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 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 choose n) $\binom{N}{n} = \dfrac{N(N-1)\cdots(N-n+1)}{n!} = \dfrac{N!}{n!(N-n)!}$. This is the number of combinations of n items selected from N distinct items (and the order of selection doesn't matter). For example, $\binom{6}{2} = \dfrac{6!}{2!4!} = \dfrac{(6)(5)(4!)}{2!4!} = \dfrac{(6)(5)}{(2)(1)} = 15$.

- 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 $P(\mathcal{S}) = 1/\binom{N}{n}$.

- 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 $\widehat{t}$ of the population total t. That is:

$$\widehat{\overline{y}_U} = \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \qquad\qquad \widehat{t} = \frac{N}{n}\sum_{i=1}^{n} y_i = \tag{10}$$

- $\widehat{\overline{y}_U}$ and $\widehat{t}$ 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:

Unit i	1	2	3	4	5
y_i	0	2	3	4	7

which has the following parameters:

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

Suppose a SRS of size $n = 2$ is selected. Then $P(\mathcal{S}) = 1/\binom{5}{2} = 1/10$ for each of the 10 possible SRSs.

All Possible Samples and Statistics from Example Population

Sample	Units	y-values	$\sum y_i$	$\widehat{\overline{y}_U} = \overline{y}$	$\widehat{t} = N\overline{y}$	$\widehat{S}^2 = s^2$	$\widehat{S} = s$
$\mathcal{S}_1$	1,2	0,2	2	1	5	2	1.4142
$\mathcal{S}_2$	1,3	0,3	3	1.5	7.5	4.5	2.1213
$\mathcal{S}_3$	1,4	0,4	4	2	10	8	2.8284
$\mathcal{S}_4$	1,5	0,7	7	3.5	17.5	24.5	4.9497
$\mathcal{S}_5$	2,3	2,3	5	2.5	12.5	.5	0.7071
$\mathcal{S}_6$	2,4	2,4	6	3	15	2	1.4142
$\mathcal{S}_7$	2,5	2,7	9	4.5	22.5	12.5	3.5355
$\mathcal{S}_8$	3,4	3,4	7	3.5	17.5	.5	0.7071
$\mathcal{S}_9$	3,5	3,7	10	5	25	8	2.8284
$\mathcal{S}_{10}$	4,5	4,7	11	5.5	27.5	4.5	2.1213
Column Sum				32	160	67	22.6274
Expected value $= E$(estimator)				$\frac{32}{10} = 3.2$ $= \overline{y}_U$	$\frac{160}{10} = 16$ $= t$	$\frac{67}{10} = 6.7$ $= S^2$	$\frac{22.6274}{10} = 2.26274$ $\neq S$

The averages for estimators $\widehat{\overline{y}_U} = \overline{y}$, $\widehat{t} = N\overline{y}$, and $\widehat{S}^2 = s^2$ equal the parameters that they are estimating. This implies that $\overline{y}$, $N\overline{y}$, and s^2 are unbiased estimators of $\overline{y}_U$, t, and S^2.
Notation: $E(\widehat{\overline{y}_U}) = \overline{y}_U$, $E(\widehat{t}) = t$, $E(\widehat{S}^2) = S^2$ or $E(\overline{y}) = \overline{y}_U$, $E(N\overline{y}) = t$, $E(s^2) = S^2$.

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

- The next problem is to study the variances of $\widehat{\overline{y}_U} = \overline{y}$ and $\widehat{t} = N\overline{y}$.
- Warning: In an introductory statistics course, you were told that the variance of the sample mean $V(\overline{Y}) = S^2/n$ $(= \sigma^2/n)$ 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 $\dfrac{N-n}{N}$ which is known as the **finite population correction (f.p.c.)**.
- Texts may rewrite the f.p.c. $\dfrac{N-n}{N}$ as either $1 - \dfrac{n}{N}$ or $1 - f$ where $f = n/N$ is the fraction of the population that was sampled. By definition :

$$V(\widehat{\overline{y}_U}) \;=\; V(\overline{y}) \;= \qquad\qquad V(\widehat{t}) \;=\; N^2 V(\overline{y}) \;=\; N(N-n)\frac{S^2}{n} \tag{11}$$

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

$$\widehat{V}(\widehat{\overline{y}_U}) \;=\; \widehat{V}(\overline{y}) \;= \qquad\qquad \widehat{V}(\widehat{t}) \;=\; N^2 \widehat{V}(\overline{y}) \;=\; N(N-n)\frac{s^2}{n} \tag{12}$$

- 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, $V(\overline{y}) = \left(\frac{N-n}{N}\right)\frac{S^2}{n} = \frac{3}{5}\frac{6.7}{2} =$ and

  $V(\widehat{t}) = N^2V(\overline{y}) = N(N-n)\frac{S^2}{n} = (5)(3)\,\frac{6.7}{2} =$.

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

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

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

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

Unit, i	1	2	3	4	5
y_i	0	2	3	4	7

which have the following parameters

$$N = 5 \qquad t = 16 \qquad \overline{y}_U = 3.2 \qquad S^2 = 6.7 \qquad S \approx 2.588$$

Estimated Variances of $\widehat{\overline{y}_U}$ and $\widehat{t}$ for All Samples

Sample	Units	y-values	s^2	$\widehat{V}(\widehat{\overline{y}_U}) = .3s^2$	$\widehat{V}(\widehat{t}) = 7.5s^2$
$\mathcal{S}_1$	1,2	0,2	2	0.6	15
$\mathcal{S}_2$	1,3	0,3	4.5	1.35	33.75
$\mathcal{S}_3$	1,4	0,4	8	2.4	60
$\mathcal{S}_4$	1,5	0,7	24.5	7.35	183.75
$\mathcal{S}_5$	2,3	2,3	.5	0.15	3.75
$\mathcal{S}_6$	2,4	2,4	2	0.6	15
$\mathcal{S}_7$	2,5	2,7	12.5	3.75	93.75
$\mathcal{S}_8$	3,4	3,4	.5	0.15	3.75
$\mathcal{S}_9$	3,5	3,7	8	2.4	60
$\mathcal{S}_{10}$	4,5	4,7	4.5	1.35	33.75
Column Sum					

- From the table we have $\mathrm{E}(\widehat{V}(\widehat{\overline{y}_U})) = 20.1/10 = 2.01 = V(\widehat{\overline{y}_U})$ and $E(\widehat{V}(\widehat{t})) = 502.5/10 = 50.25 = V(\widehat{t})$. 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: $\widehat{\overline{y}_U} = \overline{y} \qquad \widehat{V}(\widehat{\overline{y}_U}) = \widehat{V}(\widehat{\overline{y}}) = \dfrac{s^2}{n}$

- Suppose we have two estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of some parameter θ.

  $\widehat{\theta}_1$ is **less efficient** than $\widehat{\theta}_2$ for estimating θ if $V(\widehat{\theta}_1) > V(\widehat{\theta}_2)$.
  $\widehat{\theta}_1$ is **more efficient** than $\widehat{\theta}_2$ for estimating θ if $V(\widehat{\theta}_1) < V(\widehat{\theta}_2)$.

- 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 $\overline{y}_U$ and t

- In an introductory statistics course, you were given confidence interval formulas

$$\overline{y} \pm z^* \frac{s}{\sqrt{n}} \qquad \text{and} \qquad \overline{y} \pm t^* \frac{s}{\sqrt{n}} \tag{13}$$

  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 $\widehat{\overline{y}_U} = \overline{y}$ and $\widehat{t} = N\overline{y}$ have sampling distributions that are approximately normal. That is,

$$\widehat{\overline{y}_U} \mathrel{\dot\sim} N\left(\overline{y}_U\,,\ \frac{N-n}{N}\frac{S^2}{n}\right) \qquad \text{and} \qquad \widehat{t} \mathrel{\dot\sim} N\left(t\,,\ N(N-n)\frac{S^2}{n}\right)$$

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

  For $\overline{y}_U$: For t : (14)

$$\overline{y} \pm z^* \sqrt{\left(\frac{N-n}{N}\right)\frac{s^2}{n}} \qquad N\overline{y} \pm z^* \sqrt{N(N-n)\frac{s^2}{n}}$$

$$\overline{y} \pm z^* s \sqrt{\left(\frac{N-n}{N}\right)/n} \qquad N\overline{y} \pm z^* s \sqrt{N(N-n)/n} \tag{15}$$

  where z^* is the upper $\alpha/2$ critical value from the standard normal distribution. Or, in standard error (s.e.) notation,

$$\widehat{\overline{y}_U} \pm \qquad\qquad \widehat{t} \pm$$

  For 90%, 95%, and 99%, $z^* = 1.645, 1.96,$ and 2.576, respectively.

- For smaller samples, approximate $100(1-\alpha)\%$ confidence intervals for $\overline{y}_U$ and t are

For $\overline{y}_U$: For t : (16)

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

$$\overline{y} \pm t^* s\sqrt{\left(\frac{N-n}{N}\right)/n} \qquad N\overline{y} \pm t^* s\sqrt{N(N-n)/n} \qquad (17)$$

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

- The quantity being added and subtracted from $\widehat{\overline{y}_U} = \overline{y}$ or $\widehat{t} = N\overline{y}$ 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.

All Possible Samples and Confidence Intervals from Example Population

Sample	y-values	$\sum y_i$	$\widehat{\overline{y}_U} = \overline{y}$	$\widehat{t} = N\overline{y}$	$\widehat{S}^2 = s^2$	$\widehat{S} = s$	$\widehat{V}(\widehat{\overline{y}_U})$	$\widehat{V}(\widehat{t})$	90% ci for t
1	0,2	2	1	5	2	1.4142	0.6	15	(-19.45, 29.45)
2	0,3	3	1.5	7.5	4.5	2.1213	1.35	33.75	(-29.18, 44.18)
3	0,4	4	2	10	8	2.8284	2.4	60	(-38.91, 58.91)
4	0,7	7	3.5	17.5	24.5	4.9497	7.35	183.75	(-68.09, 103.09)
5	2,3	5	2.5	12.5	.5	0.7071	0.15	3.75	(0.27, 24.73)
6	2,4	6	3	15	2	1.4142	0.6	15	(-9.45, 39.45)
7	2,7	9	4.5	22.5	12.5	3.5355	3.75	93.75	(-38.63, 83.63)
8	3,4	7	3.5	17.5	.5	0.7071	0.15	3.75	(5.27, 29.73)
9	3,7	10	5	25	8	2.8284	2.4	60	(-23.91, 73.91)
10	4,7	11	5.5	27.5	4.5	2.1213	1.35	33.75	(-9.18, 64.18)

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 <u>upper</u> and <u>lower</u> $100(1-\alpha)\%$ confidence intervals for $\overline{y}_U$ and t are:

For $\overline{y}_U$: For t :

$$\left(\overline{y} - t^* s\sqrt{\left(\frac{N-n}{N}\right)/n}\ ,\ \infty\right) \qquad \left(N\overline{y} - t^* s\sqrt{N(N-n)/n}\ ,\ \infty\right) \qquad \textbf{upper}$$

$$\left(-\infty\ ,\ \overline{y} + t^* s\sqrt{\left(\frac{N-n}{N}\right)/n}\right) \qquad \left(-\infty\ ,\ N\overline{y} + t^* s\sqrt{N(N-n)/n}\right) \qquad \textbf{lower}$$

where t^* is the upper <u>α critical value</u> 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 ∞ with 0 in the upper confidence interval formulas.

- 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.

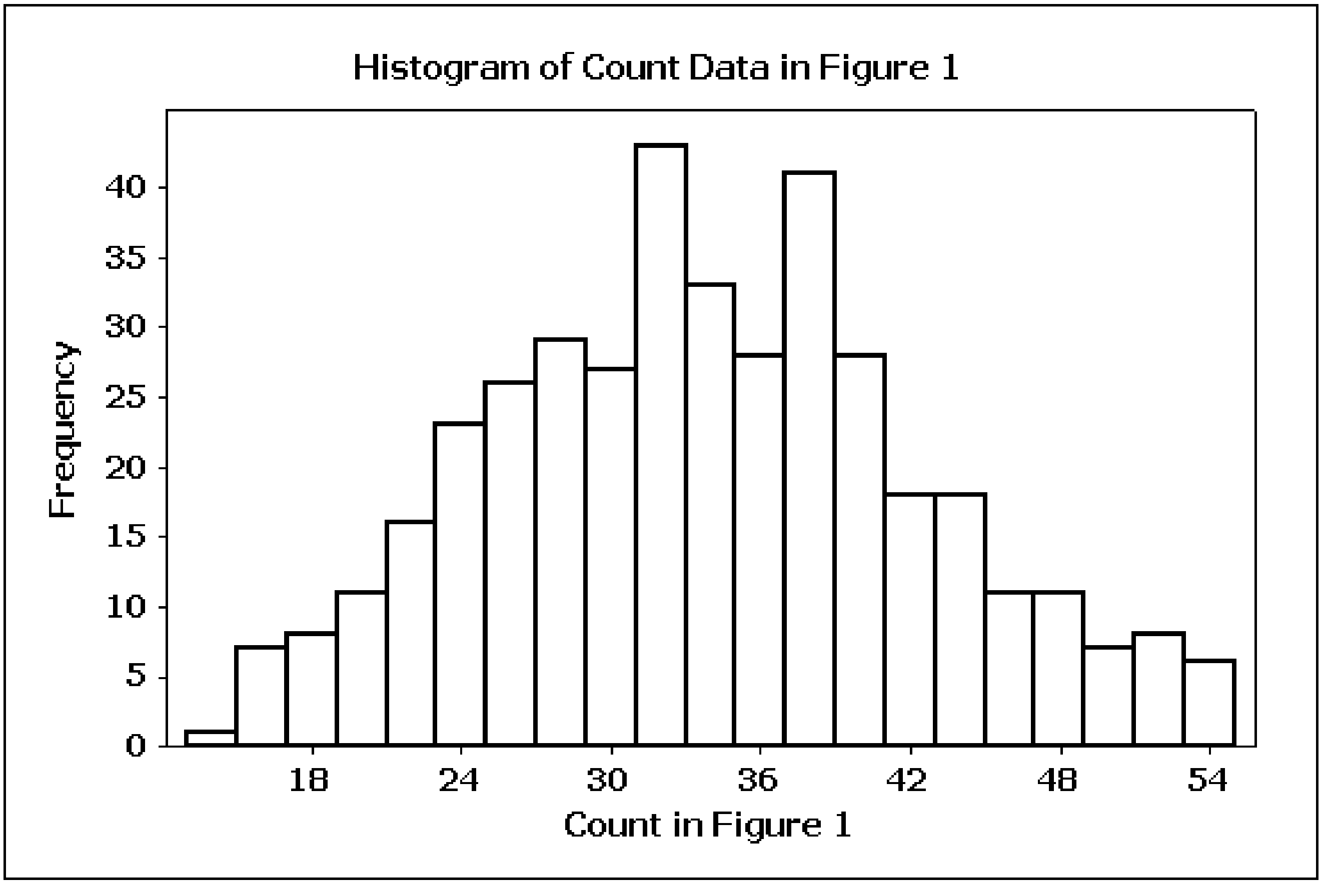
SRS Example with Strong Spatial Correlation

- To illustrate the application of simple random sampling to population mean per unit μ 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 20×20 grid of 10×10 m quadrats. The total abundance $t = 13354$ and the mean per unit is $\overline{y}_U = 33.385$. 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

18	20	15	20	20	15	19	18	24	23	20	26	29	28	28	31	31	34	28	32
13	20	16	20	15	23	19	26	21	21	24	30	23	26	25	33	31	28	32	38
16	18	20	24	25	26	22	23	26	26	22	27	25	25	34	28	37	36	38	31
17	17	16	22	21	23	22	27	27	24	28	32	29	33	27	37	37	38	35	33
15	19	23	17	21	23	21	23	24	25	31	26	32	34	32	33	31	31	36	37
21	24	20	21	28	26	30	22	31	25	29	29	27	30	29	37	35	32	38	43
23	17	24	25	24	27	31	29	31	34	27	36	29	29	34	39	37	37	40	36
18	24	21	25	27	22	32	32	31	26	28	34	34	37	35	34	38	38	37	40
22	26	28	26	24	29	33	26	27	27	34	31	39	32	36	38	37	40	44	43
23	27	28	29	26	32	25	31	35	34	32	33	37	32	42	40	40	37	42	44
23	21	31	23	30	27	31	30	32	35	30	40	32	37	37	36	40	44	44	40
26	29	31	26	30	31	34	36	30	38	36	32	38	38	37	42	42	41	40	49
28	24	28	27	26	31	32	29	32	33	38	34	39	38	40	37	41	43	42	43
32	25	31	32	29	29	35	38	38	32	36	35	39	42	39	40	44	42	41	45
27	29	35	28	35	35	31	40	35	37	38	44	40	40	47	39	49	48	51	49
30	29	32	32	33	30	36	38	42	36	35	38	44	47	45	49	41	43	44	51
28	35	35	34	34	33	41	33	34	35	39	44	44	48	44	50	49	48	53	54
29	33	32	36	39	33	33	34	35	42	46	47	48	47	46	45	44	52	54	55
28	37	38	37	33	33	34	37	45	40	39	42	42	46	47	48	52	47	46	53
38	39	39	37	34	38	39	45	39	42	45	41	44	51	46	50	52	51	51	53



SRS taken from Figure 1 $(n = 10,\ t = 13354,\ \overline{y}_U = 33.385,\ \overline{y} = 34.1,\ s^2 = 18.3\overline{2})$

18	20	15	20	20	15	19	18	24	23	20	26	29	28	28	31	31	34	28	32
13	20	16	20	15	23	19	26	21	21	24	30	23	26	25	**(33)**	31	28	32	38
16	18	20	24	25	26	22	23	26	26	22	27	25	25	34	28	37	36	38	31
17	17	16	22	21	23	22	27	27	24	28	32	29	**(33)**	27	37	37	38	35	33
15	19	23	17	21	23	21	23	24	25	31	26	32	34	32	33	31	31	36	37
21	24	20	21	28	26	**(30)**	22	31	25	29	29	27	30	29	37	35	32	38	43
23	17	24	25	24	27	31	29	31	34	27	36	29	29	34	39	37	37	40	36
18	24	21	25	27	22	32	32	31	26	28	34	34	37	35	**(34)**	38	38	37	40
22	26	28	26	24	29	33	26	27	27	34	31	**(39)**	32	36	38	37	40	44	43
23	27	28	29	26	32	25	31	35	34	32	33	37	32	42	40	40	37	42	44
23	21	31	23	30	27	31	30	32	35	30	40	32	37	37	36	40	44	44	40
26	29	31	26	30	31	34	36	30	38	36	32	38	38	37	42	42	41	40	49
28	24	28	**(27)**	26	31	32	29	32	33	38	34	39	38	40	37	41	43	42	43
32	25	31	**(32)**	29	29	35	38	38	32	**(36)**	35	39	42	39	40	44	42	41	45
27	29	35	28	35	35	31	40	35	37	38	44	40	40	47	39	49	48	51	49
30	29	32	32	33	30	36	38	42	36	35	38	44	47	45	49	41	43	44	51
28	**(35)**	35	34	34	33	41	33	34	35	39	44	44	48	44	50	49	48	53	54
29	33	32	36	39	33	33	34	35	42	46	47	48	47	46	45	44	52	54	55
28	37	38	37	33	33	34	37	45	40	39	42	**(42)**	46	47	48	52	47	46	53
38	39	39	37	34	38	39	45	39	42	45	41	44	51	46	50	52	51	51	53

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 200×200 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 20×20 grid of 10×10 m quadrats. The total abundance $t = 584$ and the mean abundance per quadrat $\overline{y}_U = 584/400 = 1.435$. 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**).

Figure 2

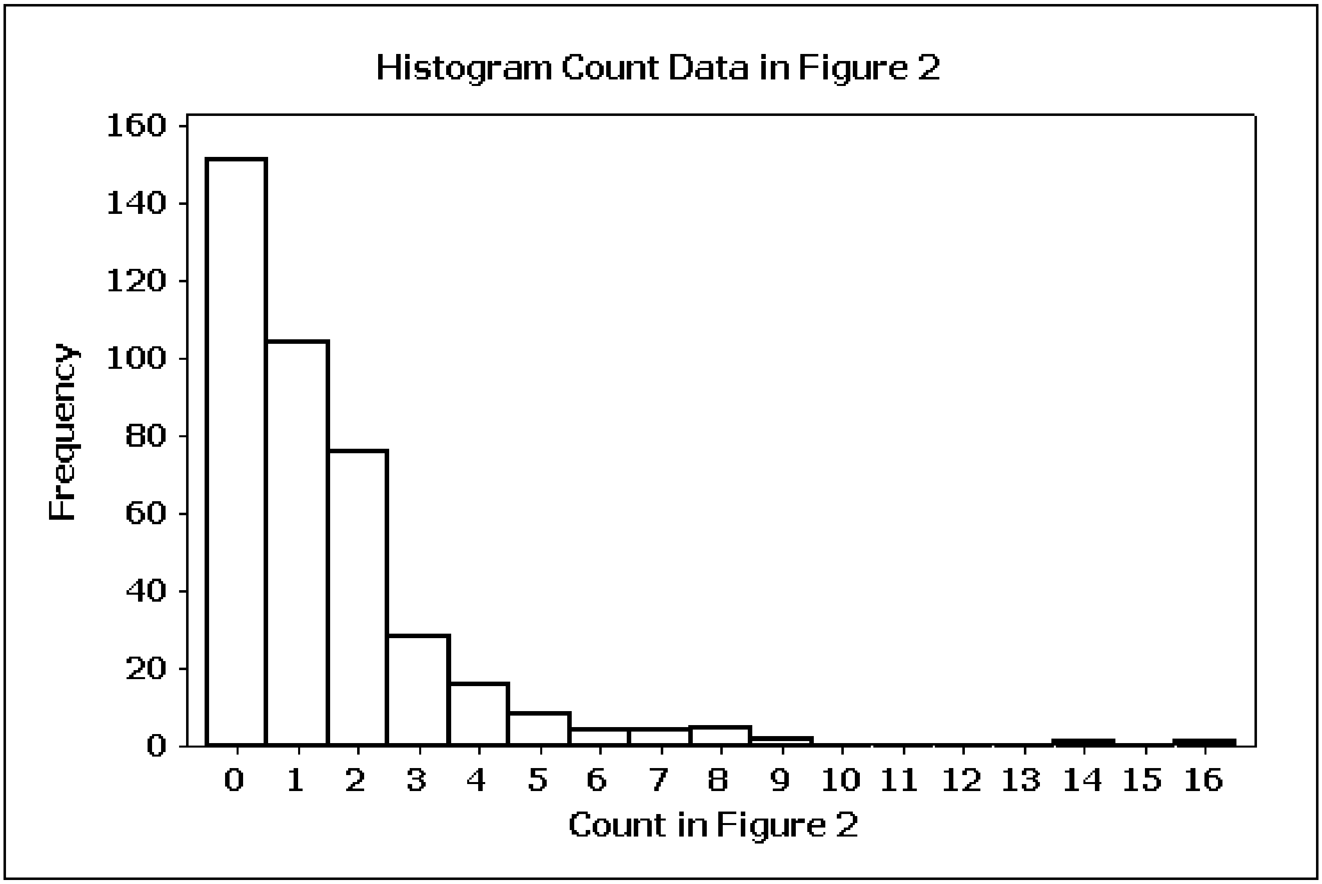
Longleaf Pine Data (Rathbun and Cressie 1994)

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	2	0	2	2	2	0	2	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	6	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	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
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	9	1	0	0	1	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	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	1	1	1	0
0	0	0	0	0	0	0	1	1	1	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

REFERENCES (for Figure 2 data)

Cressie, Noel (1991) *Statistics for Spatial Data*. Wiley, New York.

Rathbun, S.L. and Cressie, N. (1994) A space-time survival point process for a longleaf pine forest in southern Georgia. *Journal of the American Statistical Association*, **89**, 1164-1174.



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

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	2	0	2	2	2	0	2	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	6	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	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
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	9	1	0	**(0)**	1	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	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	1	1	1	0
0	0	0	0	0	0	0	1	1	1	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

      total     SE
count 13640 534.63

-------------------------------------------------------------------
mean( count ) = 13640.00000
SE( count ) = 534.62760
Two-Tailed CI for count where alpha = 0.05 with 9 df
    2.5 %        97.5 %
  12430.58835     14849.41165
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 13640.00000
SE( count ) = 534.62760
One-Tailed (Lower) CI for count where alpha = 0.05 with 9 df
     5 %        upper
  12659.96724     infinity
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 13640.00000
SE( count ) = 534.62760
One-Tailed (upper) CI for count where alpha = 0.05 with 9 df
    lower         95 %
  -infinity     14620.03276
-------------------------------------------------------------------

      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
    2.5 %        97.5 %
  31.07647     37.12353
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 34.10000
SE( count ) = 1.33657
One-Tailed (Lower) CI for count where alpha = 0.05 with 9 df
     5 %        upper
  31.64992     infinity
-------------------------------------------------------------------

-------------------------------------------------------------------
mean( count ) = 34.10000
SE( count ) = 1.33657
One-Tailed (upper) CI for count where alpha = 0.05 with 9 df
    lower         95 %
  -infinity     36.55008
-------------------------------------------------------------------
```

R Code and Output for Figure 2 SRS Analysis

```
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                              "count" "fpc"
1      1 400                              1 400
2      0 400                              0 400
3      0 400                              0 400
4     14 400                              14 400
5      1 400                              1 400
6      0 400                              0 400
7      0 400                              0 400
8      3 400                              3 400
9      0 400                              0 400
10     6 400                              6 400
11     0 400                              0 400
12     0 400                              0 400
13     0 400                              0 400
14     1 400                              1 400
15     3 400                              3 400
16     0 400                              0 400
17     0 400                              0 400
18     0 400                              0 400
19     2 400                              2 400
20     0 400                              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     SE
count 1.55 0.7228

-------------------------------------------------------------------
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
-------------------------------------------------------------------
```

2.4.3 Using SAS PROC Surveymeans 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

                             Std Error
Variable            Mean       of Mean      95% CL for Mean
--------------------------------------------------------------
count          34.100000      1.336569   31.0764709 37.1235291
--------------------------------------------------------------


Variable             Sum       Std Dev        95% CL for Sum
---------------------------------------------------------------
count              13640    534.627596    12430.5884 14849.4116
---------------------------------------------------------------
```

```
DM 'LOG;CLEAR;OUT;CLEAR';
OPTIONS NODATE NONUMBER LS=80 PS=400;

data SRS_Fig2;
      wgt= 400/20;      * wgt = N/n ;
      input trees @@;
      datalines;
1 0 0 14 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

                             Std Error
Variable           Mean        of Mean      95% CL for Mean
--------------------------------------------------------------
trees          1.550000       0.722755   0.03725610 3.06274390
--------------------------------------------------------------


Variable            Sum        Std Dev        95% CL for Sum
---------------------------------------------------------------
trees        620.000000     289.102058    14.9024382 1225.09756
---------------------------------------------------------------
```

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** $\widehat{p}$ of units that possess that attribute: $\widehat{p} = \frac{\sum_{i=1}^{n} y_i}{n} = \overline{y}$. The sample proportion $\widehat{p}$ 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 $\widehat{p}$ is

$$V(\widehat{p}) = \left(\frac{N-n}{N}\right)\frac{S^2}{n} = \left(\frac{N-n}{N}\right)\left(\frac{N}{N-1}\right)\frac{p(1-p)}{n} = \tag{18}$$

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

$$\widehat{V}(\widehat{p}) = \left(\frac{N-n}{N}\right)\frac{s^2}{n} = \tag{19}$$

- The square root of $V(\widehat{p})$ in (18) is the **standard deviation** of the estimator $\widehat{p}$.
- The square root of $\widehat{V}(\widehat{p})$ in (19) is the **standard error** of $\widehat{p}$.
- 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	1	0	0	0	1	1	0	1	0	1	1	1	0	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	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	0	0	0	1	1	1	0	1	1	1	0	1	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	1	0	1	0	0	1	1	1	0	1	1
0	0	1	1	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	1	0
1	1	1	1	1	0	1	0	1	0	1	1	0	0	1	1	1	1	1	1
0	1	1	0	0	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1
0	0	0	1	0	1	1	1	1	1	1	0	0	1	1	1	1	0	1	1
0	1	0	0	1	1	0	1	1	1	1	0	0	1	1	1	1	0	1	1
1	0	0	1	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0
0	1	0	1	1	0	1	1	0	1	0	0	1	0	0	1	1	1	1	1
0	0	0	1	1	1	0	0	0	0	0	1	1	0	1	1	1	1	1	1
1	0	0	1	0	1	0	1	0	0	1	1	1	0	0	0	1	1	1	0
0	0	0	0	0	0	0	1	1	1	0	1	0	1	0	1	0	1	1	0
1	0	0	0	0	0	0	0	1	1	0	1	1	0	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	0	1	0	0	0	1	**(1)**	1	**(1)**	0	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
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	0	0	0	1	1	1	0	1	1	1	0	1	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	1	0	1	0	0	1	1	1	0	1	1
0	0	1	1	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	1	0
1	1	1	1	1	0	1	0	1	0	1	1	0	0	**(1)**	1	1	1	1	1
0	1	1	0	0	1	1	1	0	0	0	**(1)**	1	1	0	1	1	1	1	1
0	0	0	1	0	1	1	1	1	1	1	0	**(0)**	1	1	**(1)**	**(1)**	0	1	1
0	1	0	0	1	1	0	**(1)**	1	1	1	0	0	1	1	1	1	0	1	1
1	0	0	1	1	0	0	0	1	1	1	1	1	1	0	0	1	0	0	0
(0)	1	0	**(1)**	1	0	1	1	0	1	0	0	1	0	0	1	1	1	1	1
0	0	0	1	1	1	0	0	0	0	0	1	1	0	**(1)**	1	1	1	1	1
1	0	0	1	0	1	0	1	0	0	1	1	1	0	0	0	1	1	1	0
0	0	**(0)**	0	0	0	**(0)**	1	1	1	0	1	0	1	0	1	0	1	1	0
1	0	0	0	0	0	0	0	1	1	0	1	1	0	0	1	0	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

$$\Pr(Y = j) \;=\; \frac{\binom{t}{j}\binom{N-t}{n-j}}{\binom{N}{n}}$$

  = 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, $Y \,\dot\sim\, \mathrm{BIN}(n, p)$.

- 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:

$$\widehat{p} \;\dot\sim\; N\left(p, V(\widehat{p})\right).$$

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

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

  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 $n\widehat{p}$ and $n(1-\widehat{p})$ is not too small. In most texts, it is suggested that both $n\widehat{p}$ and $n(1-\widehat{p})$ should be ≥ 5, while some texts use ≥ 10.

R Code and Output for Figure 3 Example

```
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 1 0 0 1 1 0 1 1 1 1 0 1 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

                 Class
                 Variable      Levels    Values

                 pa                 2    absent present


                                Statistics

                                       Std Error
Variable   Level       N         Mean       of Mean      90% CL for Mean
-------------------------------------------------------------------------
pa         absent      7     0.280000      0.088741   0.12817428 0.43182572
           present    18     0.720000      0.088741   0.56817428 0.87182572
-------------------------------------------------------------------------
```

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 θ, we do the following:

  - Estimate the sample size n required so that the probability of the difference between the estimator $\widehat{\theta}$ and the parameter being estimated θ exceeds some maximum allowable difference $d = |\widehat{\theta} - \theta|$ is at most α. Or, equivalently, find n such that $\Pr(|\widehat{\theta} - \theta| > d) \ < \ \alpha$.

- 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 $\widehat{\overline{y}_U} = \overline{y}$ and the population mean $\overline{y}_U$ does not exceed a maximum allowable difference d is at most α.

- Mathematically, find n such that $\Pr(|\widehat{\overline{y}_U} - \overline{y}_U| > d) \ < \ \alpha$ 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 $z_{\alpha/2}\sqrt{\left(\frac{N-n}{N}\right)\frac{S^2}{n}} \leq d$. Solving this inequality for n yields

$$n \ = \ \frac{1}{\frac{d^2}{z^2 S^2} + \frac{1}{N}} \ = \tag{21}$$

  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 $\widehat{\overline{y}_U} = \overline{y}$ is within 1 of $\overline{y}_U$ 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 $\widehat{t} = N\overline{y}$ and the population total t does not exceed a maximum allowable difference d is at most α.

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

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

$$n = \frac{1}{\frac{d^2}{N^2 z^2 S^2} + \frac{1}{N}} = \qquad (22)$$

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 $\widehat{t}$ 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 $\widehat{p}$ and the population proportion p does not exceed a maximum allowable difference d is at most α.
  - For example, consider the longleaf pine presence/absence population in Figure 3. How large a sample would be required so that $\widehat{p}$ is within .05 of p with probability at least .95?
- Mathematically, find n such that $\Pr(|\widehat{p} - p| > d) \;\leq\; \alpha$ for a specified maximum allowable difference d.
- Assuming $\widehat{p}$ is approximately normally distributed, this is equivalent to finding n so that the margin of error $z_{\alpha/2}\sqrt{\left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}} \leq d$.
- Solving this inequality for n yields

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

  where $n_0 =$ 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 $\widehat{p} = \text{s.d.}(\widehat{p}) = \sqrt{\left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}}$ 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 $\widehat{p}$ 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 $p = .5$.