

2 SIMPLE RANDOM SAMPLING

2.1 Population Parameters

- Let τ be the **population total** and μ be the **population mean** from a finite population of size N . Thus, by definition:

$$\tau = \sum_{i=1}^N y_i \quad \mu = \frac{1}{N} \sum_{i=1}^N y_i = \tau/N \quad (1)$$

- If you look at different sampling texts, the **population variance** may be defined as either:

$$(1) \sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu)^2 \quad \text{or} \quad (2) \sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$$

For example, formula (1) can be found in Thompson (1992) and Barnett (1974,1997) and formula (2) in Hedayat and Sinha (1991). Cochran (1953 page 15) presents both formulas and comments on σ^2 . Regarding (1), he states:

This convention has been used by those who approach sampling theory by means of analysis of variance. Its advantage is that most results take a slightly simpler form. Provided that the same notation is maintained consistently, all of the following results are equivalent in either notation.

- Unless otherwise stated, we will be using σ^2 in formula (1) as the finite population variance. Formula (1) is equivalent to:

$$\sigma^2 = \left(\frac{1}{N-1} \right) \left(\sum_{i=1}^N y_i^2 - \frac{\tau^2}{N} \right) = \left(\frac{1}{N-1} \right) \left(\sum_{i=1}^N y_i^2 - N\mu^2 \right)$$

Suppose we have a populations consisting of the following y - values:

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

Consequently, we have the following parameters:

$$N = 5 \quad \tau = 16 \quad \mu = 3.2 \quad \sigma^2 = 6.7 \quad \sigma \approx 2.588$$

where σ^2 is calculated using a divisor of $N - 1$ (instead of N).

2.2 Introduction to Simple Random Sampling

- Situation: The population consists of N sampling units u_1, u_2, \dots, u_N . These units are often labelled simply as $1, 2, \dots, N$.
- Associated with each of the N units is a measurable value related to the population characteristic of interest. Let y_i be the value associated with unit i .
- Sampling designs that are based on planned randomness are called **probability samples**. More formally, the design is determined by assigning a sample probability $P(s)$ to every possible sample s .

- **Simple random sampling** (without replacement) 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**.
- 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.
 - (N choose n) $\binom{N}{n} = \frac{N!}{n!(N - n)!} = \frac{N(N - 1) \cdots (N - n + 1)}{n!}$. This is the number of combinations of n items selected from N distinct items (and the order of selection doesn't matter).
- There are $\binom{N}{n}$ possible SRSs of size n selected from a population of size N .
- If s is any SRS of size n from a population of size N , then $P(s) = 1/\binom{N}{n}$.

2.3 Sample Statistics

- Let y_1, y_2, \dots, y_n be a SRS of y -values taken from a finite population. Then:
 - The **sample mean** is $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.
 - The **sample variance** is $s^2 = \frac{1}{n - 1} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n - 1} \left(\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} \right)$.
 - The **sample standard deviation** s is $\sqrt{s^2}$.
- \bar{y} , s^2 and s are the same formulas used in an introductory statistics course.

2.4 Estimation of μ and τ

- A natural estimator for the population mean μ is the **sample mean** \bar{y} . Because \bar{y} is an estimate of an individual unit's y -value, multiplication by the population size N will give us an estimate $\hat{\tau}$ of the population total τ . That is:

$$\hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \hat{\tau} = \frac{N}{n} \sum_{i=1}^n y_i = N\bar{y}. \quad (2)$$

- $\hat{\mu}$ and $\hat{\tau}$ are **design unbiased**. That is the average values of $\hat{\mu}$ and $\hat{\tau}$ taken over all possible SRS's equal μ and τ , respectively.
- In this class, the term “unbiased” will mean “design unbiased” (unless stated otherwise).

Demonstration of Unbiasedness: Suppose we have a population consisting of the following y -values:

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

which has the following parameters:

$$N = 5 \quad \tau = 16 \quad \mu = 3.2 \quad \sigma^2 = 6.7 \quad \sigma \approx 2.588$$

where σ^2 is calculated using a divisor of $N - 1$ (instead of N).

All Possible Samples and Statistics from Example Population

Sample Number	Unit Labels	y -values	$\sum y_i$	$\hat{\mu} = \bar{y}$	$\hat{\tau} = N\bar{y}$	$\hat{\sigma}^2 = s^2$	$\hat{\sigma} = s$
1	1,2	0,2	2	1	5	2	1.4142
2	1,3	0,3	3	1.5	7.5	4.5	2.1213
3	1,4	0,4	4	2	10	8	2.8284
4	1,5	0,7	7	3.5	17.5	24.5	4.9497
5	2,3	2,3	5	2.5	12.5	.5	0.7071
6	2,4	2,4	6	3	15	2	1.4142
7	2,5	2,7	9	4.5	22.5	12.5	3.5355
8	3,4	3,4	7	3.5	17.5	.5	0.7071
9	3,5	3,7	10	5	25	8	2.8284
10	4,5	4,7	11	5.5	27.5	4.5	2.1213
sum				32	160	67	22.6274
average				3.2	16	6.7	2.26274

The averages for estimators $\hat{\mu}$, $\hat{\tau}$, and $\hat{\sigma}^2$ equal the parameters that they are estimating. This implies that $\hat{\mu}$, $\hat{\tau}$, and $\hat{\sigma}^2$ are unbiased estimators of μ , τ , and σ^2 .

The average for estimator $\hat{\sigma} = s$ does not equal the parameter σ . This implies that $s = \hat{\sigma}$ is a biased estimator of σ .

- The next problem is to study the variances of $\hat{\tau}$ and $\hat{\mu}$.
- Warning: In an introductory statistics course, you were told that the variance of the sample mean $\text{var}(\bar{Y}) = \sigma^2/n$ and its standard deviation is σ/\sqrt{n} . This is appropriate if a sample was to be taken from an infinitely or extremely large population.
- However, we are dealing with finite populations that 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 :

$$\text{var}(\hat{\mu}) = \text{var}(\bar{y}) = \left(\frac{N-n}{N}\right) \frac{\sigma^2}{n} \quad \text{var}(\hat{\tau}) = N(N-n) \frac{\sigma^2}{n} \quad (3)$$

- Because σ^2 is unknown, we use s^2 to get unbiased estimators of the the variances in (3)::

$$\widehat{\text{var}}(\hat{\mu}) = \widehat{\text{var}}(\bar{y}) = \left(\frac{N-n}{N}\right) \frac{s^2}{n} \quad \widehat{\text{var}}(\hat{\tau}) = N(N-n) \frac{s^2}{n} \quad (4)$$

- Taking a square root of a variance in (3) yields the **standard deviation** of the estimator.
- Taking a square root of an estimated variance in (4) yields the **standard error** of the estimate.
- Like $\hat{\mu}$ and $\hat{\tau}$, $\widehat{\text{var}}(\hat{\mu})$ and $\widehat{\text{var}}(\hat{\tau})$ are design unbiased. That is the average values of $\widehat{\text{var}}(\hat{\mu})$ and $\widehat{\text{var}}(\hat{\tau})$ taken over all possible SRS's equal $\text{var}(\hat{\mu})$ and $\text{var}(\hat{\tau})$, respectively.

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

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

which have the following parameters

$$N = 5 \quad \tau = 16 \quad \mu = 3.2 \quad \sigma^2 = 6.7 \quad \sigma \approx 2.588$$

All Possible Samples and Statistics from Example Population

Sample Number	Unit Labels	y -values	$\sum y_i$	$\hat{\mu} = \bar{y}$	$\hat{\tau} = N\bar{y}$	$\hat{\sigma}^2 = s^2$	$\hat{\sigma} = s$	$\widehat{\text{var}}(\hat{\mu})$	$\widehat{\text{var}}(\hat{\tau})$
1	1,2	0,2	2	1	5	2	1.4142	0.6	15
2	1,3	0,3	3	1.5	7.5	4.5	2.1213	1.35	33.75
3	1,4	0,4	4	2	10	8	2.8284	2.4	60
4	1,5	0,7	7	3.5	17.5	24.5	4.9497	7.35	183.75
5	2,3	2,3	5	2.5	12.5	.5	0.7071	0.15	3.75
6	2,4	2,4	6	3	15	2	1.4142	0.6	15
7	2,5	2,7	9	4.5	22.5	12.5	3.5355	3.75	93.75
8	3,4	3,4	7	3.5	17.5	.5	0.7071	0.15	3.75
9	3,5	3,7	10	5	25	8	2.8284	2.4	60
10	4,5	4,7	11	5.5	27.5	4.5	2.1213	1.35	33.75
sum				32	160	67	22.6274	20.1	502.5
average				3.2	16	6.7	2.26274	2.01	50.25

- If N is large relative to n , then the finite population correction will be close to (but less than) 1. Omitting the finite population correction from the variance formulas will (on average) slightly overestimate the true variance. That is, there is a small positive bias. I personally would not recommend omitting the finite population correction (f.p.c.).
- If N is not large relative to n , then omitting the finite population correction from the variance formulas can seriously overestimate the true variance. That is, there can be a large positive bias. In such cases, do not omit the finite correction (f.p.c.).
- As $n \rightarrow N$, $\frac{N-n}{N} \rightarrow 0$. That is, as the sample size approaches the population size, the f.p.c. approaches 0. Thus, in (3) and (4) the variances $\rightarrow 0$ as $n \rightarrow N$.

2.5 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. Suppose 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 from a population of N units is called a **simple random sample with replacement**.
- The estimators for SRS with replacement are

$$\hat{\mu} = \bar{y} \quad \widehat{\text{var}}(\hat{\mu}) = \widehat{\text{var}}(\hat{y}) = \frac{s^2}{n}$$

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

$\hat{\theta}_1$ is **less efficient** than $\hat{\theta}_2$ for estimating θ if $\text{var}(\hat{\theta}_1) > \text{var}(\hat{\theta}_2)$.

$\hat{\theta}_1$ is **more efficient** than $\hat{\theta}_2$ for estimating θ if $\text{var}(\hat{\theta}_1) < \text{var}(\hat{\theta}_2)$.

- 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.

2.6 Two-Sided Confidence Intervals for μ and τ

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

$$\bar{y} \pm z^* \frac{s}{\sqrt{n}} \quad \text{and} \quad \bar{y} \pm t^* \frac{s}{\sqrt{n}} \quad (5)$$

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 that are not considered extremely large, we adjust our variance formulas by the finite population correction .

- Based on a finite population version of the Central Limit Theorem, we assume that the estimators $\hat{\mu}$ and $\hat{\tau}$ have sampling distributions that are approximately normal. That is,

$$\hat{\mu} \sim N\left(\mu, \frac{N-n}{N} \frac{\sigma^2}{n}\right) \quad \text{and} \quad \hat{\tau} \sim N\left(\tau, N(N-n) \frac{\sigma^2}{n}\right)$$

Or, equivalently,

$$\bar{y} \sim N\left(\mu, \frac{N-n}{N} \frac{\sigma^2}{n}\right) \quad \text{and} \quad N\bar{y} \sim N\left(\tau, N(N-n) \frac{\sigma^2}{n}\right)$$

- For large samples, approximate $100(1 - \alpha)\%$ confidence intervals for μ and τ are

For μ : For τ : (6)

$$\begin{aligned} \bar{y} \pm z^* \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}} & \qquad N\bar{y} \pm z^* \sqrt{N(N-n) \frac{s^2}{n}} \\ \bar{y} \pm z^* s \sqrt{\left(\frac{N-n}{N}\right) / n} & \qquad N\bar{y} \pm z^* s \sqrt{N(N-n) / n} \end{aligned} \quad (7)$$

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

$$\hat{\mu} \pm z^* \text{s.e.}(\hat{\mu}) \qquad \hat{\tau} \pm z^* \text{s.e.}(\hat{\tau})$$

- For smaller samples, approximate $100(1 - \alpha)\%$ confidence intervals for μ and τ are

For μ : For τ : (8)

$$\begin{aligned} \bar{y} \pm t^* \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}} & \qquad N\bar{y} \pm t^* \sqrt{N(N-n) \frac{s^2}{n}} \\ \bar{y} \pm t^* s \sqrt{\left(\frac{N-n}{N}\right) / n} & \qquad N\bar{y} \pm t^* s \sqrt{N(N-n) / n} \end{aligned} \quad (9)$$

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

- The quantity being added and subtracted from $\hat{\mu} = \bar{y}$ or $\hat{\tau} = N\bar{y}$ in the confidence interval is known as the **margin of error**.

Example: Using the small population data once again:

All Possible Samples and Confidence Intervals from Example Population

Sample Number	y -values	$\sum y_i$	$\hat{\mu} = \bar{y}$	$\hat{\tau} = N\bar{y}$	$\hat{\sigma}^2 = s^2$	$\hat{\sigma} = s$	$\widehat{\text{var}}(\hat{\mu})$	$\widehat{\text{var}}(\hat{\tau})$	90% ci for τ
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)
sum			32	160	67	22.6274	20.1	502.5	
average			3.2	16	6.7	2.262	2.01	50.25	

2.7 One-Sided Confidence Intervals for μ and τ

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

- For samples of size n approximate upper and lower $100(1 - \alpha)\%$ confidence intervals for μ and τ are:

For μ :	For τ :	
$\left(\bar{y} - t^* s \sqrt{\left(\frac{N-n}{N}\right) / n} , \infty \right)$	$\left(N\bar{y} - t^* s \sqrt{N(N-n)/n} , \infty \right)$	upper
$\left(0 , \bar{y} + t^* s \sqrt{\left(\frac{N-n}{N}\right) / n} \right)$	$\left(0 , N\bar{y} + t^* s \sqrt{N(N-n)/n} \right)$	lower

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

- If the y -values can take on negative values, replace 0 with $-\infty$ in the lower 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.

2.8 Sample Size Determination for μ and τ

- It is well known that an increase in sample size n will lead to a more precise estimator of μ or τ . 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 SRS size n required so the probability that the difference between the sample-based estimator $\hat{\theta}$ and the parameter being estimated θ exceeds some maximum allowable difference $d = |\hat{\theta} - \theta|$ is at most α . Or, equivalently, find n such that

$$\Pr(|\hat{\theta} - \theta| > d) < \alpha \tag{10}$$

- If we assume that $\hat{\theta} \sim N(\theta, \text{var}(\hat{\theta}))$ (that is, $\hat{\theta}$ is unbiased and approximately normal), then we know that the distribution of the standardized estimator is approximately standard normal:

$$\frac{\hat{\theta} - \theta}{\sqrt{\text{var}(\hat{\theta})}} \sim N(0, 1) \quad (11)$$

- Under this assumption, the sample size problem is to find n so that

$$\Pr\left(\frac{|\hat{\theta} - \theta|}{\sqrt{\text{var}(\hat{\theta})}} > z_{\alpha/2}\right) = \Pr\left(|\hat{\theta} - \theta| > z_{\alpha/2}\sqrt{\text{var}(\hat{\theta})}\right) = \alpha \quad (12)$$

- Thus, we need to find n large enough so that the margin of error $z_{\alpha/2}\sqrt{\text{var}(\hat{\theta})} \leq d$.

2.8.1 When Estimating μ

- Situation: Estimate the SRS size required so the probability that the difference between the estimator $\hat{\mu} = \bar{y}$ and the population mean μ does not exceed a maximum allowable difference d is at most α .

– For example, consider the spatially correlated population in Figure 1 on page 18. How large a sample would be required so that $\hat{\mu} = \bar{y}$ is within 1 of μ with probability at least .95 ($\alpha = .05$)? (Assume $\sigma^2 \approx 18.3$)

- In mathematical notation, find n such that $\Pr(|\hat{\mu} - \mu| > d) < \alpha$ for a specified maximum allowable difference d . Or, equivalently, find n so that the margin of error $z_{\alpha/2}\sqrt{\text{var}(\hat{\mu})} \leq d$.

- After substitution, we want to find n so that $z_{\alpha/2}\sqrt{\left(\frac{N-n}{N}\right)\frac{\sigma^2}{n}} \leq d$. Solving this inequality for n yields

$$n = \frac{1}{\frac{d^2}{z^2\sigma^2} + \frac{1}{N}} = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \quad (13)$$

where $n_0 = \frac{z^2\sigma^2}{d^2}$ and z is the critical $\alpha/2$ value from a $N(0, 1)$ distribution.

- Rounding-up the value of n in (13) 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^* .
- 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 σ^2 . Therefore, to estimate the sample size n , we need a prior estimate of σ^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 σ^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 σ^2 .
3. A Preliminary Sample: A preliminary SRS of size n_1 is taken and the sample variance s_1^2 is used to estimate σ^2 . Using s_1^2 in (13) 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 σ^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 ($\mu = \sigma^2$), all we need is a prior estimate of the population mean μ .
 - A second case occurs with estimation of a proportion p for 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 .

2.8.2 When Estimating τ

- Situation: Estimate the SRS size required so the probability that the difference between the estimator $\hat{\tau} = N\bar{y}$ and the population total τ does not exceed a maximum allowable difference d is at most α .
 - For example, consider the longleaf pine population in Figure 2 on page 19. How large a sample would be required so that $\hat{\tau}$ is within 15 of τ with probability at least .95 ($\alpha = .05$)? (Assume $\sigma^2 \approx 4$)
- In mathematical notation, find n such that $\Pr(|\hat{\tau} - \tau| > d) < \alpha$ for a specified maximum allowable difference d . Or, equivalently, find n so that $z_{\alpha/2}\sqrt{\text{var}(\hat{\tau})} \leq d$.

- After substitution, we want to find n so that $z_{\alpha/2}\sqrt{N(N-n)\frac{\sigma^2}{n}} \leq d$. Solving this inequality for n yields

$$n = \frac{1}{\frac{d^2}{N^2 z^2 \sigma^2} + \frac{1}{N}} = \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \quad (14)$$

where $n_0 = \frac{N^2 z^2 \sigma^2}{d^2}$ and z is the critical $\alpha/2$ value from a $N(0, 1)$ distribution.

- Rounding-up the value of n in (14) yields the desired sample size. If this value is < 30 , I recommend adding 2 or 3 to this value.
- If the population size N is very large, then $1/N \approx 0$. In this case, $n \approx n_0$.

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 $\tau = 13354$ and the mean per unit is $\mu = 33.385$. The population variance $\sigma^2 = 75.601$.
- 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

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 Example using Rathbun and Cressie (1994) Data

- To illustrate the application of simple random sampling to population total τ 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 $\tau = 584$ and the mean abundance per quadrat $\mu = 584/400 = 1.435$. The population variance $\sigma^2 = 3.853$.
- The pineleaf census data will be used to compare estimation properties of various sampling designs.

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

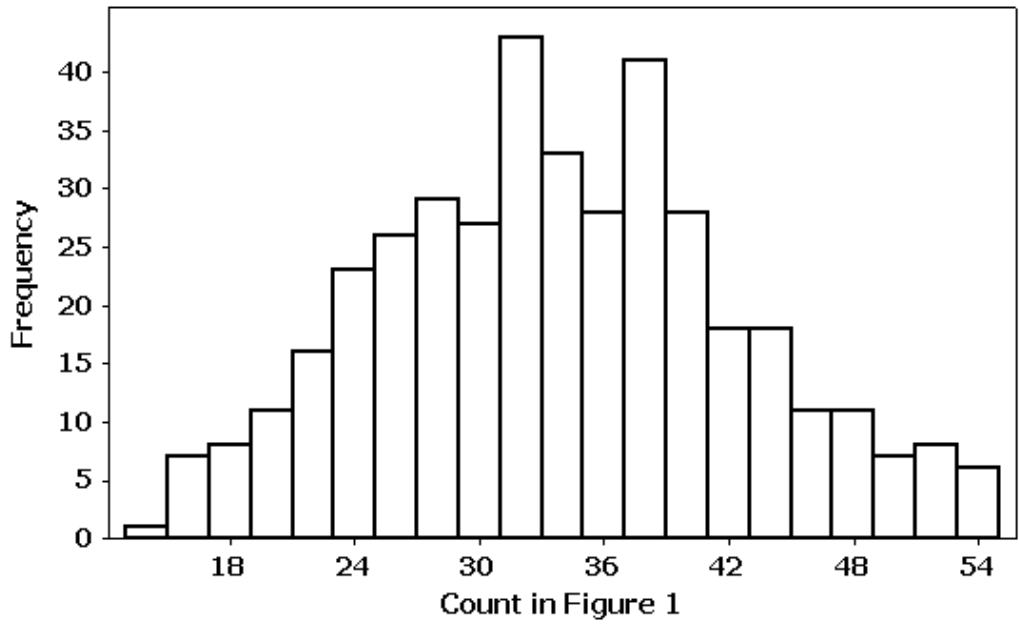
SRS taken from Figure 1 ($n = 10, \tau = 13354, \mu = 33.385, \bar{y} = 34.1, s^2 = 18.3\bar{2}$)

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

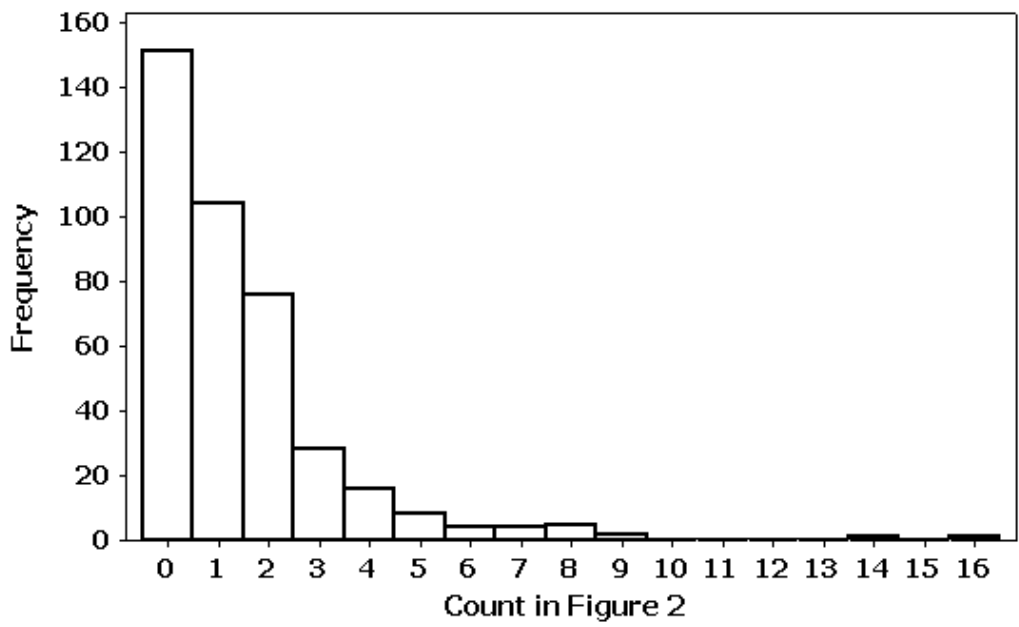
SRS taken from Figure 2 ($n = 20, \tau = 584, \mu = 1.435, \bar{y} = 1.55, s^2 = 10.9974$)

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	1	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

Histogram of Count Data in Figure 1



Histogram Count Data in Figure 2



2.9 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:

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

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

- By taking a SRS of size n , we can estimate p with the **sample proportion** \hat{p} of units that possess that attribute: $\hat{p} = \frac{\sum_{i=1}^n y_i}{n} = \bar{y}$. The sample proportion \hat{p} is unbiased for p .
- For a finite population of 0 and 1 values, the population variance

$$\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - p)^2 = \left(\frac{N}{N-1} \right) p(1-p).$$

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

$$\text{var}(\hat{p}) = \left(\frac{N-n}{N} \right) \frac{\sigma^2}{n} = \left(\frac{N-n}{N} \right) \left(\frac{N}{N-1} \right) \frac{p(1-p)}{n} = \left(\frac{N-n}{N-1} \right) \frac{p(1-p)}{n} \quad (15)$$

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

$$\widehat{\text{var}}(\hat{p}) = \left(\frac{N-n}{N} \right) \frac{s^2}{n} = \left(\frac{N-n}{N} \right) \frac{\hat{p}(1-\hat{p})}{n-1} \quad (16)$$

- The square root of $\text{var}(\hat{p})$ in (15) is the **standard deviation** of the estimator \hat{p} .
- The square root of $\widehat{\text{var}}(\hat{p})$ in (16) is the **standard error** of \hat{p} .
- Omitting the finite population correction (f.p.c.) from the formulas for large and small samples also apply here.

Figure 3: The Presence/Absence of Longleaf Pine

Rathbun/Cressie data ($\tau = 249$ $N = 400$ $p = .6225$)

1	1	1	1	1	1	1	0	0	0	1	1	0	1	0	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 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

2.9.1 Confidence Intervals for p

- Let the random variable X = 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 X follows a *hypergeometric distribution*.
- Let $\Pr(X = j)$ = the probability that the SRS of size n will have j sampling units possessing the attribute. In other words, $\Pr(X = j) = \frac{\binom{\tau}{j} \binom{N-\tau}{n-j}}{\binom{N}{n}}$ = the probability that a SRS will consist of j ones and $n - j$ zeroes selected from the population containing τ ones (1's) and $N - \tau$ zeroes (0's).
- Thompson (pages 40-41) discusses confidence interval calculation based on probability tables of hypergeometric distributions. We will use a more common approach that will apply to many sampling situations.
- Remember there are τ ones and $N - \tau$ zeros in the population. However, τ is unknown. If we can assume that n is small relative to both τ and $N - \tau$, we can use the binomial approximation to the hypergeometric distribution. That is, $X \sim \text{BIN}(n, p)$.
- Although the problem no longer depends on τ , it still depends on the unknown parameter p .
- What is commonly done is to apply the normal approximation to the binomial distribution:

$$\hat{p} \sim N(p, \text{var}(\hat{p})).$$

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

$$\hat{p} \pm z^* \sqrt{\widehat{\text{var}}(\hat{p})} \quad \text{OR} \quad \hat{p} \pm z^* \sqrt{\left(\frac{N-n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}} \quad (17)$$

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 τ or $N - \tau$. 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 ≥ 5 , while some texts use ≥ 10 .

2.9.2 Sample Size 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 α .
 - 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?
- In mathematical notation, find n such that $\Pr(|\hat{p} - p| > d) \leq \alpha$ for a specified maximum allowable difference d . Or, equivalently, find n so that the margin of error $z_{\alpha/2} \sqrt{\widehat{\text{var}}(\hat{p})} \leq d$.

- After substitution, we want to find n so that $z_{\alpha/2} \sqrt{\left(\frac{N-n}{N-1}\right) \frac{p(1-p)}{n}} \leq d$.
- Solving this inequality for n yields

$$n = \frac{Np(1-p)}{(N-1)\frac{d^2}{z^2} + p(1-p)} = \frac{1}{\frac{N-1}{Nn_0} + \frac{1}{N}} \approx \frac{1}{\frac{1}{n_0} + \frac{1}{N}} \quad (18)$$

where $n_0 = \frac{z^2 p(1-p)}{d^2}$ and z is the critical $\alpha/2$ value from a $N(0, 1)$ distribution.

- Rounding-up the value of n in (18) 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 $n \approx n_0$.
- 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 $\hat{p} = \sqrt{\left(\frac{N-n}{N-1}\right) \frac{p(1-p)}{n}}$ is largest when $p = 1/2$. Thus, it is conservative to use $p = 1/2$ in (18) if there is no prior reasonable estimate.
- Example: Consider the longleaf pine presence/absence population in Figure. 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 $p \approx$ based on the earlier SRS with $n = 25$ (ii) Assume we have no prior estimate of p and use the conservative estimate of $p = .5$.

2.10 Using SAS PROC Surveymeans for SRS data

```
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 Mean mu, Total tau from the data in Figure 1 (page 21)';
run;
```

```
=====
                               Simple Random Sample -- Example 1
Estimating Mean mu, Total tau from the data in Figure 1 (page 21)
```

The SURVEYMEANS Procedure

Data Summary

Number of Observations	10
Sum of Weights	400

Statistics

Variable	Mean	Std Error 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 Mean mu, Total tau from the data in Figure 2 (page 21)';
run;

```

=====

Simple Random Sample -- Example 2
Estimating Mean mu, Total tau from the data in Figure 2 (page 21)

The SURVEYMEANS Procedure

Data Summary

Number of Observations	20
Sum of Weights	400

Statistics

Variable	Mean	Std Error 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

```

DM 'LOG;CLEAR;OUT;CLEAR';
OPTIONS NODATE NONUMBER;

data SRS_Fig3;
    input ind @@;
    datalines;
1 0 1 0 1 1 0 0 1 1 1 1 1 0 1 1 1 1 0 0 1 1 1 1 1
;

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 ;
    var pa;
title1 'Simple Random Sample -- Example 3';
title2 'Estimating Proportion p from the SRS data in Figure 3 (page 23)';

run;

```

=====

Simple Random Sample -- Example 3
 Estimating Proportion p from the SRS data in Figure 3 (page 23)

The SURVEYMEANS Procedure

Data Summary

Number of Observations 25

Class Level Information

Class Variable	Levels	Values
pa	2	absent present

Statistics

Variable	Level	N	Mean	Std Error of Mean	95% CL for Mean
pa	absent	7	0.280000	0.088741	0.09684717 0.46315283
	present	18	0.720000	0.088741	0.53684717 0.90315283
