2.3 Simple Random Sampling

- **Simple random sampling without replacement (srswor)** of size $n$ is the probability sampling design for which a fixed number of $n$ units are selected from a population of $N$ units without replacement such that every possible sample of $n$ units has equal probability of being selected. A resulting sample is called a simple random sample or srs.

- Note: I will use SRS to denote a simple random sample and SR as an abbreviation of ‘simple random’.

- Some necessary combinatorial notation:
  
  - $(n \text{ factorial})$ $n! = n \times (n - 1) \times (n - 2) \times \cdots \times 2 \times 1$. This is the number of unique arrangements or orderings (or permutations) of $n$ distinct items. For example: $6! = 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$.
  
  - $(N \text{ choose } n)$ $\binom{N}{n} = \frac{N(N - 1) \cdots (N - n + 1)}{n!}$. This is the number of combinations of $n$ items selected from $N$ distinct items (and the order of selection doesn’t matter). For example, $\binom{6}{2} = \frac{6!}{2!4!} = \frac{(6)(5)(4!)}{2!4!} = \frac{(6)(5)}{(2)(1)} = 15$.

- There are $\binom{N}{n}$ possible SRSs of size $n$ selected from a population of size $N$.

- For any SRS of size $n$ from a population of size $N$, we have $P(S) = 1/\binom{N}{n}$.

- Unless otherwise specified, we will assume sampling is without replacement.

2.3.1 Estimation of $\overline{y}_U$ and $t$

- A natural estimator for the population mean $\overline{y}_U$ is the sample mean $\overline{y}$. Because $\overline{y}$ is an estimate of an individual unit’s $y$-value, multiplication by the population size $N$ will give us an estimate $\hat{t}$ of the population total $t$. That is:

  $$\hat{y}_U = \overline{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \quad \text{and} \quad \hat{t} = \frac{N}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} N y_i =$$ (10)

- $\hat{y}_U$ and $\hat{t}$ are design unbiased. That is, the average values of $\overline{y}$ and $N\overline{y}$ taken over all possible SRSs equal $\overline{y}_U$ and $t$, respectively.

**Demonstration of Unbiasedness:** Suppose we have a population consisting of five $y$-values:

<table>
<thead>
<tr>
<th>Unit $i$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_i$</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

which has the following parameters:

$$N = \quad t = \quad \overline{y}_U = \quad S^2 = \quad S \approx$$

Suppose a SRS of size $n = 2$ is selected. Then $P(S) = 1/\binom{5}{2} = 1/10$ for each of the 10 possible SRSs.
All Possible Samples and Statistics from Example Population

<table>
<thead>
<tr>
<th>Sample</th>
<th>Units</th>
<th>y-values</th>
<th>$\sum y_i$</th>
<th>$\bar{y}_U = \bar{y}$</th>
<th>$t = N\bar{y}$</th>
<th>$S^2 = s^2$</th>
<th>$S = s$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>1,2</td>
<td>0,2</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1.4142</td>
</tr>
<tr>
<td>$S_2$</td>
<td>1,3</td>
<td>0,3</td>
<td>3</td>
<td>1.5</td>
<td>7.5</td>
<td>4.5</td>
<td>2.1213</td>
</tr>
<tr>
<td>$S_3$</td>
<td>1,4</td>
<td>0,4</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>2.8284</td>
</tr>
<tr>
<td>$S_4$</td>
<td>1,5</td>
<td>0,7</td>
<td>7</td>
<td>3.5</td>
<td>17.5</td>
<td>24.5</td>
<td>4.9497</td>
</tr>
<tr>
<td>$S_5$</td>
<td>2,3</td>
<td>2,3</td>
<td>5</td>
<td>2.5</td>
<td>12.5</td>
<td>.5</td>
<td>.7071</td>
</tr>
<tr>
<td>$S_6$</td>
<td>2,4</td>
<td>2,4</td>
<td>6</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td>1.4142</td>
</tr>
<tr>
<td>$S_7$</td>
<td>2,5</td>
<td>2,7</td>
<td>9</td>
<td>4.5</td>
<td>22.5</td>
<td>12.5</td>
<td>3.5355</td>
</tr>
<tr>
<td>$S_8$</td>
<td>3,4</td>
<td>3,4</td>
<td>7</td>
<td>3.5</td>
<td>17.5</td>
<td>.5</td>
<td>.7071</td>
</tr>
<tr>
<td>$S_9$</td>
<td>3,5</td>
<td>3,7</td>
<td>10</td>
<td>5</td>
<td>25</td>
<td>8</td>
<td>2.8284</td>
</tr>
<tr>
<td>$S_{10}$</td>
<td>4,5</td>
<td>4,7</td>
<td>11</td>
<td>5.5</td>
<td>27.5</td>
<td>4.5</td>
<td>2.1213</td>
</tr>
</tbody>
</table>

| Column Sum | 32 | 160 | 67 | 22.6274 |
| Expected value | $= E(\text{estimator})$ | $\frac{32}{10} = 3.2$ | $\frac{160}{10} = 16$ | $\frac{67}{10} = 6.7$ | $\frac{22.6274}{10} = 2.26274$ |

The averages for estimators $\bar{y}_U = \bar{y}$, $\bar{t} = N\bar{y}$, and $\bar{S}^2 = s^2$ equal the parameters that they are estimating. This implies that $\bar{y}$, $N\bar{y}$, and $s^2$ are unbiased estimators of $\bar{y}_U$, $t$, and $S^2$.

Notation: $E(\bar{y}_U) = \bar{y}$, $E(\bar{t}) = t$, $E(\bar{S}^2) = S^2$ or $E(\bar{y}) = \bar{y}_U$, $E(N\bar{y}) = t$, $E(s^2) = S^2$.

The average for estimator $\bar{S} = s$ does not equal the parameter $S$. This implies that $s$ is a biased estimator of $S$. Notation: $E(\bar{S}) \neq S$ or $E(s) \neq S$.

- The next problem is to study the variances of $\bar{y}_U = \bar{y}$ and $\bar{t} = N\bar{y}$.
- Warning: In an introductory statistics course, you were told that the variance of the sample mean $V(Y) = S^2/n$ ($= \sigma^2/n$) and its standard deviation is $S/\sqrt{n}$ ($= \sigma/\sqrt{n}$). This is appropriate if a sample was to be taken from an infinite or extremely large population.
- However, we are dealing with finite populations that often are not considered extremely large. In such cases, we have to adjust our variance formulas by $\frac{N-n}{N}$ which is known as the finite population correction (f.p.c.).
- Texts may rewrite the f.p.c. $\frac{N-n}{N}$ as either $1 - \frac{n}{N}$ or $1 - f$ where $f = n/N$ is the fraction of the population that was sampled. By definition:

$$V(\bar{y}_U) = V(\bar{y}) = V(\bar{t}) = N^2V(\bar{y}) = N(N-n)\frac{S^2}{n} \quad (11)$$

- Because $S^2$ is unknown, we use $s^2$ to get unbiased estimators of the variances in (11):

$$\hat{V}(\bar{y}_U) = \hat{V}(\bar{y}) = \hat{V}(\bar{t}) = N^2\hat{V}(\bar{y}) = N(N-n)\frac{s^2}{n} \quad (12)$$

- Taking a square root of a variance in (11) yields the standard deviation of the estimator.
- Taking a square root of an estimated variance in (12) yields the standard error of the estimate.
• Thus, \( V(\bar{y}) = \left( \frac{N - n}{N} \right) \frac{S^2}{n} = \frac{3.67}{5} = \) and

\[
V(\hat{t}) = N^2 V(\bar{y}) = N(N - n) \frac{S^2}{n} = (5)(3) \frac{6.7}{2} = .
\]

• Like \( \widehat{\bar{y}} \) and \( \hat{t} \), the variances \( \widehat{V}(\widehat{\bar{y}}) \) and \( \widehat{V}(\hat{t}) \) are design unbiased. That is the average of \( \widehat{V}(\widehat{\bar{y}}) \) and \( \widehat{V}(\hat{t}) \) taken over all possible SRSs equal \( V(\widehat{\bar{y}}) = 2.01 \) and \( V(\hat{t}) = 50.25 \), respectively.

• For the estimated variances we have \( \widehat{V}(\widehat{\bar{y}}) = \left( \frac{N - n}{N} \right) \frac{s^2}{n} = \frac{3}{5} \frac{s^2}{2} = \) and

\[
\widehat{V}(\hat{t}) = N(N - n) \frac{s^2}{n} = (5)(3) \frac{s^2}{2} = \text{where } s^2 \text{ is a particular sample variance.}
\]

Example: We will use our population from the previous example:

<table>
<thead>
<tr>
<th>Unit, ( i )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \bar{y}_i )</td>
<td>0</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>7</td>
</tr>
</tbody>
</table>

which have the following parameters

\[ N = 5 \quad t = 16 \quad \bar{y}_U = 3.2 \quad S^2 = 6.7 \quad S \approx 2.588 \]

**Estimated Variances of \( \widehat{\bar{y}}_U \) and \( \hat{t} \) for All Samples**

<table>
<thead>
<tr>
<th>Sample</th>
<th>Units</th>
<th>( y )-values</th>
<th>( s^2 )</th>
<th>( \widehat{V}(\widehat{\bar{y}}_U) = .3s^2 )</th>
<th>( V(\hat{t}) = 7.5s^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_1 )</td>
<td>1,2</td>
<td>0.2</td>
<td>2</td>
<td>0.6</td>
<td>15</td>
</tr>
<tr>
<td>( S_2 )</td>
<td>1,3</td>
<td>0.3</td>
<td>4.5</td>
<td>1.35</td>
<td>33.75</td>
</tr>
<tr>
<td>( S_3 )</td>
<td>1,4</td>
<td>0.4</td>
<td>8</td>
<td>2.4</td>
<td>60</td>
</tr>
<tr>
<td>( S_4 )</td>
<td>1,5</td>
<td>0.7</td>
<td>24.5</td>
<td>7.35</td>
<td>183.75</td>
</tr>
<tr>
<td>( S_5 )</td>
<td>2,3</td>
<td>2.3</td>
<td>.5</td>
<td>0.15</td>
<td>3.75</td>
</tr>
<tr>
<td>( S_6 )</td>
<td>2,4</td>
<td>2.4</td>
<td>2</td>
<td>0.6</td>
<td>15</td>
</tr>
<tr>
<td>( S_7 )</td>
<td>2,5</td>
<td>2.7</td>
<td>12.5</td>
<td>3.75</td>
<td>93.75</td>
</tr>
<tr>
<td>( S_8 )</td>
<td>3,4</td>
<td>3.4</td>
<td>.5</td>
<td>0.15</td>
<td>3.75</td>
</tr>
<tr>
<td>( S_9 )</td>
<td>3,5</td>
<td>3.7</td>
<td>8</td>
<td>2.4</td>
<td>60</td>
</tr>
<tr>
<td>( S_{10} )</td>
<td>4,5</td>
<td>4.7</td>
<td>4.5</td>
<td>1.35</td>
<td>33.75</td>
</tr>
</tbody>
</table>

**Column Sum**

- From the table we have \( E(\widehat{V}(\widehat{\bar{y}}_U)) = 20.1/10 = 2.01 = V(\widehat{\bar{y}}_U) \) and \( E(\widehat{V}(\hat{t})) = 502.5/10 = 50.25 = V(\hat{t}) \). Thus, we see that both variance estimators are unbiased.

- If \( N \) is large relative to \( n \), then the finite population correction (f.p.c.) will be close to (but less than) 1. Omitting the finite population correction from the variance formulas (i.e., replacing \( (N - n)/N \) with 1) will slightly overestimate the true variance. That is, there is a small positive bias. I personally would not recommend omitting the f.p.c.

- If \( N \) is not large relative to \( n \), then omitting the f.p.c. from the variance formulas can seriously overestimate the true variance. That is, there can be a large positive bias.

- As \( n \to N, \frac{N - n}{N} \to 0 \). That is, as the sample size approaches the population size, the f.p.c. approaches 0. Thus, in (11) and (12) the variances \( \to 0 \) as \( n \to N \).
2.3.2 SRS With Replacement

- Consider a sampling procedure in which a sampling unit is randomly selected from the population, its $y$-value recorded, and is then returned to the population. This process of randomly selecting units with replacement after each stage is repeated $n$ times. Thus, a sampling unit may be sampled multiple times. A sample of $n$ units selected by such a procedure is called a **simple random sample with replacement**.

- The estimators for SRS with replacement are: $\hat{\mu} = \bar{y}$, $\hat{V}(\mu) = \hat{V}(\bar{y}) = \frac{s^2}{n}$

- Suppose we have two estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ of some parameter $\theta$.
  - $\hat{\theta}_1$ is **less efficient** than $\hat{\theta}_2$ for estimating $\theta$ if $V(\hat{\theta}_1) > V(\hat{\theta}_2)$.
  - $\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ for estimating $\theta$ if $V(\hat{\theta}_1) < V(\hat{\theta}_2)$.

- For most situations, the estimator for a SRS with replacement is **less efficient** than the estimator for a SRS without replacement.

- There will be circumstances (such as sampling proportional to size) where we will consider sampling with replacement. Unless otherwise stated, we assume that sampling is done without replacement.

2.4 Two-Sided Confidence Intervals for $\bar{y}_U$ and $t$

- In an introductory statistics course, you were given confidence interval formulas
  $$\bar{y} \pm z^* \frac{s}{\sqrt{n}}$$
  $$\bar{y} \pm t^* \frac{s}{\sqrt{n}}$$

These formulas are applicable if a sample was to be taken from an infinitely or extremely large population. But when we are dealing with finite populations, we adjust our variance formulas by the finite population correction.

- In the finite population version of the Central Limit Theorem, we assume the estimators $\hat{\mu} = \bar{y}$ and $\hat{t} = N\bar{y}$ have sampling distributions that are approximately normal. That is,
  $$\hat{\mu} \sim N\left( \mu, \frac{N-n}{N} \frac{s^2}{n} \right)$$
  $$\hat{t} \sim N\left( t, N(N-n) \frac{s^2}{n} \right)$$

- For large samples, approximate $100(1 - \alpha)\%$ confidence intervals for $\bar{y}_U$ ($\mu$) and $t$ ($\tau$) are

  For $\bar{y}_U$:
  $$\bar{y} \pm z^* \sqrt{\frac{(N-n)}{N}} \frac{s^2}{n}$$
  $$\bar{y} \pm z^* \frac{s}{\sqrt{n}}$$

  For $t$:
  $$N\bar{y} \pm z^* \frac{s}{\sqrt{n}}$$
  $$N\bar{y} \pm \frac{s}{\sqrt{n}}$$

  where $z^*$ is the upper $\alpha/2$ critical value from the standard normal distribution. Or, in standard error (s.e.) notation,
  $$\hat{\bar{y}}_U \pm \hat{t} \pm$$

For 90%, 95%, and 99%, $z^* = 1.645, 1.96, and 2.576$, respectively.
• For smaller samples, approximate 100(1 - α)% confidence intervals for \( \overline{y}_U \) and \( t \) are

For \( \overline{y}_U \) :

\[
\overline{y} \pm t^* \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}}
\]

For \( t \) :

\[
N\overline{y} \pm t^* \frac{s \sqrt{N(N-n)/n}}{n}
\]

(16)

where \( t^* \) is the upper \( \alpha/2 \) critical value from the \( t(n-1) \) distribution.

• The quantity being added and subtracted from \( \overline{y}_U = \overline{y} \) or \( \overline{t} = N\overline{y} \) in the confidence interval is known as the margin of error.

Example: Use the small population data again. For \( n = 2 \), \( t^* \approx 6.314 \) for a nominal 90% confidence level.

<table>
<thead>
<tr>
<th>Sample</th>
<th>( y )-values</th>
<th>( \sum y_i )</th>
<th>( \overline{y} )</th>
<th>( \overline{t} = N \overline{y} )</th>
<th>( S^2 = s^2 )</th>
<th>( S = s )</th>
<th>( V(\overline{y}_U) )</th>
<th>( V(t) )</th>
<th>90% ci for ( t )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1.4142</td>
<td>0.6</td>
<td>15</td>
<td>(-19.45, 29.45)</td>
</tr>
<tr>
<td>2</td>
<td>0.3</td>
<td>3</td>
<td>1.5</td>
<td>7.5</td>
<td>4.5</td>
<td>2.1213</td>
<td>1.35</td>
<td>33.75</td>
<td>(-29.18, 44.18)</td>
</tr>
<tr>
<td>3</td>
<td>0.4</td>
<td>4</td>
<td>2</td>
<td>10</td>
<td>8</td>
<td>2.8284</td>
<td>2.4</td>
<td>60</td>
<td>(-38.91, 58.91)</td>
</tr>
<tr>
<td>4</td>
<td>0.7</td>
<td>7</td>
<td>3.5</td>
<td>17.5</td>
<td>24.5</td>
<td>4.9497</td>
<td>7.35</td>
<td>183.75</td>
<td>(-68.09, 103.09)</td>
</tr>
<tr>
<td>5</td>
<td>2.3</td>
<td>5</td>
<td>2.5</td>
<td>12.5</td>
<td>.5</td>
<td>0.7071</td>
<td>0.15</td>
<td>3.75</td>
<td>(0.27, 24.73)</td>
</tr>
<tr>
<td>6</td>
<td>2.4</td>
<td>6</td>
<td>3</td>
<td>15</td>
<td>2</td>
<td>1.4142</td>
<td>0.6</td>
<td>15</td>
<td>(-9.45, 39.45)</td>
</tr>
<tr>
<td>7</td>
<td>2.7</td>
<td>9</td>
<td>4.5</td>
<td>22.5</td>
<td>12.5</td>
<td>3.5355</td>
<td>3.75</td>
<td>93.75</td>
<td>(-38.63, 83.63)</td>
</tr>
<tr>
<td>8</td>
<td>3.4</td>
<td>7</td>
<td>3.5</td>
<td>17.5</td>
<td>.5</td>
<td>0.7071</td>
<td>0.15</td>
<td>3.75</td>
<td>(5.27, 29.73)</td>
</tr>
<tr>
<td>9</td>
<td>3.7</td>
<td>10</td>
<td>5</td>
<td>25</td>
<td>8</td>
<td>2.8284</td>
<td>2.4</td>
<td>60</td>
<td>(-23.91, 73.91)</td>
</tr>
<tr>
<td>10</td>
<td>4.7</td>
<td>11</td>
<td>5.5</td>
<td>27.5</td>
<td>4.5</td>
<td>2.1213</td>
<td>1.35</td>
<td>33.75</td>
<td>(-9.18, 64.18)</td>
</tr>
</tbody>
</table>

2.4.1 One-Sided Confidence Intervals for \( \overline{y}_U \) and \( t \)

• Occasionally, a researcher may want a one-sided confidence interval. There are two types of one-sided confidence intervals: upper and lower.

• Approximate upper and lower 100(1 - α)% confidence intervals for \( \overline{y}_U \) and \( t \) are:

For \( \overline{y}_U \) :

\[
\left( \overline{y} - t^* s \sqrt{\left(\frac{N-n}{N}\right) /n} , \infty \right)
\]

For \( t \) :

\[
\left( -\infty , \overline{y} + t^* s \sqrt{N(N-n)/n} \right)
\]

(upper)

\[
\left( N\overline{y} - t^* s \sqrt{N(N-n)/n} , \infty \right)
\]

(lower)

where \( t^* \) is the upper \( \alpha \) critical value from the \( t(n-1) \) distribution.

• If the \( y \)-values cannot be negative, replace \( -\infty \) with 0 in the lower confidence interval formulas. If the \( y \)-values cannot be positive, replace \( \infty \) with 0 in the upper confidence interval formulas.

26
• Later, we will discuss another method of generating a confidence interval called **bootstrapping**. This will be useful when the sample size may be small and the central limit theorem cannot be applied.

**SRS Example with Strong Spatial Correlation**

• To illustrate the application of simple random sampling to population mean per unit \( \mu \) estimation, consider the abundance data in Figure 1. The abundance counts are artificial but show a strong diagonal spatial correlation.

• The region has been gridded into a 20 \( \times \) 20 grid of 10 \( \times \) 10 m quadrats. The total abundance \( t = 13354 \) and the mean per unit is \( \bar{y}_U = 33.385 \). The population variance \( S^2 = 75.601 \).

• This data will be used to compare estimation properties of various sampling designs when data are spatially correlated.

**Figure 1**

Data Exhibiting Strong Spatial Correlation

<table>
<thead>
<tr>
<th>18</th>
<th>20</th>
<th>15</th>
<th>20</th>
<th>15</th>
<th>19</th>
<th>18</th>
<th>24</th>
<th>23</th>
<th>20</th>
<th>26</th>
<th>29</th>
<th>28</th>
<th>28</th>
<th>31</th>
<th>31</th>
<th>34</th>
<th>28</th>
<th>32</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>20</td>
<td>16</td>
<td>20</td>
<td>15</td>
<td>23</td>
<td>19</td>
<td>26</td>
<td>21</td>
<td>21</td>
<td>24</td>
<td>30</td>
<td>23</td>
<td>26</td>
<td>25</td>
<td>33</td>
<td>31</td>
<td>28</td>
<td>32</td>
<td>38</td>
</tr>
<tr>
<td>16</td>
<td>18</td>
<td>20</td>
<td>24</td>
<td>25</td>
<td>26</td>
<td>22</td>
<td>23</td>
<td>26</td>
<td>26</td>
<td>22</td>
<td>27</td>
<td>25</td>
<td>25</td>
<td>34</td>
<td>28</td>
<td>37</td>
<td>36</td>
<td>38</td>
<td>31</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>16</td>
<td>22</td>
<td>21</td>
<td>23</td>
<td>22</td>
<td>27</td>
<td>27</td>
<td>24</td>
<td>28</td>
<td>32</td>
<td>29</td>
<td>33</td>
<td>27</td>
<td>37</td>
<td>37</td>
<td>38</td>
<td>35</td>
<td>33</td>
</tr>
<tr>
<td>15</td>
<td>19</td>
<td>23</td>
<td>17</td>
<td>21</td>
<td>23</td>
<td>21</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>31</td>
<td>26</td>
<td>32</td>
<td>34</td>
<td>32</td>
<td>33</td>
<td>31</td>
<td>31</td>
<td>36</td>
<td>37</td>
</tr>
<tr>
<td>21</td>
<td>24</td>
<td>20</td>
<td>21</td>
<td>28</td>
<td>26</td>
<td>30</td>
<td>22</td>
<td>31</td>
<td>25</td>
<td>29</td>
<td>29</td>
<td>27</td>
<td>30</td>
<td>29</td>
<td>37</td>
<td>35</td>
<td>32</td>
<td>38</td>
<td>43</td>
</tr>
<tr>
<td>23</td>
<td>17</td>
<td>24</td>
<td>25</td>
<td>24</td>
<td>27</td>
<td>31</td>
<td>29</td>
<td>31</td>
<td>34</td>
<td>27</td>
<td>36</td>
<td>29</td>
<td>29</td>
<td>34</td>
<td>39</td>
<td>37</td>
<td>37</td>
<td>40</td>
<td>36</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
<td>21</td>
<td>25</td>
<td>27</td>
<td>22</td>
<td>32</td>
<td>32</td>
<td>31</td>
<td>26</td>
<td>28</td>
<td>34</td>
<td>34</td>
<td>37</td>
<td>35</td>
<td>34</td>
<td>38</td>
<td>38</td>
<td>37</td>
<td>40</td>
</tr>
<tr>
<td>22</td>
<td>26</td>
<td>28</td>
<td>26</td>
<td>24</td>
<td>29</td>
<td>33</td>
<td>26</td>
<td>27</td>
<td>27</td>
<td>34</td>
<td>31</td>
<td>39</td>
<td>32</td>
<td>36</td>
<td>38</td>
<td>37</td>
<td>40</td>
<td>44</td>
<td>43</td>
</tr>
<tr>
<td>23</td>
<td>27</td>
<td>28</td>
<td>29</td>
<td>26</td>
<td>32</td>
<td>25</td>
<td>31</td>
<td>35</td>
<td>34</td>
<td>32</td>
<td>33</td>
<td>37</td>
<td>32</td>
<td>42</td>
<td>40</td>
<td>40</td>
<td>37</td>
<td>42</td>
<td>44</td>
</tr>
<tr>
<td>23</td>
<td>21</td>
<td>31</td>
<td>23</td>
<td>30</td>
<td>27</td>
<td>31</td>
<td>30</td>
<td>32</td>
<td>35</td>
<td>30</td>
<td>40</td>
<td>32</td>
<td>37</td>
<td>37</td>
<td>36</td>
<td>40</td>
<td>44</td>
<td>44</td>
<td>40</td>
</tr>
<tr>
<td>26</td>
<td>29</td>
<td>31</td>
<td>26</td>
<td>30</td>
<td>31</td>
<td>34</td>
<td>36</td>
<td>30</td>
<td>38</td>
<td>36</td>
<td>32</td>
<td>38</td>
<td>38</td>
<td>37</td>
<td>42</td>
<td>42</td>
<td>41</td>
<td>40</td>
<td>49</td>
</tr>
<tr>
<td>28</td>
<td>24</td>
<td>28</td>
<td>27</td>
<td>26</td>
<td>31</td>
<td>32</td>
<td>29</td>
<td>32</td>
<td>33</td>
<td>38</td>
<td>34</td>
<td>39</td>
<td>38</td>
<td>40</td>
<td>37</td>
<td>41</td>
<td>43</td>
<td>42</td>
<td>43</td>
</tr>
<tr>
<td>32</td>
<td>25</td>
<td>31</td>
<td>32</td>
<td>29</td>
<td>29</td>
<td>35</td>
<td>38</td>
<td>38</td>
<td>32</td>
<td>36</td>
<td>35</td>
<td>39</td>
<td>42</td>
<td>39</td>
<td>40</td>
<td>44</td>
<td>42</td>
<td>41</td>
<td>45</td>
</tr>
<tr>
<td>27</td>
<td>29</td>
<td>35</td>
<td>28</td>
<td>35</td>
<td>35</td>
<td>31</td>
<td>40</td>
<td>35</td>
<td>37</td>
<td>38</td>
<td>44</td>
<td>40</td>
<td>40</td>
<td>47</td>
<td>39</td>
<td>49</td>
<td>48</td>
<td>51</td>
<td>49</td>
</tr>
<tr>
<td>30</td>
<td>29</td>
<td>32</td>
<td>32</td>
<td>33</td>
<td>30</td>
<td>36</td>
<td>38</td>
<td>42</td>
<td>36</td>
<td>35</td>
<td>38</td>
<td>44</td>
<td>47</td>
<td>45</td>
<td>49</td>
<td>41</td>
<td>43</td>
<td>44</td>
<td>51</td>
</tr>
<tr>
<td>28</td>
<td>35</td>
<td>35</td>
<td>34</td>
<td>34</td>
<td>33</td>
<td>41</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>39</td>
<td>44</td>
<td>44</td>
<td>48</td>
<td>44</td>
<td>50</td>
<td>49</td>
<td>48</td>
<td>53</td>
<td>54</td>
</tr>
<tr>
<td>29</td>
<td>33</td>
<td>32</td>
<td>36</td>
<td>39</td>
<td>33</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>42</td>
<td>46</td>
<td>47</td>
<td>48</td>
<td>47</td>
<td>46</td>
<td>45</td>
<td>44</td>
<td>52</td>
<td>54</td>
<td>55</td>
</tr>
<tr>
<td>28</td>
<td>37</td>
<td>38</td>
<td>37</td>
<td>33</td>
<td>33</td>
<td>34</td>
<td>37</td>
<td>45</td>
<td>40</td>
<td>39</td>
<td>42</td>
<td>42</td>
<td>46</td>
<td>47</td>
<td>48</td>
<td>52</td>
<td>47</td>
<td>46</td>
<td>53</td>
</tr>
<tr>
<td>38</td>
<td>39</td>
<td>39</td>
<td>37</td>
<td>34</td>
<td>38</td>
<td>39</td>
<td>45</td>
<td>39</td>
<td>42</td>
<td>45</td>
<td>41</td>
<td>44</td>
<td>51</td>
<td>46</td>
<td>50</td>
<td>52</td>
<td>51</td>
<td>51</td>
<td>53</td>
</tr>
</tbody>
</table>
SRS taken from Figure 1 \((n = 10, t = 13354, \mu_U = 33.385, \bar{y} = 34.1, s^2 = 18.32)\)
SRS Example using Rathbun and Cressie (1994) Data

• To illustrate the application of simple random sampling to population total $t$ estimation, consider the abundance data in Figure 2. The abundance counts correspond to the census data studied by Rathbun and Cressie (1994).

• This $200 \times 200$ m study region is located in an old-growth forest in Thomas County, Georgia. This data represents the number of longleaf pine trees located in each quadrat. The coordinates of the 584 tree locations are given in Cressie (1991).

• I have gridded the region into a $20 \times 20$ grid of $10 \times 10$ m quadrats. The total abundance $t = 584$ and the mean abundance per quadrat $\bar{y}_U = 584/400 = 1.435$. The population variance $S^2 = 3.853$.

• There is only a weak spatial correlation of tree counts within the study region.

• The pineleaf census data will be used to compare estimation properties of various sampling designs.

• Note the two relatively large boldfaced values (14 and 16).

| 1 | 1 | 1 | 1 | 1 | 2 | 1 | 0 | 0 | 0 | 4 | 5 | 0 | 1 | 0 | 1 | 2 | 1 | 0 | 1 |
| 3 | 2 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 2 | 2 | 0 | 2 | 2 | 0 | 2 | 0 | 0 | 1 |
| 7 | 4 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 2 | 2 | 0 | 4 | 3 | 2 | 4 | 2 | 1 | 2 | 2 |
| 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 | 4 | 5 | 1 | 5 | 0 | 0 | 0 | 2 | 1 | 2 | 0 |
| 1 | 1 | 0 | 2 | 3 | 2 | 0 | 0 | 2 | 1 | 3 | 1 | 4 | 1 | 1 | 2 | 2 | 1 | 1 |
| 2 | 0 | 0 | 0 | 4 | 3 | 3 | 0 | 1 | 16 | 5 | 0 | 1 | 3 | 8 | 0 | 0 | 1 | 3 | 3 |

Referenced for Figure 2 data
SRS taken from Figure 2 \((n = 20, \ t = 584, \overline{Y}_U = 1.435, \overline{y} = 1.55, \ s^2 = 10.9974)\)

![Histogram Count Data in Figure 2](image)

| 1 1 1 1 1 2 1 0 0 0 4 5 0 1 0 1 2 1 0 1 0 1 |
| 3 2 1 0 1 0 0 0 1 2 2 2 0 2 2 2 0 2 0 1 1 |
| 7 4 1 1 1 1 1 0 0 0 2 2 0 4 3 2 4 2 1 2 2 |
| 0 1 2 0 0 0 (0) 0 0 4 6 5 1 5 0 0 0 0 2 1 2 0 |
| 1 (1) 0 2 3 2 (0) 0 2 1 3 1 4 1 1 1 2 2 1 1 |
| 2 0 0 0 0 0 1 1 1 0 1 6 5 0 1 (3) 8 0 0 1 3 3 |
| 0 (0) 1 (14) 3 (3) 1 2 0 8 (0) 2 0 3 9 0 4 2 1 0 |
| 0 0 5 (1) 8 7 (6) 6 6 1 0 4 0 0 1 2 2 0 1 2 |
| 0 0 2 2 3 2 2 3 1 1 1 3 0 0 2 2 0 3 4 (0) |
| 0 0 0 0 1 0 3 1 1 1 2 0 2 0 2 (0) 2 1 1 0 |
| 1 8 7 7 8 0 5 0 1 (0) 1 2 0 (0) 2 4 2 2 2 4 |
| 0 0 0 0 1 0 1 1 0 0 0 1 2 4 0 2 1 3 3 1 |
| 0 0 0 1 0 2 4 3 1 2 2 0 0 1 1 2 2 0 2 4 |
| 0 1 0 0 1 2 0 2 3 5 2 0 0 2 1 1 2 0 1 3 |
| 1 0 0 1 1 0 0 0 2 2 2 (1) 1 1 0 0 (2) 0 0 0 |
| 0 2 0 2 2 0 1 1 0 2 0 0 1 0 0 1 1 1 5 3 |
| 0 0 0 3 2 1 0 0 0 0 0 0 2 1 0 1 1 1 3 1 2 |
| 1 (0) 0 1 0 3 (0) 1 0 0 2 1 2 0 0 0 0 1 1 1 0 |
| 0 0 0 0 0 0 0 1 1 5 0 0 1 0 3 0 2 0 1 1 0 |
| 2 0 0 0 0 0 0 0 1 2 0 1 3 (0) 0 1 0 1 2 4 |
2.4.2 Using the R Survey Package for a SRS

R Code and Output for Figure 1 SRS Analysis

"count" "fpc" <- This is the contents of the data file fig1.txt
33 400 <- The first column are the recorded responses
33 400 <- The second column is the population size N
30 400
34 400
39 400
27 400
32 400
36 400
35 400
42 400

R Code

source("c:/courses/st446/rcode/confintt.r")

# t-based confidence intervals for SRS in Figure 1

library(survey)
srsdat <- read.table("c:/courses/st446/rcode/fig1.txt", header=T)
srsdat

srs_design <- svydesign(id=~1, fpc=~fpc, data=srsdat)
srs_design

esttotal <- svytotal(~count,srs_design)
print(esttotal,digits=15)
confint.t(esttotal,degf(srs_design),level=.95)
confint.t(esttotal,degf(srs_design),level=.95,tails='lower')
confint.t(esttotal,degf(srs_design),level=.95,tails='upper')

estmean <- svymean(~count,srs_design)
print(estmean,digits=15)
confint.t(estmean,degf(srs_design),level=.95)
confint.t(estmean,degf(srs_design),level=.95,tails='lower')
confint.t(estmean,degf(srs_design),level=.95,tails='upper')

R output for t-based confidence interval for SRS

> srsdat
   count fpc
 1   33 400
 2   33 400
 3   30 400
 4   34 400
 5   39 400
 6   27 400
 7   32 400
 8   36 400
 9   35 400
 10  42 400
Independent Sampling design

<table>
<thead>
<tr>
<th></th>
<th>total</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>13640</td>
<td>534.63</td>
</tr>
</tbody>
</table>

---

mean( count ) = 13640.00000
SE( count ) = 534.62760
Two-Tailed CI for count where alpha = 0.05 with 9 df

<table>
<thead>
<tr>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>12430.58835</td>
<td>14849.41165</td>
</tr>
</tbody>
</table>

---

mean( count ) = 13640.00000
SE( count ) = 534.62760
One-Tailed (Lower) CI for count where alpha = 0.05 with 9 df

<table>
<thead>
<tr>
<th>5 % upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>12659.96724</td>
</tr>
</tbody>
</table>

---

mean( count ) = 13640.00000
SE( count ) = 534.62760
One-Tailed (upper) CI for count where alpha = 0.05 with 9 df

<table>
<thead>
<tr>
<th>lower 95 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>-infinity</td>
</tr>
</tbody>
</table>

---

mean SE
count 34.1 1.3366

---

mean( count ) = 34.10000
SE( count ) = 1.33657
Two-Tailed CI for count where alpha = 0.05 with 9 df

<table>
<thead>
<tr>
<th>2.5 %</th>
<th>97.5 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>31.07647</td>
<td>37.12353</td>
</tr>
</tbody>
</table>

---

mean( count ) = 34.10000
SE( count ) = 1.33657
One-Tailed (Lower) CI for count where alpha = 0.05 with 9 df

<table>
<thead>
<tr>
<th>5 % upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>31.64992</td>
</tr>
</tbody>
</table>

---

mean( count ) = 34.10000
SE( count ) = 1.33657
One-Tailed (upper) CI for count where alpha = 0.05 with 9 df

<table>
<thead>
<tr>
<th>lower 95 %</th>
</tr>
</thead>
<tbody>
<tr>
<td>-infinity</td>
</tr>
</tbody>
</table>
R Code and Output for Figure 2 SRS Analysis

```r
source("c:/courses/st446/rcode/confintt.r")

# t-based confidence intervals for SRS in Figure 2

library(survey)
srsdat <- read.table("c:/courses/st446/rcode/fig2.txt", header=T)
srsdat

srs_design <- svydesign(id=~1, fpc=~fpc, data=srsdat)
srs_design

esttotal <- svytotal(~count, srs_design)
print(esttotal,digits=15)
confint.t(esttotal,degf(srs_design),level=.95)
confint.t(esttotal,degf(srs_design),level=.95,tails='lower')
confint.t(esttotal,degf(srs_design),level=.95,tails='upper')

estmean <- svymean(~count, srs_design)
print(estmean,digits=15)
confint.t(estmean,degf(srs_design),level=.95)
confint.t(estmean,degf(srs_design),level=.95,tails='lower')
confint.t(estmean,degf(srs_design),level=.95,tails='upper')
```

R output for t-based confidence interval for SRS

```
The data file:
"count" "fpc"
1 1 400
2 0 400
3 0 400
4 14 400
5 1 400
6 0 400
7 0 400
8 3 400
9 0 400
10 6 400
11 0 400
12 0 400
13 0 400
14 1 400
15 3 400
16 0 400
17 0 400
18 0 400
19 2 400
20 0 400

   total   SE
count  620  289.1
```
mean( count ) = 620.00000
SE( count ) = 289.10206
Two-Tailed CI for count where alpha = 0.05 with 19 df
  2.5 %  97.5 %
    14.90244  1225.09756

mean( count ) = 620.00000
SE( count ) = 289.10206
One-Tailed (Lower) CI for count where alpha = 0.05 with 19 df
  5 % upper
       120.10415    infinity

mean( count ) = 620.00000
SE( count ) = 289.10206
One-Tailed (upper) CI for count where alpha = 0.05 with 19 df
  lower 95 %
         -infinity  1119.89585

mean( count ) = 1.55000
SE( count ) = 0.72276
Two-Tailed CI for count where alpha = 0.05 with 19 df
  2.5 %  97.5 %
    0.03726   3.06274

mean( count ) = 1.55000
SE( count ) = 0.72276
One-Tailed (Lower) CI for count where alpha = 0.05 with 19 df
  5 % upper
       0.30026    infinity

mean( count ) = 1.55000
SE( count ) = 0.72276
One-Tailed (upper) CI for count where alpha = 0.05 with 19 df
  lower 95 %
         -infinity   2.79974

mean SE
count 1.55 0.7228
2.4.3 Using SAS PROC Surveymeasures for a SRS

DM 'LOG;CLEAR;OUT;CLEAR'; *** I recommend putting these two lines of code;
OPTIONS NODATE NONUMBER; *** at the beginning of every SAS program  

data SRS_Fig1;
  wgt= 400/10; * wgt = N/n ;
  input count @@;
datalines;
33 33 30 34 39 27 32 36 35 42
;
proc surveymeans data=SRS_Fig1 total=400 mean clm sum clsum;
  var count;
  weight wgt;
title1 'Simple Random Sample -- Example 1';
title2 'Estimating the population mean and total from the data in Figure 1';
run;
===========================================================================

Simple Random Sample -- Example 1
Estimating the population mean and total from the data in Figure 1

The SURVEYMEANS Procedure

Data Summary

Number of Observations 10
Sum of Weights 400

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>34.100000</td>
<td>1.336569</td>
<td>31.0764709 37.1235291</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sum</th>
<th>Std Dev</th>
<th>95% CL for Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>13640</td>
<td>534.627596</td>
<td>12430.5884 14849.4116</td>
</tr>
</tbody>
</table>
data SRS_Fig2;
  wgt= 400/20; * wgt = N/n ;
  input trees @@;
  datalines;
  1 0 0 1 4 1 0 0 3 0 6 0 0 0 1 3 0 0 0 2 0
  ;
proc surveymeans data=SRS_Fig2 total=400 mean clm sum clsum;
  var trees;
  weight wgt;
title1 'Simple Random Sample -- Example 2';
title2 'Estimating the population mean and total from the data in Figure 2';
run;

Simple Random Sample -- Example 2
Estimating the population mean and total from the data in Figure 2

The SURVEYMEANS Procedure

Data Summary

  Number of Observations 20
  Sum of Weights 400

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>95% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>trees</td>
<td>1.550000</td>
<td>0.722755</td>
<td>0.03725610 3.06274390</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Sum</th>
<th>Std Dev</th>
<th>95% CL for Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>trees</td>
<td>620.000000</td>
<td>289.102058</td>
<td>14.9024382 1225.09756</td>
</tr>
</tbody>
</table>
2.5 Attribute Proportion Estimation

- Suppose we are interested in an attribute (characteristic) associated with the sampling units. The **population proportion** $p$ is the proportion of population units having that attribute.

- Statistically, the goal is to estimate proportion $p$.

- Examples: the proportion of females (or males) in an animal population, the proportion of consumers who own motorcycles, the proportion of married couples with at least 1 child...

- Statistically, we use an *indicator function* that assigns a $y_i$ value to unit $i$ as follows:

$$ y_i = \begin{cases} 
1 & \text{if unit } i \text{ possesses the attribute} \\
0 & \text{otherwise} 
\end{cases} $$

Then $t = \sum_{i=1}^{N} y_i$ and $\overline{y}_U = \frac{1}{N} \sum_{i=1}^{N} y_i = p$. The population proportion $p$ can be expressed as a population mean $\overline{y}_U$. Therefore, we will, under certain conditions, be able to apply the SRS methods for estimating $\overline{y}_U$.

- By taking a SRS of size $n$, we can estimate $p$ with the **sample proportion** $\hat{p}$ of units that possess that attribute: $\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}$. The sample proportion $\hat{p}$ is unbiased for $p$.

- For a finite population of 0 and 1 values, the population variance

$$ S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (y_i - p)^2 = $$

- Therefore, the variance of $\hat{p}$ is

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

(18)

- Because $S^2$ is unknown, we estimate it with $s^2 = \frac{n}{n-1} \hat{p}(1-\hat{p})$. Substitution provides the unbiased estimator of $V(\hat{p})$:

$$ \hat{V}(\hat{p}) = \left( \frac{N-n}{N} \right) \frac{s^2}{n} = $$

(19)

- The square root of $V(\hat{p})$ in (18) is the **standard deviation** of the estimator $\hat{p}$.

- The square root of $\hat{V}(\hat{p})$ in (19) is the **standard error** of $\hat{p}$.

- The effects of omitting the finite population correction (f.p.c.) from the formulas for large and small samples apply here as they did earlier.
Figure 3: The Presence/Absence of Longleaf Pine

Rathbun/Cressie data \((t = 249 \quad N = 400 \quad p = .6225)\)

| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 |
| 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 |
| 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 |

A simple random sample of size \(n = 25\)

| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | (1) | 0 | 1 | 1 |
| (1) | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | (1) | 1 | (1) | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | (1) | 1 | 1 | 1 | (1) | 1 | 1 |
| 0 | (1) | 1 | 0 | 0 | 0 | 0 | (0) | 1 | 1 | 1 | 1 | 1 | 0 | 0 | (0) | 1 | 1 | 1 | 0 |
| 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | (1) |
| 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 |
| 0 | 0 | 1 | 1 | (1) | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | (0) | 1 | 1 | 1 | 0 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | (1) | 1 | 0 | 1 | 1 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | (1) | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 1 |
| 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | (0) | 1 | 1 | 1 | 0 | 0 | (1) | 1 | 1 | 1 | 1 | 1 |
| 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 |
| 0 | 1 | 0 | 0 | (1) | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 0 | 0 | (0) | 0 | 0 | 0 | (0) | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 |
2.5.1 Confidence Intervals for $p$

- Let the random variable $Y =$ the number of units in a SRS of size $n$ that possess the attribute of interest. We know (in theory) that the sampling distribution of $Y$ follows a hypergeometric distribution.

- Hypergeometric distribution for a SRS: The probability that a SRS of size $n$ will have exactly $j$ sampling units possessing the attribute is

$$\Pr(Y = j) = \frac{{^{t}(N-t) \choose {n-j}} \choose {N \choose n}}$$

= the probability that a SRS will consist of $j$ ones and $n - j$ zeroes selected from the population containing $t$ ones (1’s) and $N - t$ zeroes (0’s).

- Although confidence interval calculations can be based on probability tables of hypergeometric distributions, we will use a more common approach that will apply to many sampling situations.

- Remember there are $t$ ones and $N - t$ zeros in the population. However, $t$ is unknown. If we can assume that $n$ is small relative to both $t$ and $N - t$, we can use the binomial approximation to the hypergeometric distribution. That is, $Y \sim \text{BIN}(n, p)$.

- Although the problem no longer depends on $t$, it still depends on the unknown proportion parameter $p$.

- What is commonly done is to apply the normal approximation to the binomial distribution:

$$\hat{p} \sim N(p, V(\hat{p}))$$

- Thus, if the sample size $n$ is large enough, we use $\hat{V}(\hat{p})$ to estimate $V(\hat{p})$. An approximate $100(1 - \alpha)\%$ confidence interval for $p$ is:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{V}(\hat{p})}{n}} \text{ OR } \hat{p} \pm z^* \sqrt{\frac{n - \hat{n}}{N} \frac{\hat{p}(1 - \hat{p})}{n - 1}} \tag{20}$$

where $z^*$ is the upper $\alpha/2$ critical value from the standard normal distribution. Sample sizes are typically large enough to use $z^*$ instead of $t^*$.

- The normal approximation will be reasonable given

1. $n$ is not too large relative to $t$ or $N - t$. This will be a problem if $p$ is close to 0 or 1.

2. The smaller of $n\hat{p}$ and $n(1 - \hat{p})$ is not too small. In most texts, it is suggested that both $n\hat{p}$ and $n(1 - \hat{p})$ should be $\geq 5$, while some texts use $\geq 10$. 
R Code and Output for Figure 3 Example

```r
source("c:/courses/st446/rcode/confintt.r")

# t-based confidence intervals for SRS in Figure 3

library(survey)
srsdat <- read.table("c:/courses/st446/rcode/fig3.txt", header=T)
srsdat

srs_design <- svydesign(id=~1, fpc=~fpc, data=srsdat)
estmean <- svymean(~presence,srs_design)
print(estmean,digits=15)
confint.t(estmean,degf(srs_design),level=.90)
confint.t(estmean,degf(srs_design),level=.90,tails='lower')
confint.t(estmean,degf(srs_design),level=.90,tails='upper')
```

R output for t-based confidence interval for SRS

```
> srsdat
   presence fpc
 1        1 400
 2        1 400
 3        1 400
   :     : ::
23       0 400
24       0 400
25       0 400

mean    SE
presence 0.72  0.0887

---------------------------------------------
mean( presence ) = 0.72000
SE( presence ) = 0.08874
Two-Tailed CI for presence where alpha = 0.1 with 24 df
  5 %     95 %
  0.56817  0.87183
---------------------------------------------

---------------------------------------------
mean( presence ) = 0.72000
SE( presence ) = 0.08874
One-Tailed (Lower) CI for presence where alpha = 0.1 with 24 df
   10 %  upper
   0.60305   infinity
---------------------------------------------

---------------------------------------------
mean( presence ) = 0.72000
SE( presence ) = 0.08874
One-Tailed (upper) CI for presence where alpha = 0.1 with 24 df
   lower     90 %
 -infinity  0.83695
---------------------------------------------
```


SAS Code and Output for Figure 3 Example

```
DM 'LOG;CLEAR;OUT;CLEAR';
OPTIONS NODATE NONUMBER LS=72 PS=54;

DATA SRS_Fig3;
  INPUT ind @@;
DATALINES;
1 1 1 1 1 1 0 0 1 1 0 1 1 1 0 1 1 0 1 1 0 0
;
DATA SRS_Fig3; set SRS_Fig3;
  IF ind = 0 then pa = 'absent ';
  IF ind = 1 then pa = 'present';

PROC SURVEYMEANS DATA=SRS_Fig3 TOTAL = 400 ALPHA = .10;
  VAR pa;
TITLE 'Simple Random Sample -- Figure 3';
TITLE2 'Estimating population proportion p';
RUN;
```

Simple Random Sample -- Figure 3
Estimating population proportion p

The SURVEYMEANS Procedure

Data Summary

Number of Observations 25

Class Level Information

<table>
<thead>
<tr>
<th>Class Variable</th>
<th>Levels</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td>pa</td>
<td>2</td>
<td>absent, present</td>
</tr>
</tbody>
</table>

Statistics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Level</th>
<th>N</th>
<th>Mean</th>
<th>Std Error of Mean</th>
<th>90% CL for Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>pa</td>
<td>absent</td>
<td>7</td>
<td>0.280000</td>
<td>0.088741</td>
<td>0.12817428 0.43182572</td>
</tr>
<tr>
<td></td>
<td>present</td>
<td>18</td>
<td>0.720000</td>
<td>0.088741</td>
<td>0.56817428 0.87182572</td>
</tr>
</tbody>
</table>
2.6 Sample Size Determination with Simple Random Sampling

- It is well known that an increase in sample size \( n \) will lead to a more precise estimator of \( \overline{y}_U \) or \( t \). It is also obvious that an increase in the sample size \( n \) will make the sample more expensive to collect. There will, however, be a limited amount of resources available (allocated, budgeted) for data collection.

- When designing a sampling plan, the researcher wants to achieve a desired degree of reliability at the lowest possible cost while satisfying the resource limitations for data collection. That is, the goal is to get the most information given resources and constraints.

- To do this, the researcher tries to achieve a balance to avoid the following mistakes:
  - Oversampling: The sampling plan may provide more precision than is needed. Oversampling will lead to increased sampling effort, time, and cost.
  - Undersampling: The sampling plan may yield insufficient precision resulting in producing overly-wide confidence intervals. Undersampling will lead to wasted time and money.

- To determine a sample size \( n \) when estimating a parameter \( \theta \), we do the following:
  - Estimate the sample size \( n \) required so that the probability of the difference between the estimator \( \hat{\theta} \) and the parameter being estimated \( \theta \) exceeds some maximum allowable difference \( d = |\hat{\theta} - \theta| \) is at most \( \alpha \). Or, equivalently, find \( n \) such that \( \Pr(|\hat{\theta} - \theta| > d) < \alpha \).

- This is equivalent to finding \( n \) large enough so that the margin of error

2.6.1 When Estimating \( \overline{y}_U \)

- **Situation:** Estimate the SRS size required so the probability that the difference between the estimator \( \hat{\overline{y}}_U = \overline{y} \) and the population mean \( \overline{y}_U \) does not exceed a maximum allowable difference \( d \) is at most \( \alpha \).

- Mathematically, find \( n \) such that \( \Pr(|\hat{\overline{y}}_U - \overline{y}_U| > d) < \alpha \) for a specified maximum allowable difference \( d \).

- Assuming \( \overline{y} \) is approximately normally distributed, this is equivalent to finding \( n \) so that the margin of error \( \frac{z_{\alpha/2} \sqrt{\left(1 - \frac{n}{N}\right) \frac{S^2}{n}}}{\sqrt{N}} \leq d \). Solving this inequality for \( n \) yields

\[
\frac{1}{z^2S^2 + \frac{1}{N}} = n = \frac{1}{\frac{d^2}{z^2S^2} + \frac{1}{N}} \quad (21)
\]

where \( n_0 = \) and \( z \) is the critical \( \alpha/2 \) value from a \( N(0, 1) \) distribution.

- Rounding up the value of \( n \) in (21) yields the desired sample size. If this value is < 30, I recommend adding 2 or 3 to this value to account for the use of the large sample \( z^* \) in the previous formulas instead of a smaller sample \( t^* \).
For example, consider the spatially correlated population in Figure 1. How large a sample would be required so that \( \hat{y}_{U} = \bar{y} \) is within 1 of \( \bar{y}_U \) with probability at least .95 (\( \alpha = .05 \))? (Assume \( S^2 \approx 18.3 \))

• If the population size \( N \) is very large, then \( 1/N \approx 0 \). In this case, \( n \approx n_0 \). This is the formula given in introductory statistics books.

• There remains one major problem. This sample size formula assumes that you know the population variance \( S^2 \). Therefore, to estimate the sample size \( n \), we need a prior estimate of \( S^2 \). Barnett (1997, pages 33-34) describes 4 ways to do this:

1. A Pilot Study: A small sample size pilot study can be conducted prior to the primary study to provide an estimate of \( S^2 \).

2. Previous Studies: Other similar studies may have been conducted elsewhere and appear in the professional journals. Measures of variability from earlier studies may provide an estimate of \( S^2 \).

3. Double Sampling: A preliminary SRS of size \( n_1 \) is taken and the sample variance \( s_1^2 \) is used to estimate \( S^2 \). Using \( s_1^2 \) in (21) will approximate an adequate sample size \( n \). Then, a further SRS of size \( n - n_1 \) is taken from the remaining unsampled \( N - n_1 \) sampling units. This is an example of double sampling.

4. Exploiting the structure of the population: Sometimes we may have some knowledge of the structure of the population which can provide information about \( S^2 \).
   - A common case is when you have count data and it is reasonable to assume the distribution of counts follows a Poisson distribution. Because the mean and the variance of a Poisson distribution are the same, all we need is a prior estimate of the population mean.
   - A second case occurs with estimation of a proportion \( p \) for a binomial distribution. If we have a prior estimate of \( p \), we also have a prior estimate of the variance which is a function of \( p \).

2.6.2 When Estimating \( t \)

• Situation: Estimate the SRS size required so the probability that the difference between the estimator \( \hat{t} = N\bar{y} \) and the population total \( t \) does not exceed a maximum allowable difference \( d \) is at most \( \alpha \).

• Mathematically, find \( n \) such that \( \Pr(|\hat{t} - t| > d) < \alpha \) for a specified maximum allowable difference \( d \).

• Assuming \( N\bar{y} \) is approximately normally distributed, this is equivalent to finding \( n \) so that the margin of error \( z_{\alpha/2} \sqrt{N(N - n) \frac{S^2}{n}} \leq d \). Solving this inequality for \( n \) yields

\[
\frac{1}{\frac{d^2}{N^2} + \frac{1}{N}} = \frac{1}{\frac{d^2}{N^2} + \frac{1}{N}}
\]

where \( n_0 = \) and \( z \) is the critical \( \alpha/2 \) value from a \( N(0, 1) \) distribution.
• Rounding-up the value of $n$ in (22) yields the desired sample size. If this value is $< 30$, I recommend adding 2 or 3 to this value.

  – For example, consider the longleaf pine population in Figure 2. How large a sample would be required so that $\hat{t}$ is within 15 of $t$ with probability at least .95 ($\alpha = .05$)? (Assume $S^2 \approx 4$)

• If the population size $N$ is very large, then $1/N \approx 0$. In this case, $n \approx n_0$.

2.6.3 When Estimating $p$

• Situation: Estimate the SRS size required so the probability that the difference between the sample proportion $\hat{p}$ and the population proportion $p$ does not exceed a maximum allowable difference $d$ is at most $\alpha$.

  – For example, consider the longleaf pine presence/absence population in Figure 3. How large a sample would be required so that $\hat{p}$ is within .05 of $p$ with probability at least .95?

• Mathematically, find $n$ such that $\Pr(|\hat{p} - p| > d) \leq \alpha$ for a specified maximum allowable difference $d$.

• Assuming $\hat{p}$ is approximately normally distributed, this is equivalent to finding $n$ so that the margin of error $z_{\alpha/2} \sqrt{\frac{(N-n)}{N-1} \frac{p(1-p)}{n}} \leq d$.

• Solving this inequality for $n$ yields

$$n = \frac{Np(1-p)}{(N-1)\frac{d^2}{z^2} + p(1-p)} = \approx \frac{1}{n_0} + \frac{1}{N} \quad (23)$$

where $n_0 = \ldots$ and $z$ is the critical $\alpha/2$ value from a $N(0, 1)$ distribution.

• Rounding-up the value of $n$ in (23) yields the desired sample size.

• Because $N$ is typically large when estimating $p$, it is common to ignore the f.p.c. If you, the estimated sample size is $n \approx n_0$.

• Unfortunately, the sample size formulas assume you know the population proportion $p$, the quantity you are trying to estimate. Thus, to estimate an adequate sample size, we need a prior estimate of $p$. In addition to the four methods of Barnett (pp 33-34), there is also the following conservative approach.

• Note that the standard deviation of $\hat{p} = \text{s.d.}(\hat{p}) = \sqrt{\frac{(N-n)}{N-1} \frac{p(1-p)}{n}}$ is largest when $p = 1/2$. Thus, it is conservative to use $p = 1/2$ in (23) if there is no prior reasonable estimate.

• Example: Consider the longleaf pine presence/absence population in Figure 3. How large a sample would be required so that $\hat{p}$ is within .05 of $p$ with probability at least .95?

  (i) Assume we use $p \approx .72$ based on the earlier SRS with $n = 25$.

  (ii) Assume we have no prior estimate of $p$ and use the conservative estimate of $p = .5$. 

44