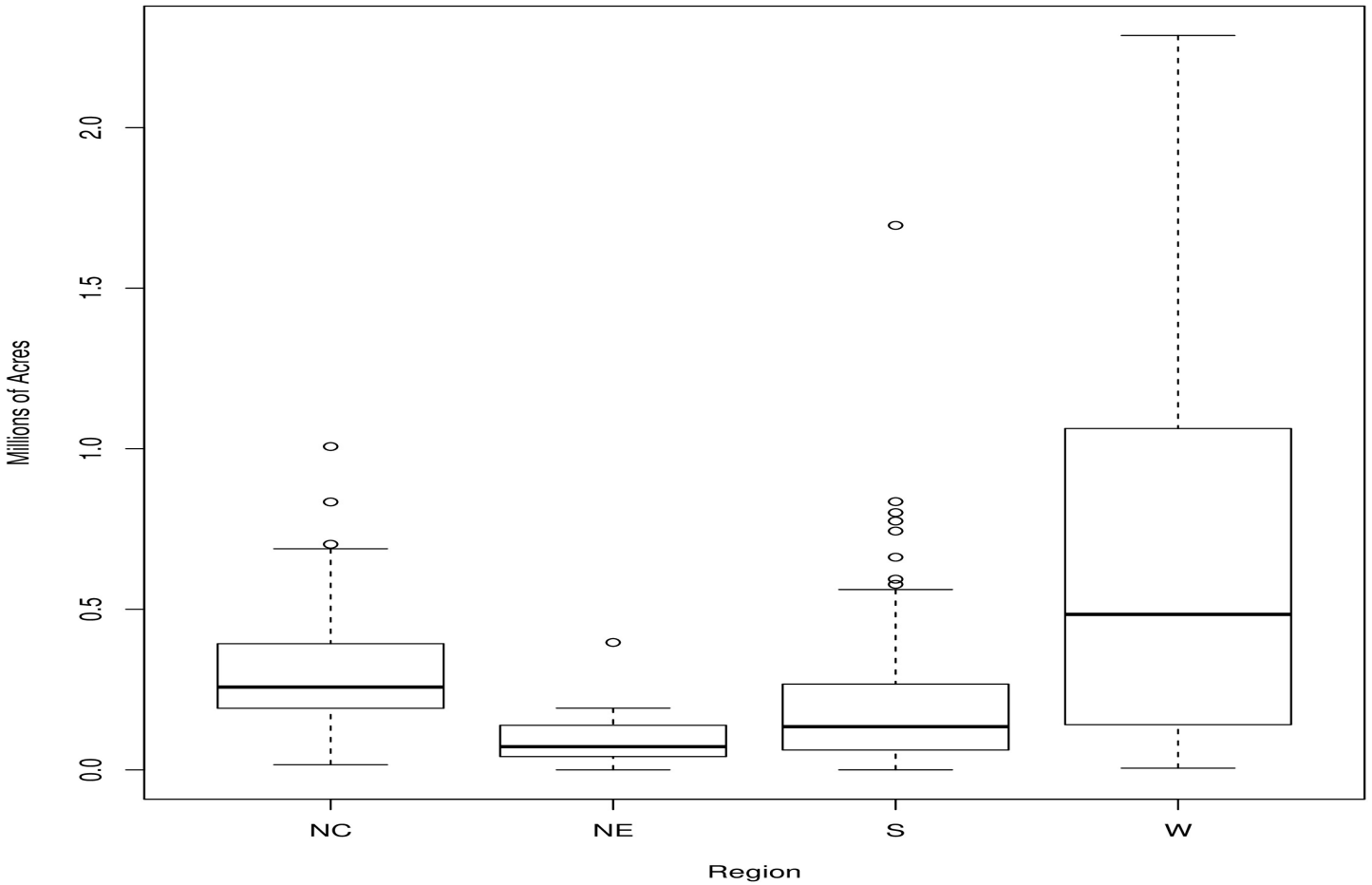

Figure 2: Boxplots of farm acreage by region: - The West has higher median and variance. - All regions exhibit positively skewed distributions.

Sampling Design

Stratum	# of Counties	# in Sample
Northeast	220	21
North Central	1054	103
South	1382	135
West	422	41
Total	3078	300

Table 1: Sample Allocation Across Strata

Summary Statistics by Stratum

Region	Stratum Size	Sample Size	Average	Variance
Northeast	220	21	97,629.8	7,647,472,708
North Central	1045	103	300,504.2	29,618,183,543
South	1382	135	211,315.0	53,587,487,856
West	422	41	662,295.5	396,185,950,266

Table 2: Stratum-Level Summary Statistics

- Example for Northeast Region:

$$\hat{t}_1 = 220 \cdot 97,629.8 = 21,478,558.2$$

$$\hat{V}(\hat{t}_1) = 220^2 \cdot \left(1 - \frac{21}{220}\right) \cdot \frac{7,647,472,708}{21} = 1.59 \times 10^{13}$$

Population Estimates by Stratum

Region	Estimated Total	Estimated Variance
Northeast	$21,478,558.2$	1.59×10^{13}
North Central	$316,731,379.4$	2.88×10^{14}
South	$292,037,390.8$	6.84×10^{14}
West	$279,488,706.1$	1.55×10^{15}
Total	$909,736,034.4$	2.54×10^{15}

Table 3: Stratum-Level and Overall Estimates

Comparison: SRS vs. Stratified Sampling

	Sample Size	$\hat{t}$	SE
SRS	300	916,927,110	58,169,381
Stratified	300	909,736,034	50,417,248

Table 4: Comparison of Total Estimates and Variances

- Stratified sampling reduces variability in the population estimate.
- Variance ratio:

$$\frac{\text{Stratified Variance}}{\text{SRS Variance}} = \frac{2.54 \times 10^{15}}{3.38 \times 10^{15}} = 0.75$$

- Equivalent precision achieved with fewer observations:

$$300 \cdot 0.75 = 225$$

Comments on Stratified Sampling

- Reduces variability by avoiding poorly representative samples.
- Enables precise estimates for subgroups of interest.
- Often more cost-effective and practical.
- Reduces overall variance for population estimates.

Theory of Stratified Sampling

Stratum	1	2	$\cdots$	H	Total
Population Size	N_1	N_2	$\cdots$	N_H	$\sum_{h=1}^{H} N_h = N$
Sample Size	n_1	n_2	$\cdots$	n_H	$\sum_{h=1}^{H} n_h = n$
Population Total	t_1	t_2	$\cdots$	t_H	

- Take an SRS of size n_h from stratum h.
- Population total and estimator:

$$t_{str} = t_1 + t_2 + \cdots + t_H, \quad \hat{t}_{str} = \hat{t}_1 + \hat{t}_2 + \cdots + \hat{t}_H = \sum_{h=1}^{H} N_h \bar{y}_h$$

- Population mean (a weighted average of the stratum means) and estimator

$$\bar{y}_U = \frac{t_{str}}{N}, \quad \hat{\bar{y}}_U = \bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$$

Population quantities and estimators

Let y_{hj} denote the value of the jth unit in stratum h.

Population Quantities	**Estimators**
Stratum mean $\bar{y}_{hU} = \dfrac{\sum_{j=1}^{N_h} y_{hj}}{N_h}$	$\bar{y}_h = \dfrac{1}{n_h}\sum_{j\in S_h} y_{hj}$
Stratum total $t_h = \sum_{j=1}^{N_h} y_{hj}$	$\hat{t}_h = \dfrac{N_h}{n_h}\sum_{j\in S_h} y_{hj} = N_h\bar{y}_h$
Population total $t = \sum_{h=1}^{H} t_h$	$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h\bar{y}_h$
Population mean $\bar{y}_U = \dfrac{t}{N} = \dfrac{\sum_{h=1}^{H}\sum_{j=1}^{N_h} y_{hj}}{N}$	$\bar{y}_{str} = \dfrac{\hat{t}_{str}}{N} = \sum_{h=1}^{H}\dfrac{N_h}{N}\bar{y}_h$
Stratum variance $S_h^2 = \dfrac{\sum_{j=1}^{N_h}(y_{hj} - \bar{y}_{hU})^2}{N_h - 1}$	$s_h^2 = \dfrac{\sum_{j\in S_h}(y_{hj} - \bar{y}_h)^2}{n_h - 1}$

- Variance estimator of the total:

$$\hat{V}(\hat{t}_{str}) = \hat{V}(\hat{t}_1) + \hat{V}(\hat{t}_2) + \cdots + \hat{V}(\hat{t}_H)$$

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

- Variance estimator of the mean:

$$\hat{V}(\bar{y}_{str}) = \frac{1}{N^2}(\hat{V}(\hat{t}_1) + \hat{V}(\hat{t}_2) + \cdots + \hat{V}(\hat{t}_H))$$

$$\hat{V}(\bar{y}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{s_h^2}{n_h}$$

Properties of the estimators:

- $E[\hat{t}_{str}] = t$
- $E[\bar{y}_{str}] = \bar{y}_U$
- $\hat{V}(\hat{t}_{str})$ is an unbiased estimator of $V(\hat{t}_{str})$
- $\hat{V}(\bar{y}_{str})$ is an unbiased estimator of $V(\bar{y}_{str})$

$$
\begin{aligned}
E[\hat{t}_{str}] &= E\left[\sum_{h=1}^{H} N_h \bar{y}_h\right] \\
&= \sum_{h=1}^{H} N_h E(\bar{y}_h) \\
&= \sum_{h=1}^{H} N_h \bar{y}_{hU} = \sum_{h=1}^{H} t_h = t
\end{aligned}
$$

$$
\begin{aligned}
E[\bar{y}_{str}] &= E\left[\frac{\hat{t}_{str}}{N}\right] \\
&= \frac{t}{N} \\
&= \bar{y}_U
\end{aligned}
$$

Confidence Intervals for Stratified Samples

- Approximation applies under the following conditions:
  - (1) The sample sizes within each stratum are large, **or**
  - (2) The sampling design includes a large number of strata.
- According to the central limit theorem (Krewski and Rao, 1981), an approximate $100(1-\alpha)\%$ confidence interval for the population mean $\bar{y}_U$ is:

$$\bar{y}_{str} \pm z_{\alpha/2} SE(\bar{y}_{str})$$

- Some survey software packages use the percentile of a t-distribution with $n - H$ degrees of freedom instead of the percentile of the normal distribution.

Stratified Sampling for Proportions

- A special case of the mean occurs when:

$$y_i = \begin{cases} 1 & \text{if the unit has the characteristic,} \\ 0 & \text{otherwise.} \end{cases}$$

- In this case, we estimate mean as proportions.
  —-Within each stratum:

$$\bar{y}_h = \hat{p}_h, \quad s_h^2 = \frac{n_h}{n_h - 1}\hat{p}_h(1 - \hat{p}_h)$$

Formulas for Stratified Sampling of Proportions

- Stratified proportion:

$$\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$$

- Variance of stratified proportion:

$$\hat{V}(\hat{p}_{str}) = \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}$$

- Stratified total and its variance:

$$\hat{t}_{str} = \sum_{h=1}^{H} N_h \hat{p}_h$$

$$\hat{V}(\hat{t}_{str}) = N^2 \hat{V}(\hat{p}_{str})$$

Example 3.4: ACLS Stratified Random Sampling

- The American Council of Learned Societies (ACLS) conducted a stratified random sample of societies across seven disciplines.
- The study aimed to analyze publication patterns, computer use, library use, and female membership in these disciplines.
- The data is summarized in the following table:

ACLS Sampling Data

Discipline	Membership N_h	# mailed	valid returns n_h	female members(%)
Literature	9,100	915	636	38
Classics	1,950	633	451	27
Philosophy	5,500	658	481	18
History	10,850	855	611	19
Linguistics	2,100	667	493	36
Political Science	5,500	833	575	13
Sociology	9,000	824	588	26
Totals	44,000	5,385	3,835	

- Goal: Estimate the percentage and total number of female members across the societies.
- Assumptions: Nonresponse is ignored, and there are no duplicate memberships.

Estimating Female Membership

- Proportion of female members:

$$\begin{aligned}\hat{p}_{str} &= \sum_{h=1}^{7} \frac{N_h}{N} \hat{p}_h \\ &= \frac{9,100}{44,000} \times 0.38 + \cdots + \frac{9,000}{44,000} \times 0.26 \\ &= 0.2465\end{aligned}$$

- Standard error of $\hat{p}_{str}$:

$$\begin{aligned}SE(\hat{p}_{str}) &= \sqrt{\sum_{h=1}^{7} \left(1 - \frac{n_h}{N_h}\right) \left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_h(1-\hat{p}_h)}{n_h - 1}} \\ &= 0.0071\end{aligned}$$

- Estimated total number of female members:

$$\begin{aligned}\hat{t}_{str} &= 44,000 \times 0.2465 = 10,847 \\ SE(\hat{t}_{str}) &= 44,000 \times 0.0071 = 312\end{aligned}$$

Using Weights in Stratified Sampling

- **Sampling Weights:** The number of units in the population represented by each sample member (h, j), where:
  - h: Stratum index,
  - j: Element index within the stratum
- Estimator for Total:

$$\begin{aligned}
\hat{t}_{str} &= \sum_{h=1}^{H} N_h \bar{y}_h \\
&= \sum_{h=1}^{H} \sum_{j \in S_h} \frac{N_h}{n_h} y_{hj} \\
&= \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj},
\end{aligned}$$

where $w_{hj} = \frac{N_h}{n_h}$.

- Estimator for Mean:

$$\bar{y}_{str} = \frac{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj}}.$$

Example: Sampling Weights

- Population Details: A population of 2000 units consists of:
  - 1600 males (Stratum 1)
  - 400 females (Stratum 2)
- Sample Details: The sample contains 400 units:
  - 200 units from each stratum.
- Calculations:

$$\pi_{1j} = \frac{200}{1600} = \frac{1}{8}, \quad w_{1j} = \frac{1}{\pi_{1j}} = 8,$$
$$\pi_{2j} = \frac{200}{400} = \frac{1}{2}, \quad w_{2j} = \frac{1}{\pi_{2j}} = 2.$$

- Interpretation:
  - Each man in the sample represents 8 men in the population.
  - Each woman in the sample represents 2 women in the population.

Extension of Weights Notation to Stratified Random Sampling

Strata	1	2	$\cdots$	H	Total
Population size	N_1	N_2	$\cdots$	N_H	$\sum_{h=1}^{H} N_h = N$
Sample size	n_1	n_2	$\cdots$	n_H	$\sum_{h=1}^{H} n_h = n$
Population total	t_1	t_2	$\cdots$	t_H	

Key Formulas for Stratified Random Sampling

- $\pi_{hj} = \frac{n_h}{N_h}$
- $w_{hj} = \frac{N_h}{n_h}$
- Estimator for Population Total:

$$\hat{t}_{str} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} N_h \bar{y}_h = \sum_{h=1}^{H} \sum_{j \in S_h} \frac{N_h}{n_h} y_{hj} = \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}$$

- Variance of $\hat{t}_{str}$:

$$V(\hat{t}_{str}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

- Estimator for Population Mean:

$$\bar{y}_{str} = \frac{\hat{t}_{str}}{N} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \frac{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} y_{hj}}{\sum_{h=1}^{H} \sum_{j \in S_h} w_{hj}}$$

- Variance of $\bar{y}_{str}$:

$$V(\bar{y}_{str}) = \frac{V(\hat{t}_{str})}{N^2} = \sum_{h=1}^{H} \frac{N_h^2}{N^2} \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

Comments on Weights in Stratified Sampling

- Let π_{hj} denote the probability of selecting unit j from stratum h:

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

- The sum of weights equals the population size:

$$\begin{aligned} \sum_{h=1}^{H} \sum_{j \in S_h} w_{hj} &= \sum_{h=1}^{H} \sum_{j \in S_h} \frac{N_h}{n_h} \\ &= \sum_{h=1}^{H} N_h = N \end{aligned}$$

The weights ensure that the sample represents the entire population.

Proportional Allocation:

The number of sampled units in each stratum is proportional to the size of the stratum

$$\frac{n_h}{N_h} = \frac{n}{N}, \quad n_h = n \cdot \frac{N_h}{N}$$

$$\pi_{hj} = \frac{n_h}{N_h} = \frac{n}{N} \quad \text{and} \quad w_{hj} = \frac{1}{\pi_{hj}} = \frac{N}{n}$$

Sample is self-weighting

$$\begin{aligned} \bar{y}_{str} &= \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \frac{\sum_{j \in S_h} y_{hj}}{n_h} \\ &= \sum_{h=1}^{H} \frac{1}{n} \sum_{j \in S_h} y_{hj} = \frac{1}{n} \sum_{h=1}^{H} \sum_{j \in S_h} y_{hj} \\ &= \bar{y} \end{aligned}$$

Example

- Population Details: A population of 2000 units consists of:
  - 1600 males (Stratum 1)
  - 400 females (Stratum 2)
- Different Sample Details from previous example: A sample containing 400 units is drawn using proportional allocation:
  - 320 males from Stratum 1
  - 80 females from Stratum 2
- Calculations:

$$\begin{aligned} \pi_{1j} &= \frac{320}{1600} = \frac{1}{5}, \quad w_{1j} = \frac{1}{\pi_{1j}} = 5, \\ \pi_{2j} &= \frac{80}{400} = \frac{1}{5}, \quad w_{2j} = \frac{1}{\pi_{2j}} = 5. \end{aligned}$$

- Key Feature: The number of sampled units in each stratum is proportional to the size of the stratum, 20%. Sample is self-weighting with a weight of 5.

Variances of mean and total from proportional stratified random sampling:

$$V_{prop}(\bar{y}_{str}) = \left(1 - \frac{n}{N}\right) \frac{1}{n} \sum_h \frac{N_h}{N} S_h^2$$

$$V_{prop}(\hat{t}_{str}) = \left(1 - \frac{n}{N}\right) \frac{N}{n} \sum_h N_h S_h^2$$

ANOVA Table

SS	df	Sum of Squares
Between strata SSB	$H-1$	$\sum_{h=1}^{H}\sum_{j=1}^{N_h}(\bar{y}_{hU}-\bar{y}_U)^2 = \sum_{h=1}^{H} N_h(\bar{y}_{hU}-\bar{y}_U)^2$
Within strata SSW	$N-H$	$\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_{hU})^2 = \sum_{h=1}^{H}(N_h-1)S_h^2$
Total corrected SSTO	$N-1$	$\sum_{h=1}^{H}\sum_{j=1}^{N_h}(y_{hj}-\bar{y}_U)^2 = (N-1)S^2$

SSTO = SSB +SSW

Comparison between SRS and proportional allocation

$$
\begin{aligned}
V_{prop}(\hat{t}_{str}) &= V\left(\sum_{h=1}^{H} N_h \bar{y}_h\right) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\
&= \sum_{h=1}^{H} \left(1 - \frac{n}{N}\right) \frac{N}{nN_h} N_h^2 S_h^2 = \sum_{h=1}^{H} \left(1 - \frac{n}{N}\right) \frac{N}{n} N_h S_h^2 \\
&= \left(1 - \frac{n}{N}\right) \frac{N}{n} \left[SSW + \sum_{h=1}^{H} S_h^2\right]
\end{aligned}
$$

$$
\begin{aligned}
V(\hat{t}_{srs}) &= \left(1 - \frac{n}{N}\right) N^2 \frac{S^2}{n} \\
&= \left(1 - \frac{n}{N}\right) \frac{N^2}{n} \frac{1}{N-1} (SSW + SSB) \\
&\approx \left(1 - \frac{n}{N}\right) \frac{N}{n} (SSW + SSB)
\end{aligned}
$$

Efficiency of Proportional Stratification

- Proportional stratification is more efficient if:

$$\sum_{h=1}^{H} S_h^2 < SSB$$

where $SSB = \sum_{h=1}^{H} N_h(\bar{y}_{hU} - \bar{y}_U)^2$.

- This condition is usually satisfied since the large population sizes of the strata force:

$$N_h(\bar{y}_{hU} - \bar{y}_U)^2 > S_h^2$$

Comments:

- The variance of the estimator $\hat{t}$ from a stratified sample with proportional allocation is generally smaller than the variance of the estimator from SRS, given the same number of observations.
- The more unequal the stratum means $\bar{y}_{hU}$ and the more homogeneous the units within each stratum, the greater the precision gained from proportional allocation.

Optimal Allocation

Example: Sampling American Corporations to Estimate Trade with Europe

- The variation among large corporations is often greater than the variation among small ones:
  - Large units tend to exhibit more variability than smaller units.
- To account for this variability:
  - A higher percentage of large corporations needs to be sampled.
- Limitations of Proportional Allocation:
  - Proportional allocation samples the same percentage within each stratum.
  - Works well if the variances S_h^2 are similar across strata.
  - Performs poorly when the variances S_h^2 differ significantly between strata.
- Optimal Allocation:
  - Take more units from strata with larger variances S_h^2.
  - Balances sampling effort with variability to achieve greater precision.

Cost Function

- Cost function:

$$c = c_0 + \sum_{h=1}^{H} c_h n_h$$

where:

  - c_0: Overhead costs (e.g., maintaining an office)
  - c_h: Cost of sampling a unit in stratum h
  - n_h: Number of sampled units in stratum h

- Objective:
  - Minimize $V(\hat{t}_{str})$ for a fixed cost c
  - Minimize c for a fixed $V(\hat{t}_{str})$

- Variance of $\hat{t}_{str}$:

$$V(\hat{t}_{str}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h}$$

Expanding and simplifying:

$$V(\hat{t}_{str}) = \sum_{h=1}^{H} N_h^2 \frac{S_h^2}{n_h} - \sum_{h=1}^{H} N_h S_h^2$$

- Equivalent to minimizing:

$$\sum_{h=1}^{H} N_h^2 \frac{S_h^2}{n_h}$$

Deriving Optimal Allocation

- Objective function:

$$f = \sum_{h=1}^{H} N_h^2 \frac{S_h^2}{n_h} + \lambda \left(c_0 + \sum_{h=1}^{H} c_h n_h - c \right)$$

- Take the partial derivative with respect to n_h and set to 0:

$$\frac{\partial f}{\partial n_h} = \frac{-N_h^2 S_h^2}{n_h^2} + \lambda c_h = 0$$

- Solving for n_h:

$$n_h = \frac{N_h S_h}{\sqrt{c_h \lambda}}$$

- Using the constraint $\sum_{h=1}^{H} n_h = n$, we find:

$$\frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_{l=1}^{H} \frac{N_l S_l}{\sqrt{c_l}}}$$

- Final result for optimal allocation:

$$n_{h,opt} = n \times \left(\frac{N_h S_h / \sqrt{c_h}}{\sum_{l=1}^{H} N_l S_l / \sqrt{c_l}} \right)$$

- Simplified proportionality:

$$n_{h,opt} \propto \frac{N_h S_h}{\sqrt{c_h}}$$

We take a larger sample from stratum h if

- The stratum size N_h is large
- The variance within the stratum S_h is large
- The sampling within the stratum c_h is inexpensive

Neyman Allocation

- Optimal allocation formula:

$$n_{h,opt} = n \times \left(\frac{N_h S_h / \sqrt{c_h}}{\sum_{l=1}^{H} N_l S_l / \sqrt{c_l}} \right)$$

- **Neyman Allocation**: Assumes equal costs for all strata (c_h's are equal):

$$n_{h,\text{Neyman}} = n \times \left(\frac{N_h S_h}{\sum_{l=1}^{H} N_l S_l} \right)$$

- Simplify with a constant factor a:

$$a = \frac{n}{\sum_{l=1}^{H} N_l S_l}$$

$$n_{h,\text{Neyman}} = a \times N_h S_h$$

- Key insight:
  - The sample size for each stratum is proportional to the product of the population size N_h and standard deviation S_h.

Variance of Neyman Allocation

$$
\begin{aligned}
V(\hat{t}_{str,\text{Neyman}}) &= \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2 S_h^2}{n_h} \\
&= \sum_{h=1}^{H} \left(1 - \frac{aN_h S_h}{N_h}\right) \frac{N_h^2 S_h^2}{aN_h S_h} \\
&= \sum_{h=1}^{H} (1 - aS_h) \frac{N_h S_h}{a} \\
&= \sum_{h=1}^{H} \left(1 - \frac{n}{\sum_{l=1}^{H} N_l S_l} S_h\right) \frac{N_h S_h \sum_{l=1}^{H} N_l S_l}{n} \\
&= \frac{\sum_{h=1}^{H} N_h S_h \sum_{l=1}^{H} N_l S_l}{n} - \sum_{h=1}^{H} N_h S_h^2
\end{aligned}
$$

Variance of Proportional Allocation

$$
\begin{aligned}
V(\hat{t}_{str,\text{Prop}}) &= \sum_{h=1}^{H} \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{n_h} S_h^2 \\
&= \sum_{h=1}^{H} \left(1 - \frac{n}{N}\right) \frac{N}{n} N_h S_h^2 \\
&= \sum_{h=1}^{H} \frac{N}{n} N_h S_h^2 - \sum_{h=1}^{H} N_h S_h^2
\end{aligned}
$$

Comparison of Neyman and proportional allocation, and SRS given the same sample size

$$\begin{aligned}\sum_{h=1}^{H} N_h S_h \sum_{l=1}^{H} N_l S_l &= \sum_{h=1}^{H} N_h^2 S_h^2 + 2\sum_{i=1}^{H}\sum_{j>i}^{H} N_i N_j S_i S_j \\ \sum_{h=1}^{H} N_h S_h^2 &= \sum_{h=1}^{H} N_h^2 S_h^2 + \sum_{i=1}^{H}\sum_{j>i}^{H} N_i N_j (S_i^2 + S_j^2)\end{aligned}$$

$$V(\hat{t}_{str,\text{Neyman}}) \leq V(\hat{t}_{str,\text{Prop}})$$

Relative Precision: $V(\hat{t}_{str,\text{Neyman}}) \leq V(\hat{t}_{str,\text{Prop}}) \leq V_{srs}(\hat{t})$

Example 3.3: Siniff and Skoog (1964) - Estimation of Nelchina Herd

Siniff and Skoog (1964) used stratified random sampling to estimate the size of the Nelchina herd of Alaska caribou in February of 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.

Example 3.10: Caribou Survey - Neyman Allocation

The caribou survey in Example 3.3 used Neyman allocation to determine the n_h.

- Before taking the survey, the investigators constructed strata to be relatively homogeneous in terms of population density.
- Total sample size was set as $n = 225$.
- Estimated count in each stratum is served as a rough estimate of the standard deviation.
- Wanted the sampling fraction to be at least $1/3$ in smaller strata.
- Used the Neyman allocations as a guideline for determining the final sample sizes in the last column of Table 5.

Stratum	N_h	s_h	$N_h s_h$	$n_h = 225\frac{N_h s_h}{\sum N_l s_l}$	Sample Size
A	400	3,000	1,200,000	96.26	98
B	30	2,000	60,000	4.81	10
C	61	9,000	549,000	44.04	37
D	18	2,000	36,000	2.89	6
E	70	12,000	840,000	67.38	39
F	120	1,000	120,000	9.63	21
Total	**699**		**2,805,000**	**225.00**	**221**

Table 5: Optimal allocation for Caribou Herd Example

Some design issues of stratified random sampling

- Allocating observations to strata
  —-Proportional allocation: $\frac{n_h}{N_h} = \frac{n}{N}$
  —-Neyman allocation (special case of optimal allocation when c_h's are all equal)

$$n_{h,\text{Neyman}} = n \left(\frac{N_h S_h}{\sum_{l=1}^{H} N_l S_l} \right)$$

- Sample size
- Defining strata: variables and number of strata

Example 3.12: Stratified Sampling for U.S. Colleges and Universities

The file `college.csv`, abstracted from online data published by the U.S. Department of Education (2020), contains selected variables from a population of 1,372 U.S. colleges and universities.

- The goal is to estimate the total instructional budget or the number of students who have full-time employment five years after graduation. These variables are thought to be related to the size of the institution.
- The institutions are grouped into ten strata based on size and residential status, as defined by the variable `ccsizset`.
- For each stratum h, S_h represents the population standard deviation of the undergraduate enrollment (`ugds`).

TABLE 3.8
Allocations for a stratified sample of size 200 in Example 3.12.

Stratum	N_h	S_h	Proportional	Neyman
Very small	195	251	28	3
Small, primarily nonresidential	45	784	7	2
Small, primarily residential	123	515	18	3
Small, highly residential	347	508	51	10
Medium, primarily nonresidential	80	2490	12	11
Medium, primarily residential	160	2150	23	19
Medium, highly residential	158	1473	23	13
Large, primarily nonresidential	95	11273	14	59
Large, primarily residential	126	9178	18	64
Large, highly residential	43	6844	6	16
Total	1372		200	200

Figure 3: Proportional and Neyman allocation example for college data

Source: Table 3.8 of Sampling: Design and Analysis, 3rd edition, by Sharon L. Lohr

Determining sample size

$$
\begin{aligned}
V(\hat{t}_{str}) &= \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{S_h^2}{n_h} \\
&\leq \sum_{h=1}^{H} N_h^2 \cdot \frac{S_h^2}{n_h} = \frac{1}{n} \sum_{h=1}^{H} \frac{n}{n_h} N_h^2 S_h^2 = v/n
\end{aligned}
$$

- v depends on stratum size N_h, variances S_h^2, and on the relative sample sizes n_h/n
- v can be thought of as the "average" variability per observation unit in a stratified random sample with the specified allocation
  95 % CI: $\hat{t}_{str} \pm z_{\alpha/2}\sqrt{v/n}$
  $z_{\alpha/2}\sqrt{v/n} = e$, $n = z_{\alpha/2}^2 v/e^2$

Defining Strata

1. Variables for Stratification

- Variables should be highly associated with the variables of interest.
- For example:
  - To estimate total business expenditures on advertising, we might stratify by the number of employees, size of the business, and type of product or service.
  - For farm income, we could use farm size as a stratifying variable, as larger farms are expected to have higher incomes.
- The stratifying variables must be known for all sampling units in the population.

2. **Number of Strata:**

- The number of strata depends on several factors:
  - The difficulty in constructing a sampling frame with stratifying information.
  - The cost of stratification.
- Various formulas in the literature provide guidelines for determining the number of strata.
- A pilot study can help in deciding the number of strata.
- General rule: The more information you have about the population, the more strata you should use. If little prior information about the target population is available, you should consider using simple random sampling (SRS).

Stratification or SRS?

Recall: Relative precision of stratification and SRS

$$V(\hat{t}_{str,\text{opt}}) \leq V(\hat{t}_{str,\text{prop}}) \leq V_{srs}(\hat{t})$$

1. Stratified sampling provides higher precision than SRS, but why conduct SRS?
   - Stratification adds complexity to the survey, which may not justify the small gain in precision.
   - Information is needed about which units belong to each stratum and how many units there are.
2. When is stratified sampling efficient?
   - SSB (between-strata variation) is large (strata means differ greatly).
   - SSW (within-strata variation) is small (low variability within stratum).

Example 3.13: Homeless Counts in the United States

In the United States, state and local agencies responsible for coordinating homeless services produce annual estimates of the number and characteristics of persons experiencing homelessness.

- Each agency selects one of the last ten nights in January to conduct the count of the:
  1. Sheltered population (persons in emergency shelters or transitional housing).
  2. Unsheltered population (persons staying in places not intended for human habitation, such as on the streets, under bridges, in cars, tents, bus or subway stations, or abandoned buildings).
- The survey is conducted on one night to reduce possible double-counting and minimize the effort required for teams of volunteers who canvass areas to survey the unsheltered population.

Example 3.13 (Cont.): Survey Design

- Most agencies also use the survey to disseminate information about resources and social services available for persons experiencing homelessness.
- Collect data about needs and demographics for future community outreach efforts.
- The goal is to have a sampling plan that:
  1. Produces an accurate estimate of the unsheltered population's size.
  2. Allows contact with as many unsheltered persons as possible.
- In many cities, the large land area makes it impossible for volunteers to visit every location in the city during one night.

Homeless Counts in New York City

New York City, with a land area of about 800 square kilometers, uses a carefully designed stratified random sample for its annual point-in-time count (Schneider et al., 2016; New York City Department of Homeless Services, 2019).

- The city is divided into approximately 7,000 areas, each classified as "high-density" or "low-density":
  1. High-density: Areas in Manhattan or the subway system are considered high-density if expected to contain at least two persons experiencing homelessness.
  2. High-density: Areas in the Bronx, Queens, Staten Island, or Brooklyn are considered high-density if expected to contain at least one person.
  3. Low-density: Areas not classified as high-density.
- Strata: Six high-density strata, six low-density strata.

Sampling Strategy in New York City

- Sampling fraction: 100% for high-density strata (all areas in high-density strata are canvassed).
- Sampling fractions for low-density strata vary by borough and year, designed to give high-precision estimates for each borough and the subway system.
- In 2019, about 21% of the 7,000 areas in the city were canvassed.

Advantages of Stratified Sampling in NYC

The stratified sampling design allows New York City to:

- Achieve its objectives for the one-night count with the number of volunteers available (typically between 2,000 and 2,500).
- Deploy most volunteers to areas thought to contain unsheltered persons, obtaining a large number of unsheltered persons in the sample, and gather more information about their characteristics. This still produces statistically valid estimates due to random sampling from each low-density stratum.
- However, the point-in-time count still has measurement error for estimating the number of persons experiencing homelessness on the night of the count:
  1. Some persons remain out of sight or in locations deemed unsafe for volunteers to visit.
  2. Volunteers may mistakenly identify persons as not homeless when they are, or as homeless when they are not.
  3. Volunteers may fail to survey some individuals along their routes.

Model-based inference for stratified sampling

- The one-way ANOVA model with fixed effects provides an underlying structure for stratified sampling:

$$y_{hj} = \mu_h + \epsilon_{hj} \tag{1}$$

  where ϵ_{hj} are independent with mean 0 and variance σ_h^2.
- The least squares estimator of μ_h is $\bar{y}_h$, the average in stratum h

Estimators and Properties:

- $T_h = \sum_{j=1}^{N_h} y_{hj}$: the total in stratum h
- $T = \sum_{h=1}^{H} T_h$: the overall total
- Note that both T_h and T are random variables under model (1)
- The best linear unbiased estimator for T_h is $\hat{T}_h = \frac{N_h}{n_h} \sum_{j \in S_h} y_{hj}$.
- $E_M[\hat{T}_h - T_h] = 0$
- $E_M[(\hat{T}_h - T_h)^2] = N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{\sigma_h^2}{n_h}$

By the fact that observations in different strata are independent under model (1),

$$
\begin{aligned}
Var(\hat{T}) &= E_M\left[\left\{\sum_{h=1}^{H}(\hat{T}_h - T_h)\right\}^2\right] \\
&= E_M\left[\sum_{h=1}^{H}(\hat{T}_h - T_h)^2 + \sum_{h=1}^{H}\sum_{k\neq h}(\hat{T}_h - T_h)(\hat{T}_k - T_k)\right] \\
&= E_M\left[\sum_{h=1}^{H}(\hat{T}_h - T_h)^2\right] \\
&= \sum_{h=1}^{H} N_h^2\left(1 - \frac{n_h}{N_h}\right)\frac{\sigma_h^2}{n_h}
\end{aligned}
$$

Comments:

- The theoretical variance σ_h^2 can be estimated by s_h^2.
- Adopting the model in (1) results in the same estimators for t and its standard error as found under randomization theory.
- However, if a different model is used, different estimators will be obtained.

Disclaimer

Slides are intended for a course based on the book: *Sampling: Design and Analysis, Third Edition* by Sharon L. Lohr
All examples, datasets, and pages directly scanned and included in this chapter are from the textbook.
References to papers or books cited in these slides can be found in the Bibliography section of the textbook.