Stat472/572 Sampling: Theory and Practice

Instructor: Yan Lu

University of New Mexico

Chapter 2: Simple Probability Samples

Probability Sampling

A sampling method in which every member of a population has a known, non-zero chance (or probability) of being selected.
Commonly used probability sampling methods:

$$\begin{cases} \text{simple random sampling} \\ \text{stratified sampling} \\ \text{cluster sampling} \end{cases}$$

Assumption through Chapter 7:

- sampled population = target population
- no missing data
- no measurement error

Simple Random Sample without replacement (SRS):

Randomly select a size n sample from size N population

- Every possible subset of n units in the population has the same chance to be in the sample
- Each individual is equally likely to be in the sample

Framework for Probability Sampling:

- $U = \{1; 2; \cdots ; N\}$: universe finite population
- S: sample
- $\pi_i = p(\text{unit } i \text{ in the sample})$
-

	Population quantities	Sample quantities
Size	N	n
Mean	$\bar{y}_U = \sum_{i=1}^{N} y_i / N$	$\bar{y}_S = \sum_{i \in S} y_i / n$
Total	$t = \sum_{i=1}^{N} y_i$	$\hat{t} = N\bar{y}_S$
Variance	$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \bar{y}_U)^2$	$s^2 = \frac{1}{n-1} \sum_{i \in S} (y_i - \bar{y}_S)^2$

- $E[\hat{t}] = \sum_S \hat{t}_S p(S)$
- Bias $[\hat{t}] = E[\hat{t}] - t$
-

$$\begin{aligned} V(\hat{t}) &= E[(\hat{t} - E[\hat{t}])^2] \\ &= \sum_S p(S)(\hat{t}_S - E[\hat{t}])^2 \end{aligned}$$

-

$$\begin{aligned} MSE(\hat{t}) &= E[(\hat{t} - t)^2] \\ &= E[\hat{t} - E(\hat{t}) + E(\hat{t}) - t]^2 \\ &= E[(\hat{t} - E(\hat{t}))^2] + (E(\hat{t}) - t)^2 \\ &= V(\hat{t}) + (\text{bias}[\hat{t}])^2 \end{aligned}$$

- Unbiased, $E[\hat{t}] = t$ (Archer A)
- Precise, $V[\hat{t}] = E[(\hat{t} - E(\hat{t}))^2]$ is small, measures how close estimates from different samples are to each other (Archer B)
- Accurate, $MSE[\hat{t}] = E[(\hat{t} - t)^2]$ is small, measures how close the estimate is to the true value (Archer C)

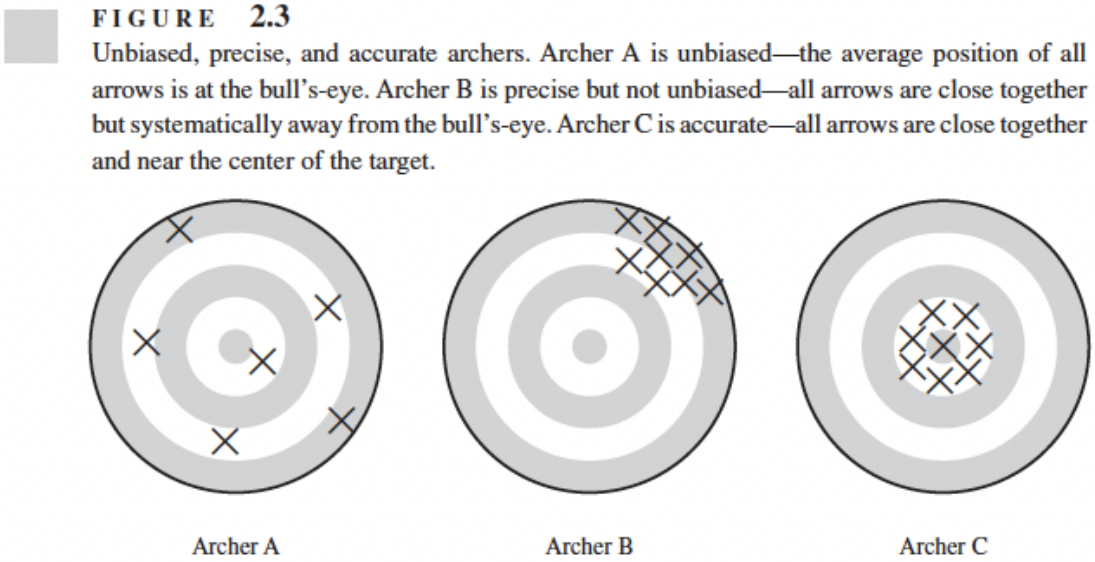


Figure 1: Unbiased, precise and accurate archers

Source: Figure 2.3 of Sampling: Design and Analysis, 3rd edition, by Sharon L. Lohr

Example 1:

$U = \{1, 2, 3, 4\}, N = 4, n = 2$

$\begin{pmatrix} 4 \\ 2 \end{pmatrix}$ possible samples

$S_1 = \{1, 2\}, S_2 = \{1, 3\}, S_3 = \{1, 4\}$

$S_4 = \{2, 3\}, S_5 = \{2, 4\}, S_6 = \{3, 4\}$

In probability sampling, each of the possible samples S from the population has a known probability $p(S)$ of being selected, and $\sum p(S) = 1$

SRS: $p(S_i) = 1/\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 1/6$

$\pi_i = p(\text{unit } i \text{ in the sample})$

$= \begin{pmatrix} N-1 \\ n-1 \end{pmatrix} / \begin{pmatrix} N \\ n \end{pmatrix} = n/N = 2/4 = 1/2$

We can also set up different selection probability for different samples

Example 2: $U = \{1, 2, 3, 4\}, N = 4, n = 2$

$S_1 = \{1, 2\}, S_2 = \{1, 3\}, S_3 = \{1, 4\}$

$S_4 = \{2, 3\}, S_5 = \{2, 4\}, S_6 = \{3, 4\}$

Assume a list of letters, a, a, b, f, f, f

- if a is chosen, S_1 is the sample;
  if b is chosen, S_2 is the sample;
  if f is chosen, S_6 is the sample
- $P(S_1) = 1/3$, $P(S_2) = 1/6$, $P(S_6) = 1/2$,
  $P(S_3) = P(S_4) = P(S_5) = 0$
  What are the π_is?
  $\pi_i = p(\text{unit } i \text{ in the sample})$

$$
\begin{aligned}
\pi_1 &= P(\text{unit 1 is in the sample}) \\
&= P(\text{unit 1} \in S_1 \text{ or } S_2 \text{ or } S_3) \\
&= P(S_1) + P(S_2) + P(S_3) \\
&= \frac{1}{3} + \frac{1}{6} + 0 \\
&= \frac{1}{2}
\end{aligned}
$$

$$\pi_2 = p(S_1) + p(S_4) + p(S_5) = 1/3 + 0 + 0 = 1/3$$
$$\pi_3 = p(S_2) + p(S_4) + p(S_6) = 1/6 + 0 + 1/2 = 2/3$$
$$\pi_4 = p(S_3) + p(S_5) + p(S_6) = 0 + 0 + 1/2 = 1/2$$
$$\sum_{i=1}^{4} \pi_i = 2$$

In general, $\sum_{i=1}^{N} \pi_i = n$

Proof: Let

$$Z_i = \begin{cases} 1 & \text{if unit} \quad i \in S \\ 0 & \text{otherwise} \end{cases}$$

$p(Z_i = 1) = \pi_i$

$E(Z_i) = 1 * p(Z_i = 1) = \pi_i$

$$\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} E(Z_i) = E\left(\sum_{i=1}^{N} Z_i\right) = n$$

Continue with Example 2, suppose values of y_i is equivalent to i,

i	1	2	3	4
y_i	1	2	3	4

Possible Sample	$p(S)$	sample values	$\bar{y}_S$	$\hat{t}_S = N\bar{y}_S$
$S_1 = \{1, 2\}$	$1/3$	1,2	$\bar{y}_{S_1} = 1.5$	$\hat{t}_1 = 6$
$S_2 = \{1, 3\}$	$1/6$	1,3	$\bar{y}_{S_2} = 2$	$\hat{t}_2 = 8$
$S_6 = \{3, 4\}$	$1/2$	3,4	$\bar{y}_{S_3} = 3.5$	$\hat{t}_3 = 14$

$$t = \sum_{i=1}^{4} y_i = 1 + 2 + 3 + 4 = 10$$

$$\begin{aligned} E[\hat{t}] &= \sum_S \hat{t}_S p(S) \\ &= \hat{t}_1 p(S_1) + \hat{t}_2 p(S_2) + \hat{t}_6 p(S_6) \\ &= 6 \times \frac{1}{3} + 8 \times \frac{1}{6} + 14 \times \frac{1}{2} = \frac{31}{3} \end{aligned}$$

$\text{Bias}(\hat{t}) = E(\hat{t}) - t = 31/3 - 10 = 1/3$, therefore, $\hat{t}$ is biased.

Continue with Example 2, now with same probability of selection

Possible S	$p(S)$	sample values	$\bar{y}_S$	$\hat{t}_S$
$S_1 = \{1, 2\}$	1/6	1,2	$\bar{y}_{S_1} = 1.5$	$\hat{t}_1 = 6$
$S_2 = \{1, 3\}$	1/6	1,3	$\bar{y}_{S_2} = 2$	$\hat{t}_2 = 8$
$S_3 = \{1, 4\}$	1/6	1,4	$\bar{y}_{S_3} = 2.5$	$\hat{t}_3 = 10$
$S_3 = \{2, 3\}$	1/6	2,3	$\bar{y}_{S_3} = 2.5$	$\hat{t}_3 = 10$
$S_3 = \{2, 4\}$	1/6	2,4	$\bar{y}_{S_3} = 3$	$\hat{t}_3 = 12$
$S_3 = \{3, 4\}$	1/6	3,4	$\bar{y}_{S_3} = 3.5$	$\hat{t}_3 = 14$

$$t = \sum_{i=1}^{4} y_i = 1 + 2 + 3 + 4 = 10$$

$$\begin{aligned} E[\hat{t}] &= \sum_{S} \hat{t}_S p(S) \\ &= \hat{t}_1 p(S_1) + \hat{t}_2 p(S_2) + \hat{t}_3 p(S_3) + \hat{t}_4 p(S_4) + \hat{t}_5 p(S_5) + \hat{t}_6 p(S_6) \\ &= 6 \times \frac{1}{6} + 8 \times \frac{1}{6} + 10 \times \frac{1}{6} + 10 \times \frac{1}{6} + 12 \times \frac{1}{6} + 14 \times \frac{1}{6} \\ &= 10 \end{aligned}$$

$\hat{t}$ is unbiased for t

$$
\begin{aligned}
V(\hat{t}) &= \sum_S p(S)(\hat{t}_S - E[\hat{t}])^2 \\
&= p(S_1)(\hat{t}_{S_1} - E[\hat{t}])^2 + p(S_2)(\hat{t}_{S_2} - E[\hat{t}])^2 \\
&+ \cdots + p(S_6)(\hat{t}_{S_6} - E[\hat{t}])^2 \\
&= \frac{20}{3}
\end{aligned}
$$

$$
\mathrm{MSE}(\hat{t}) = V(\hat{t}) + \mathrm{bias}^2 = \frac{20}{3}
$$

Remarks on Probability Samples:

- Probability samples ensure that every population unit has a nonzero chance of being selected.
- They allow us to quantify the likelihood that our sample is a "good" one:
  - A single probability sample is not guaranteed to be representative of the population for a specific criterion.
  - However, we can quantify how often the sample meets a criterion of representativeness, similar to interpreting a 95% confidence interval as containing the true parameter value in 95% of cases.
- In practice, we do not perform detailed computations for every sampling design, but the underlying concepts apply to all probability sampling methods.
- Typically, not all population values $y_1, y_2, \cdots, y_N$ are known; only the y_i values in the sample are observed.
- The primary objective is to estimate key population parameters, such as the total or mean.

Recall:

Simple Random Sampling Without Replacement (SRS)

Simple random sampling is the most fundamental form of probability sampling. It serves as the theoretical foundation for more complex sampling designs.

- The population is denoted by $U = \{1, 2, \cdots, N\}$.
- Each of the $\binom{N}{n}$ possible samples has an equal probability of being selected:

$$P(\text{any sample}) = \frac{1}{\binom{N}{n}}$$

- Every unit i in the population has the same probability of being selected:

$$\pi_i = \frac{n}{N}$$

Question:

Does $\pi_i =$ constant for all i imply an SRS?

Answer: $\pi_i =$ constant for all i does not necessarily imply a Simple Random Sample (SRS).

Example: 100 integers are divided into 20 clusters: $\{1,2,3,4,5\}, \{6,7,8,9,10\}, \cdots, \{96,97,98,99,100\}$. Randomly select 4 clusters, and observe the 5 observations within the selected clusters.

$$\begin{aligned}
\pi_i &= P(i \in \text{cluster } j \text{ and cluster } j \text{ is selected}) \\
&= P(i \in \text{cluster } j \mid \text{cluster } j \text{ is selected}) \cdot P(\text{cluster } j \text{ is selected}) \\
&= 1 \cdot \frac{4}{20} \\
&= \frac{1}{5}
\end{aligned}$$

This is a **cluster sample**, not an SRS.

Parameters of interest:

- Population mean: $\bar{y}_U = \sum_{i=1}^{N} y_i / N$
- Population total: $t = \sum_{i=1}^{N} y_i = N\bar{y}_U$

Estimators based on SRS:

- $\hat{\bar{y}}_U = \bar{y}_S = \frac{1}{n}\sum_{i \in S} y_i$
- $\hat{t} = N\bar{y}_S$

—-often use $\bar{y}$ to denote $\bar{y}_S$

Properties of $\bar{y}$:

1. **Expected Value:** $E(\bar{y}) = \bar{y}_U$
   - $\bar{y}$ is an unbiased estimator of the population mean $\bar{y}_U$.
2. **Variance:** $V(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{S^2}{n}$, where $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2$
   - $\left(1 - \frac{n}{N}\right)$ is called the *finite population correction (fpc)*.
   - If $n = N$, the variance is 0 because there is no sampling variability.
   - If $\frac{n}{N}$ is large, the variance is small.
   - If N is large and n is small compared to N, the fpc is approximately 1.
3. **Unbiased Estimator of Variance:** $\hat{V}(\bar{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$ is an unbiased estimator of $V(\bar{y})$, where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$.

Proof of Properties 1 and 2:

$$Z_i = \begin{cases} 1 & \text{if unit } i \in S \\ 0 & \text{otherwise} \end{cases}$$

$$\bar{y} = \sum_{i \in S} \frac{y_i}{n} = \sum_{i=1}^{N} Z_i \frac{y_i}{n}$$

$$P(Z_i = 1) = P(\text{unit } i \in S) = \frac{n}{N}, \quad E(Z_i) = E(Z_i^2) = \frac{n}{N}$$

$$\begin{aligned} V(Z_i) &= E(Z_i^2) - (E[Z_i])^2 \\ &= \frac{n}{N} - \left(\frac{n}{N}\right)^2 \\ &= \frac{n}{N}\left(1 - \frac{n}{N}\right) \end{aligned}$$

For $i \neq j$:

$$\begin{aligned} E[Z_i Z_j] &= P(Z_i = 1 \text{ and } Z_j = 1) \\ &= P(Z_j = 1 \mid Z_i = 1) \cdot P(Z_i = 1) \\ &= \left(\frac{n-1}{N-1}\right) \cdot \frac{n}{N} \end{aligned}$$

$$\begin{aligned} \text{Cov}(Z_i, Z_j) &= E[Z_i Z_j] - E[Z_i]E[Z_j] \\ &= \left(\frac{n-1}{N-1}\right) \cdot \frac{n}{N} - \left(\frac{n}{N}\right)^2 \\ &= -\frac{1}{N-1}\left(1 - \frac{n}{N}\right) \cdot \frac{n}{N} \end{aligned}$$

$$
\begin{aligned}
E(\bar{y}) &= E\left[\sum_{i=1}^{N} Z_i \frac{y_i}{n}\right] \\
&= \sum_{i=1}^{N} \frac{n}{N} \cdot \frac{y_i}{n} \\
&= \sum_{i=1}^{N} \frac{y_i}{N} \\
&= \bar{y}_U
\end{aligned}
$$

$$
\begin{aligned}
V(\bar{y}) &= \frac{1}{n^2} V\left(\sum_{i=1}^{N} Z_i y_i\right) \\
&= \frac{1}{n^2} Cov\left(\sum_{i=1}^{N} Z_i y_i, \sum_{j=1}^{N} Z_j y_j\right) \\
&= \frac{1}{n^2}\left[\sum_{i=1}^{N} y_i^2 V(Z_i) + \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j Cov(Z_i, Z_j)\right] \\
&= \frac{1}{n^2}\left[\frac{n}{N}(1-\frac{n}{N})\sum_{i=1}^{N} y_i^2 - \sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j \frac{1}{N-1}(1-\frac{n}{N})(\frac{n}{N})\right] \\
&= \frac{1}{n^2}\frac{n}{N}\left(1-\frac{n}{N}\right)\left[\sum_{i=1}^{N} y_i^2 - \frac{1}{N-1}\sum_{i=1}^{N}\sum_{j\neq i}^{N} y_i y_j\right] \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left[(N-1)\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2 + \sum_{i=1}^{N} y_i^2\right] \\
&= \frac{1}{n}\left(1-\frac{n}{N}\right)\frac{1}{N(N-1)}\left[N\sum_{i=1}^{N} y_i^2 - (\sum_{i=1}^{N} y_i)^2\right] = \left(1-\frac{n}{N}\right)\frac{S^2}{n}
\end{aligned}
$$

Estimating a Proportion: A Special Case of Estimating a Mean

Define:

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

The population proportion is:

$$p = \frac{\text{number of units with the characteristic in the population}}{N}$$

$$= \frac{\sum_{i=1}^{N} y_i}{N} = \bar{y}_U$$

The sample proportion is:

$$\hat{p} = \bar{y} = \frac{\sum_{i \in S} y_i}{n}$$

Example 2.4: Estimating a Proportion

Given:

$$U = \{1, 2, 3, 4, 5, 6, 7, 8\}, \quad N = 8, \quad n = 4$$

i	1	2	3	4	5	6	7	8
Values	1	2	4	4	7	7	7	8

Define:

$$y_i = \begin{cases} 1 & \text{if unit } i \text{ has the value 7,} \\ 0 & \text{otherwise.} \end{cases}$$

The population proportion is:

$$p = \frac{3}{8}$$

The sample proportion is:

$$\hat{p}_S = \frac{\sum_{i \in S} y_i}{4}$$

This represents the proportion of 7s in the sample.

Find $E(\hat{p})$.

Probability Distribution and Expectation of $\hat{p}$

k	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$
$p(\hat{p}=k)$	$\frac{5}{70}$	$\frac{30}{70}$	$\frac{30}{70}$	$\frac{5}{70}$

Probability Calculation: For $k = 0$:

$$p(\hat{p}=0) = \frac{\binom{5}{4}\binom{3}{0}}{\binom{8}{4}} = \frac{5}{70}$$

Expectation of $\hat{p}$:

$$E(\hat{p}) = 0 \cdot \frac{5}{70} + \frac{1}{4} \cdot \frac{30}{70} + \frac{1}{2} \cdot \frac{30}{70} + \frac{3}{4} \cdot \frac{5}{70} = \frac{3}{8}$$

Variance Calculation:

$$
\begin{aligned}
S^2 &= \frac{\sum_{i=1}^{N}(y_i - p)^2}{N-1} \\
&= \frac{\sum_{i=1}^{N} y_i^2 - 2p\sum_{i=1}^{N} y_i + Np^2}{N-1} \\
&= \frac{Np - 2p \cdot Np + Np^2}{N-1} \\
&= \frac{N}{N-1} p(1-p)
\end{aligned}
$$

$$
\begin{aligned}
V(\hat{p}) &= \left(1 - \frac{n}{N}\right) \frac{1}{n} S^2 \\
&= \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{N}{N-1} p(1-p) \\
&= \frac{N-n}{N-1} \times \frac{p(1-p)}{n}
\end{aligned}
$$

In our example,

$$\begin{aligned} V(\hat{p}) &= \frac{N-n}{N-1} \times \frac{p(1-p)}{n} \\ &= \frac{8-4}{8-1} \times \frac{\frac{3}{8}(1-\frac{3}{8})}{4} \\ &= \frac{15}{448} \end{aligned}$$

Sample Variance and Variance Estimate
The sample variance is given by:

$$\begin{aligned} s^2 &= \frac{\sum_{i \in S} (y_i - \hat{p})^2}{n-1} \\ &= \frac{\sum_{i \in S} y_i^2 - 2\hat{p} \sum_{i \in S} y_i + n\hat{p}^2}{n-1} \\ &= \frac{n\hat{p} - 2\hat{p} \cdot n\hat{p} + n\hat{p}^2}{n-1} \\ &= \frac{n}{n-1} \hat{p}(1-\hat{p}) \end{aligned}$$

The variance estimate is:

$$\begin{aligned} \hat{V}(\hat{p}) &= \left(1 - \frac{n}{N}\right) \frac{1}{n} s^2 \\ &= \left(1 - \frac{n}{N}\right) \frac{1}{n} \frac{n}{n-1} \hat{p}(1-\hat{p}) \\ &= \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1} \end{aligned}$$

Here, $\hat{p}$ is derived from the sample.

Coefficient of Variation (CV)

Under Simple Random Sampling (SRS) and $\bar{y}_U \neq 0$:

$$CV(\bar{y}) = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})} = \sqrt{1 - \frac{n}{N}} \cdot \frac{S}{\sqrt{n} \cdot \bar{y}_U}$$

$$\widehat{CV}(\bar{y}) = \frac{SE(\bar{y})}{\bar{y}} = \sqrt{1 - \frac{n}{N}} \cdot \frac{s}{\sqrt{n} \cdot \bar{y}}$$

— The standard error expressed as a percentage of the mean.

- A standardized measure of dispersion in a probability or frequency distribution.
- Measures relative variability.
- Does not depend on the unit of measurement.
- $CV(\hat{t}) = CV(\bar{y})$.

Confidence Intervals

- Repeatedly draw samples and construct confidence intervals for each sample.
- Expect approximately 95% of the resulting intervals to include the true value of the parameter (for a 95% confidence level).
- In probability sampling from a finite population:
  - Only a finite number of possible samples exist.
  - The probability of choosing each sample is known.
- This allows for the exact confidence level of a confidence interval procedure to be calculated.

Example 2.10

$U = \{1, 2, 3, 4, 5, 6, 7, 8\}, N = 8, n = 4$

i	1	2	3	4	5	6	7	8
values	1	2	4	4	7	7	7	8

- choose an arbitrary procedure for calculating a confidence interval

$$CI(\mathcal{S}) = [\hat{t}_{\mathcal{S}} - 4s_{\mathcal{S}}, \hat{t}_{\mathcal{S}} + 4s_{\mathcal{S}}] \tag{1}$$

Use (1) to illustrate the concept of a confidence interval. Let

$$\mu(\mathcal{S}) = \begin{cases} 1 & CI(\mathcal{S}) \text{ contains the population value t} = 40 \\ 0 & \text{otherwise} \end{cases}$$

Table 1: Confidence intervals for possible samples from this small population

$\mathcal{S}$	$y_i, i \in \mathcal{S}$	$\hat{t}_{\mathcal{S}}$	$s_{\mathcal{S}}$	$CI(\mathcal{S})$	$\mu(\mathcal{S})$
1, 2, 3, 4	1,2,4,4	22	1.50	[16.00, 28.00]	0
1, 2, 3, 5	1,2,4,7	28	2.65	[17.42, 38.58]	0
1, 2, 3,6	1,2,4,7	28	2.65	[17.42, 38.58]	0
1, 2, 3, 7	1,2,4,7	28	2.65	[17.42, 38.58]	0
1, 2, 3, 8	1,2,4,8	30	3.10	[17.62, 42.38]	1
$\vdots$		$\vdots$			$\vdots$

Confidence level: $\sum_{S} p(\mathcal{S})\mu(\mathcal{S}) = 0.77$

Confidence Intervals in Practice

- When taking an SRS of four elements without replacement from a population of eight elements:
  - There is a 77% chance that our sample is one of the "good" ones, where the confidence interval contains the true value (e.g., 40).
- This procedure results in a 77% confidence interval using (1).
- In practice, we take only one sample and do not know the population total t.
- Consequently, we cannot determine if our sample is:
  - One of the "good" samples (e.g., $\mathcal{S} = \{2, 3, 5, 6\}$), or
  - One of the "bad" samples (e.g., $S = \{4, 6, 7, 8\}$).
- The confidence interval provides only a probabilistic statement about how often we expect to be correct.

Assumptions for Asymptotic Results

- The population is finite.
- Aim: Apply asymptotic results in finite population sampling.
- Approach:
  - Treat the population as part of a larger **superpopulation**.
  - Assume the superpopulation is a subset of an even larger superpopulation, and so on.
  - This embedding into a sequence of increasingly large superpopulations provides properties like:
    - **Consistency**
    - **Asymptotic Normality**

Hajek's Central Limit Theorem for SRS Without Replacement (1960)

- If certain conditions hold and if n, N, and $N - n$ are all "sufficiently large," then:

$$\frac{\bar{y} - \bar{y}_U}{\sqrt{\left(1 - \frac{n}{N}\right)} \cdot \frac{S}{\sqrt{n}}}$$

  is approximately normal with:
  - Mean 0
  - Variance 1
- A large-sample $100(1 - \alpha)\%$ confidence interval (CI) for the population mean is given by:

$$\left[\bar{y} - z_{\alpha/2}\sqrt{\left(1 - \frac{n}{N}\right)} \cdot \frac{S}{\sqrt{n}}, \bar{y} + z_{\alpha/2}\sqrt{\left(1 - \frac{n}{N}\right)} \cdot \frac{S}{\sqrt{n}}\right]$$

  where:
  - $z_{\alpha/2}$ is the $(1 - \alpha/2)$th percentile of the standard normal distribution.

Sample Size Requirement for Normal Approximation

- **Sugden et al. (2000):** Extend Cochran's rule (Cochran, 1977, p. 42) for determining the sample size required for the normal approximation to be adequate.
- The minimum sample size is given by:

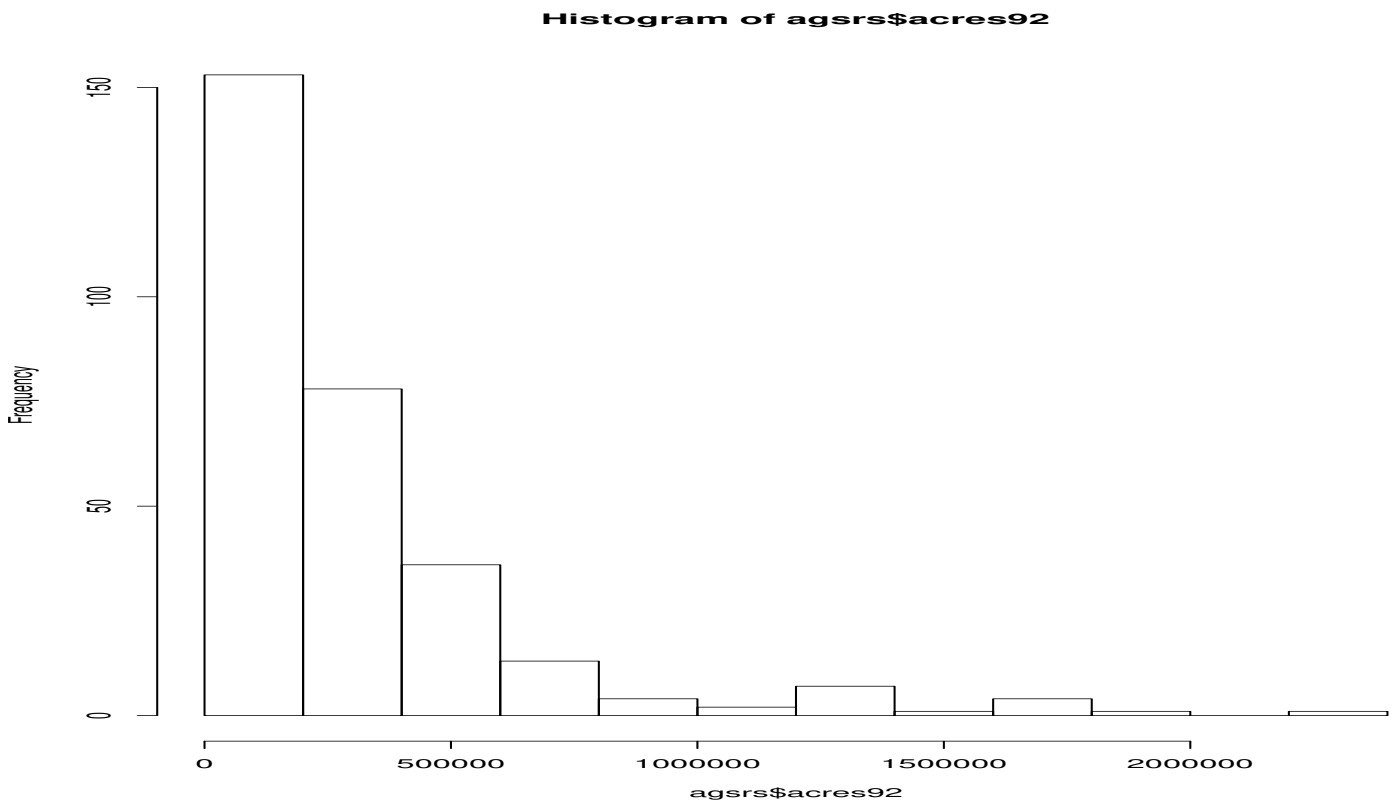
$$n_{\text{min}} = 28 + 25 \left(\frac{\sum_{i=1}^{N} (y_i - \bar{y}_U)^3}{NS^3} \right)^2$$

where:

- $\frac{\sum_{i=1}^{N} (y_i - \bar{y}_U)^3}{NS^3}$ is the **skewness** of the distribution.

Example 2.11

Figure 2: Histogram: number of acres devoted to farms in 1992, for an SRS of 300 counties. Note the skewness of the data. Most of the counties have fewer than 500,000 acres in farms; some counties, however, have more than 1.5 million acres in farms.



$$s = 344,551.9$$

$$\sum_{i=1}^{300}(y_i - \bar{y})^3/n = 1.05036 \times 10^{17}$$

$$n_{\text{min}} = 28 + 25\left(\frac{1.05036 \times 10^{17}}{(344,551.9)^3}\right)^2 \approx 193$$

The sample of size 300 appears to be sufficiently large for the sampling distribution of $\bar{y}$ to be approximately normal.

- In practice, often, substitute $t_{\alpha/2,n-1}$, the $(1-\alpha/2)$th percentile of a t distribution with $n-1$ degrees of freedom, for $z_{\alpha/2}$
  —-large sample, $t_{\alpha/2,n-1} \sim z_{\alpha/2}$
  —-small sample, t produce a wider CI
- Most software use

$$\left[\bar{y} - t_{\alpha/2,n-1}\sqrt{\left(1-\frac{n}{N}\right)}\cdot\frac{s}{\sqrt{n}}, \quad \bar{y} + t_{\alpha/2,n-1}\sqrt{\left(1-\frac{n}{N}\right)}\cdot\frac{s}{\sqrt{n}}\right].$$

Examples 2.7, 2.8, 2.9 (Using Rhandout)

Data Description:

- The U.S government conducts a census of agriculture every five years, collecting data on all farms in the 50 states. The census of agriculture provides data on number of farms, the total acreage devoted to farms, farm size, yield of different crops, and a wide variety of other agriculture measures for $N = 3078$ counties and county-equivalent in the United States.
- agpop.csv is the population data
- agsrs.csv is a simple random sample of size 300 from agpop.csv

Consider

- Response: number of acres devoted to farms in 1992

$$N = 3078, n = 300$$

$$\bar{y} = 297,896$$

$$s = \sqrt{\frac{\sum_{i=1}^{300}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{300}(y_i - 297,896)^2}{299}} = 344,551.9$$

$$\hat{t} = N\bar{y} = 3078 \times 297897 = 916,927,110$$

$$SE(\bar{y}) = \sqrt{\left(1 - \frac{300}{3078}\right)\frac{s^2}{300}} = 18,898.434428$$

$$SE(\hat{t}) = 3078 * SE(\bar{y}) = 3078 * 18,898.434428 = 58,169,381$$

$$\widehat{CV}(\hat{t}) = \widehat{CV}(\bar{y}) = \frac{SE(\bar{y})}{\bar{y}} = \frac{18,898.434428}{297,897} = 0.06344$$

Data is highly skewed, median number of farm acres in a county is $196,717$

An approximate 95% CI for $\bar{y}_U$, using $t_{\alpha/2,299} = 1.968$ is

$$297,896 \pm (1.968)(18,898.434) \quad \text{or} \quad [260706, 335088].$$

An approximate 95% CI for population total t is

$$916,927,110 \pm 1.968(58,169,381) \quad \text{or} \quad [8.02 \times 10^8, 1.03 \times 10^9]$$

Estimated proportion of counties with fewer than $200,000$ acres in farms,

$$\hat{p} = 153/300 = 0.51$$

Recall that

$$\hat{V}(\hat{p}) = \left(1 - \frac{n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}$$

$$SE(\hat{p}) = \sqrt{\left(1 - \frac{300}{3078}\right) \frac{0.51 \cdot 0.49}{299}} = 0.0275$$

For estimating proportions, the usual criterion that the sample size is large enough to use the normal distribution is if both $np \geq 5$ and $n(1-p) \geq 5$.
A 95% CI for the proportion of counties with fewer than 200,000 acres in farms is
$0.51 \pm 1.968(0.0275)$, or $[0.456, 0.564]$.

Sampling Weights

- π_i: **Selection Probability** – the probability that unit i is selected to be in the sample.
- w_i: **Sampling Weight** – defined as $w_i = 1/\pi_i$.
- **In a Simple Random Sample (SRS):**

$$\pi_i = \frac{n}{N}, \quad w_i = \frac{N}{n}$$

- Key Properties:
  - $\sum_{i \in S} w_i = n \cdot \frac{N}{n} = N$
  - $\sum_{i \in S} w_i y_i = \sum_{i \in S} \frac{N}{n} y_i = N \frac{\sum_{i \in S} y_i}{n} = N\bar{y} = \hat{t}$
  - $\frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i} = \frac{\hat{t}}{N} = \bar{y}$
- **Self-Weighting:**
  - In SRS, all weights are equal ($w_i = \frac{N}{n}$).
  - Every unit in the sample represents the same number of units in the population.

Simple Random Sampling with Replacement (SRSWR):

An SRSWR of size n from a population of size N can be thought of as n independent selections of one unit each.

- Select one unit from the population with probability $1/N$.
- After selection, the unit is replaced in the population, and a second unit is randomly selected with the same probability $1/N$.
- The procedure is repeated until the sample contains n units.
- **Note:** In this sampling method, a unit may be selected more than once.

Estimators based on SRSWR:

Let Q_i = the number of time unit i appears in the sample, $i = 1, 2, \cdots, N$

- Estimator of the Population Mean:

$$\bar{y}_r = \frac{1}{n} \sum_{i=1}^{N} Q_i y_i$$

- The Estimator of the Population Total:

$$\hat{t}_{yr} = N \bar{y}_r$$

Properties of $Q_1, Q_2, \cdots, Q_N$

- The joint distribution of $(Q_1, Q_2, \cdots, Q_N)$ is N-variate multinomial with n trials and $p_1 = p_2 = \cdots = p_N = 1/N$

$$f(Q_1, Q_2, \cdots, Q_N) = \frac{n!}{Q_1! Q_2! \cdots Q_N!} p_1^{Q_1} p_2^{Q_2} \cdots p_N^{Q_N}$$

- $0 \leq Q_i \leq n, i = 1, 2, \cdots, N$
- $E(Q_i) = np_i = \frac{n}{N}$
- $V(Q_i) = np_i(1 - p_i) = \frac{n(N-1)}{N^2}$
- $Cov(Q_i, Q_j) = -\frac{n}{N^2}$

Prove: $Cov(Q_i, Q_j) = -\frac{n}{N^2}$

$Q_i \sim Bin(n, p_i)$ and $Q_j \sim Bin(n, p_j)$. So that $Q_i + Q_j \sim Bin(n, p_i + p_j)$. From the variance of the sum:

$$V(Q_i + Q_j) = n(p_i + p_j)(1 - p_i - p_j) \tag{2}$$

Using the variance and covariance relationship:

$$\begin{aligned} V(Q_i + Q_j) &= V(Q_i) + V(Q_j) + 2\,Cov(Q_i, Q_j) \\ &= np_i(1 - p_i) + np_j(1 - p_j) + 2\,Cov(Q_i, Q_j) \end{aligned} \tag{3}$$

Setting Equation (2) equal to Equation (3):

$$Cov(Q_i, Q_j) = -np_ip_j$$

For $p_i = p_j = 1/N$:

$$Cov(Q_i, Q_j) = -\frac{n}{N^2}$$

Note: For fixed n, an increase in one component of a multinomial vector requires a decrease in another component.

Properties of the Estimators in SRSWR:

- $E(\bar{y}_r) = \bar{y}_U$
- $E(\hat{t}_{yr}) = t_y$
- $V(\bar{y}_r) = \frac{N-1}{N} \cdot \frac{S^2}{n}$
- $V(\hat{t}_{yr}) = N(N-1) \cdot \frac{S^2}{n}$

Proof of $E(\bar{y}_r) = \bar{y}_U$:

$$\begin{aligned} E(\bar{y}_r) &= E\left(\frac{1}{n}\sum_{i=1}^{N} Q_i y_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{N} y_i E(Q_i) \\ &= \frac{1}{n} \cdot \frac{n}{N}\sum_{i=1}^{N} y_i \\ &= \bar{y}_U \end{aligned}$$

Comparison of SRS and SRSWR

$$
\begin{aligned}
V(\bar{y}) - V(\bar{y}_r) &= \frac{N-n}{N} \cdot \frac{S^2}{n} - \frac{N-1}{N} \cdot \frac{S^2}{n} \\
&= \frac{S^2}{Nn}\left[N - n - N + 1\right] \\
&= -\frac{n-1}{Nn} \cdot S^2 < 0
\end{aligned}
$$

$$
\mathrm{MSE}(\bar{y}) - \mathrm{MSE}(\bar{y}_r) < 0
$$

- SRS is more efficient than SRSWR.

Sample size estimation

The simplest equation relating the precision and sample size comes from the confidence intervals. Specify the tolerable error

$$p(|\bar{y} - \bar{y}_U| \leq e) = 1 - \alpha$$

where e is called the margin of error.
For an SRSWR, the variance is given by:

$$V(\bar{y}_r) = \frac{N-1}{N} \cdot \frac{S^2}{n}$$

The margin of error e is:

$$e = z_{\alpha/2} \sqrt{\frac{N-1}{N}} \cdot \frac{S}{\sqrt{n}}$$

Thus, the sample size n_0 is approximately:

$$n_0 \approx \left(\frac{z_{\alpha/2} S}{e} \right)^2$$

For an SRSWOR, the margin of error e is given by:

$$e = z_{\alpha/2}\sqrt{\left(1 - \frac{n}{N}\right)} \cdot \frac{S}{\sqrt{n}}$$

The sample size n can be calculated as:

$$n = \frac{z_{\alpha/2}^2 S^2}{e^2 + \frac{z_{\alpha/2}^2 S^2}{N}}$$

or equivalently:

$$n = \frac{\frac{z_{\alpha/2}^2 S^2}{e^2}}{1 + \frac{\frac{z_{\alpha/2}^2 S^2}{e^2}}{N}}$$

which simplifies to:

$$n = \frac{n_0}{1 + \frac{n_0}{N}}$$

Example 2.12 sample size estimation

Suppose we want to estimate the proportion of recipes in the *better homes & gardens New cook book* that do not involve animal products. We plan to take an SRS from the $N = 1251$ kitchen-tested recipes, and want to use a 95% CI with margin of error .03.

$e = .03, \alpha = .05, z_{\alpha/2} = 1.96$

Proportion:

$$S^2 = \frac{N}{N-1} p(1-p) \approx p(1-p),$$

when N is large

S^2 reaches maximum when $p = 1/2$

$$n_0 = \left(\frac{z_{\alpha/2} S}{e}\right)^2 = \frac{1.96^2}{4 \times .03^2} \approx 1067$$

will result in a 95% CI with width at most $2e$

Note: The sample size ignoring the fpc is large compared with the population size, so in this case we would make the fpc adjustment and use

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{1067}{1 + \frac{1067}{1251}} = 576$$

Example 2.12 Sample size estimation

Suppose we want to estimate the proportion of recipes in the *Better Homes & Gardens New Cookbook* that do not involve animal products. We plan to take an SRS from the $N = 1251$ kitchen-tested recipes, and want to use a 95% confidence interval (CI) with a margin of error of 0.03, i.e.,

$$e = 0.03, \quad \alpha = 0.05, \quad z_{\alpha/2} = 1.96$$

For the proportion, the variance is:

$$S^2 = \frac{N}{N-1} p(1-p) \approx p(1-p), \quad \text{when } N \text{ is large}$$

The variance S^2 reaches its maximum when $p = \frac{1}{2}$.
Thus, the sample size without considering the finite population correction (fpc) is:

$$n_0 = \left(\frac{z_{\alpha/2} S}{e} \right)^2 = \frac{1.96^2}{4 \times 0.03^2} \approx 1067$$

This will result in a 95% CI with a width at most $2e$.

Note: The sample size ignoring the finite population correction (fpc) is large compared with the population size. Therefore, we apply the fpc adjustment and use the following formula:

$$n = \frac{n_0}{1 + \frac{n_0}{N}} = \frac{1067}{1 + \frac{1067}{1251}} = 576$$

Thus, the adjusted sample size is 576.

Estimate S^2:

- When estimating a proportion, use $\frac{1}{4}$ as an upper bound for S^2. For other quantities, S^2 must be estimated or approximated.
- Pilot sample: A small sample taken to provide information and guidance for the design of the main survey can be used to estimate quantities needed for setting the sample size.
- Use previous studies or data available in the literature for estimates.
- Alternatively, estimate or guess the variance based on the context.

- Use the Coefficient of Variation (CV) to estimate the standard error (SE):
  - The CV is computed as the standard error (SE) divided by the estimate.
  - Specifying a CV instead of a target margin of error accounts for the fact that the total estimate and the standard error are likely to change together. Thus, different accounting methodologies that produce different estimates and standard errors can be expected to have similar CVs.

$$CV(\hat{t}) = \frac{\sqrt{V(\hat{t})}}{E(\hat{t})}$$

$$\widehat{CV}(\hat{t}) = \frac{SE(\hat{t})}{\hat{t}} = N\sqrt{1 - \frac{n}{N}} \cdot \frac{s}{\sqrt{n} \cdot \hat{t}}$$

Example: Suppose the total revenue estimate is \$1 billion. A 10% CV means the standard error (SE) is \$100 million. At a 95% confidence level, the margin of error is \$196 million.

- Adjust for budget considerations when estimating sample size and error.

A Prediction Approach for SRS

- The randomization theory (also called design-based theory) provides one framework for sampling methods.
- An alternative is model-based theory.
  Model (*): Assume $y_1, y_2, \ldots, y_N$ are independent with

$$E_M[y_i] = u \quad \text{and} \quad V_M[y_i] = \sigma^2$$

$$t = \sum_{i=1}^{N} y_i = \sum_{i \in S} y_i + \sum_{i \notin S} y_i$$

$$\begin{aligned} \hat{t} &= \sum_{i \in S} y_i + \sum_{i \notin S} \hat{y}_i \\ &= \sum_{i \in S} y_i + \frac{N-n}{n} \sum_{i \in S} y_i \\ &= \frac{N}{n} \sum_{i \in S} y_i \end{aligned}$$

$$
\begin{aligned}
E_M[\hat{t} - t] &= \frac{N}{n} \sum_{i \in S} E_M(y_i) - \sum_{i=1}^{N} E_M[y_i] \\
&= \frac{N}{n} nu - Nu = 0
\end{aligned}
$$

Thus, $\hat{t}$ is model-unbiased for t. Also, we can show that

$$
V(\hat{t}) = N^2 \left(1 - \frac{n}{N}\right) \frac{\sigma^2}{n}
$$

$$
\begin{aligned}
V(\hat{t}) &= E_M[(\hat{t}-t)^2] \\
&= E_M\left[\left(\frac{N}{n}\sum_{i\in S} y_i - \sum_{i=1}^{N} y_i\right)^2\right] \\
&= E_M\left[\left(\frac{N}{n}-1\right)\sum_{i\in S} y_i - \sum_{i\notin S} y_i\right]^2 \\
&= E_M\left[\left(\frac{N}{n}-1\right)\sum_{i\in S} y_i - \sum_{i\notin S} y_i - \left(\frac{N}{n}-1\right)nu + (N-n)u\right]^2 \\
&= E_M\left[\left(\frac{N}{n}-1\right)^2\left(\sum_{i\in S} y_i - nu\right)^2 + \left(\sum_{i\notin S} y_i - (N-n)u\right)^2\right] \\
&= \left(\frac{N}{n}-1\right)^2 n\sigma^2 + (N-n)\sigma^2 \\
&= N^2\left(1-\frac{n}{N}\right)\frac{\sigma^2}{n}
\end{aligned}
$$

Note: Under model (*), the randomization and model-based approaches give:

- The same estimator of the population total.
- The same variance estimator.
- The same confidence interval.

Design-based approach v.s model-based approach

Design-Based Approach

- Data: $\{y_i, i = 1, 2, \cdots, N\}$ (fixed values)
- Random variables: Z_i (1 if selected, 0 otherwise)
- Inference: Based on repeated sampling from a finite population
- No assumptions on distribution of y_is (nonparametric)

Model-Based Approach (using model $(*)$)

- Data: $\{y_i, i = 1, 2, \cdots, N\}$ (independent random variables)
- Inference: Based on probabilities using CLT approximation
- Assumptions: y_is are independent with the same mean (u) and variance (σ^2)

Chapter 2 Summary: Simple Random Sampling

- $\pi_i = P(i \in S) = \dfrac{n}{N}$
- $w_i = \dfrac{1}{\pi_i} = \dfrac{N}{n}$
- $\bar{y} = \dfrac{1}{n}\sum_{i \in S} y_i$
- $s^2 = \dfrac{1}{n-1}\sum_{i \in S}(y_i - \bar{y})^2$

Estimators and Standard Errors

Population Quantity	Estimator	Standard Error
$t = \sum_{i=1}^{N} y_i$	$\hat{t} = \sum_{i \in S} w_i y_i = N\bar{y}$	$N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$
$\bar{y}_U = \frac{t}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i$	$\bar{y} = \frac{\hat{t}}{N} = \frac{\sum_{i \in S} w_i y_i}{\sum_{i \in S} w_i}$	$\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$
p	$\hat{p}$	$\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1}}$

Most software use the following formula for confidence intervals:

$$\left[\bar{y} - t_{\alpha/2,n-1}\sqrt{\left(1 - \frac{n}{N}\right)\frac{s}{\sqrt{n}}}, \quad \bar{y} + t_{\alpha/2,n-1}\sqrt{\left(1 - \frac{n}{N}\right)\frac{s}{\sqrt{n}}}\right]$$

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.