

Simple Random Sampling (Chapter 2)

Simple random sampling (SRS) is a sampling design where n units are selected (without replacement) from a population of N units, such that all samples of size n are equally likely to be selected. First, though, we introduce some general ideas and terminology.

Population Notation:

Finite Population Values: y_1, \dots, y_N

Population Mean: $\mu = \frac{1}{N} \sum_{i=1}^N y_i$

Population Total: $\tau = \sum_{i=1}^N y_i$

Population Median: $M = \text{median}(y_1, \dots, y_N)$

Population Variance: $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^N (y_i - \mu)^2$

Sample: y_1, \dots, y_n , where $n =$ sample size ($n \leq N$).

Definition: An estimator (or statistic) is a function of the sample values.

Let: $\theta =$ a population parameter (e.g.: μ).

$\hat{\theta} =$ an estimator of θ (e.g.: \bar{y}).

Definition: The sampling distribution of $\hat{\theta}$ refers to the distribution of values of $\hat{\theta}$ for all possible samples of size n from the population. The sampling distribution depends on the sampling plan being used.

Bias, Variance, and MSE of Estimators:

- Definition: The bias of an estimator $\hat{\theta}$ of θ is given by:

$$\text{Bias}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

An estimator of θ is unbiased if its bias is 0 for all values of θ , that is, if $E(\hat{\theta}) = \theta$.

Note: Unbiasedness is a property of the estimator, not of the sampling plan.

- $\text{Var}(\hat{\theta}) = E\left[(\hat{\theta} - E(\hat{\theta}))^2\right]$ = a measure of the precision of an estimator, often used to compare estimators when they are unbiased or when the bias is small.
- The mean squared error (MSE) of an estimator is:

$$\begin{aligned} \text{MSE}(\hat{\theta}) &= E\left[(\hat{\theta} - \theta)^2\right] = \\ (\text{accuracy}) &= \end{aligned}$$

- The MSE incorporates both the bias and the precision of an estimator into a measure of overall accuracy and can be used to compare estimators whether they are unbiased or not.

- Many standard estimators are unbiased or nearly unbiased so the variance of estimators is the most common means of comparison.

Example: Consider a population of size $N = 4$, with $y_1 = 6$, $y_2 = 4$, $y_3 = 16$, & $y_4 = 10$.

Population Mean $\mu = 36/4 = 9$

Population Variance $\sigma^2 = \frac{1}{3} [(6 - 9)^2 + (4 - 9)^2 + (16 - 9)^2 + (10 - 9)^2] = \frac{1}{3}(84) = 28$.

- Consider the following comparison of simple random sampling and stratified random sampling for samples of size 2 from this population.

1. Take an SRS of size $n = 2$: Possible samples? We want to estimate μ with \bar{y} .

	Sample	\bar{y}	Prob.
$E(\bar{y}) = \frac{1}{6}(5 + \dots + 13) = \frac{54}{6} = 9$, so \bar{y} is an unbiased estimator of μ .			
$\text{Var}(\bar{y}) = \frac{1}{6}(5 - 9)^2 + \dots + \frac{1}{6}(13 - 9)^2 = \frac{42}{6} = 7$,			
$\text{SD}(\bar{y}) = \sqrt{\text{Var}(\bar{y})} = 2.65$.			

2. Take a stratified random sample: Suppose we have 2 strata: 6, 4|16, 10. Does this seem like a good choice of strata?

- Suppose we take a sample of size 1 from each stratum, and compute the stratified estimator of the population mean: $\bar{y}_s = \frac{1}{2}(y_1 + y_2)$. Possible samples?

	Sample	\bar{y}_s	Prob.
$E(\bar{y}_s) = \frac{1}{4}(11 + 8 + 10 + 7) = 9$ (also unbiased for μ)			
$\text{Var}(\bar{y}_s) = \frac{1}{4} [(11 - 9)^2 + \dots + (7 - 9)^2] = 2.5$			
$\text{SD}(\bar{y}_s) = \sqrt{2.5} = 1.58$.			

- Since both estimators (SRS and stratified) are unbiased, but \bar{y}_s has the smaller variance, a stratified random sample is better here.
- Suppose instead that the 2 strata were defined as: 4, 16|6, 10.

	Possible Samples	\bar{y}_s	Prob.
$E(\bar{y}_s) = 9$, $\text{Var}(\bar{y}_s) = 10$, and $\text{SD}(\bar{y}_s) = 3.16$,	(4,6)	5	1/4
which is worse than the SRS in this case. Why?	(4,10)	7	1/4
	(16,6)	11	1/4
	(16,10)	13	1/4

- In the 1st set of strata, we have low variability within strata (homogeneous within) & high variability between strata (heterogeneous between).
- In the 2nd set of strata, we have high variability within strata (heterogeneous within) & low variability between strata (homogeneous between).
- This example illustrates the importance of choosing strata which are heterogeneous between and homogeneous within. It should also serve as a warning that stratification should not be used in sampling unless there are clear reasons for believing the defined strata are quite different from one another. In other words, don't stratify without a good reason!

Back to Simple Random Sampling (N = population size, n = sample size)

Estimating the Mean

- We know that the population mean μ is estimated by \bar{y} . What is the variance $\text{Var}(\bar{y})$?

- If $N \gg n$, then the fpc ≈ 1 and is unnecessary. In this case, we revert to the usual formula for the variance of the sample mean: $\text{Var}(\bar{y}) = \sigma^2/n$.
- If $N = n$ (as in a census), then $\text{Var}(\bar{y}) =$
- The derivation of the finite population corrected variance is somewhat tedious and is covered on pages 21-22 of the text.
- The standard deviation of the sample mean is given by: $\text{SD}(\bar{y}) = \sqrt{\left(\frac{N-n}{N}\right) \frac{\sigma^2}{n}}$.
What is the problem with this as an estimator of variability?
- The population variance σ^2 is unknown, so we need to estimate it. This results in the estimated standard deviation of \bar{y} , more commonly referred to as the standard error of \bar{y} , given by:

$$\text{SE}(\bar{y}) = \widehat{\text{SD}}(\bar{y}) = \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}} \quad \text{where} \quad s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

Estimating the Total: Estimate $\tau = \sum_{i=1}^N y_i$ with $\hat{\tau} = \frac{N}{n} \sum_{i=1}^n y_i = N\bar{y}$.

$\text{Var}(\hat{\tau}) =$

- In typical problems, we report the sample mean \bar{y} & the standard error $\text{SE}(\bar{y})$, and form a confidence interval (CI) for μ as:

$$(\text{Estimator}) \pm \begin{pmatrix} \text{Critical} \\ \text{Value} \end{pmatrix} \begin{pmatrix} \text{SE of} \\ \text{Estimator} \end{pmatrix} \implies \bar{y} \pm t \cdot \text{SE}(\bar{y}) \implies \bar{y} \pm t \cdot \sqrt{\left(\frac{N-n}{N}\right) \frac{s^2}{n}}.$$

- Why is this interval t-based and not standard normal (z)-based?
- What assumptions are we are making in the use of this CI?
 - 1.
 - 2.
- If the first of these is violated, we can appeal to the Central Limit Theorem (CLT) to obtain an approximate $(1 - \alpha) \times 100\%$ CI.

Usual Central Limit Theorem (CLT): If y_1, \dots, y_n are iid random variables with mean μ and variance $\sigma^2 < \infty$, then: $\bar{y} \sim N(\mu, \sigma^2/n)$ as n gets large.

- This version of the CLT is for an infinite population, where sampling with and without replacement are the same.

Finite Population CLT: If y_1, \dots, y_n are an SRS without replacement from a finite population, then:

$$\bar{y} \sim N\left(\mu, \text{Var}(\bar{y}) = \left(\frac{N-n}{N}\right) \frac{\sigma^2}{n}\right) \text{ as both } n \text{ and } N-n \text{ get large.}$$

Example: Suppose a study is undertaken to estimate the number of pellet groups for white-tailed deer in a 20-acre field. Sampling was done along 10 randomly located belt transects of dimensions 3'x50', where the number of pellet groups in each belt was counted.

- What is the sampling unit here?
- Suppose the following summary statistics were obtained: $\bar{y} = 5.55$ groups, $s^2 = 14.06$, $s = 3.75$ groups, where:

μ = the mean number of pellet groups in a 3'x50' transect,

τ = the total number of pellet groups in the 20 acres (1 acre = 43560 square feet).

- The goal here is to find an estimate $\hat{\tau}$ of the total number of pellet groups using a confidence interval. Computing:

$\bar{y} = 5.55$ per 150 square feet \implies there are 5.55/150 pellet groups per square foot.

$$\implies \hat{\tau} = \underbrace{\frac{5.55}{150}}_{\left(\begin{array}{l} \# \text{ groups per} \\ \text{square foot} \end{array} \right)} \cdot \underbrace{43560}_{\left(\begin{array}{l} \# \text{ square feet} \\ \text{per acre} \end{array} \right)} \cdot \underbrace{20}_{\# \text{ acres}} = \underline{32234.4 \text{ groups}}$$

$$\implies N = \frac{43560 \cdot 20}{150} = \underline{5808} \text{ (150 square foot areas in the 20-acre field).}$$

- Note: This is not an infinite population because the transects have area; there are, however, an infinite number of possible transects.
- The standard error of the sample mean number of groups per transect is:

$$SE(\bar{y}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{s^2}{n}} = \sqrt{\underbrace{\left(1 - \frac{10}{5808}\right)}_{.9983 \approx 1} \frac{14.06}{10}} = \underline{1.186}.$$

- This gives 95% confidence intervals for the mean and total as:

$$95\% \text{ CI for } \mu : \quad \bar{y} \pm t_{.025}(9) \cdot SE(\bar{y}) = 5.55 \pm 2.262(1.186) = \underline{(2.87, 8.23)}.$$

$$95\% \text{ CI for } \tau : \quad (2.87 \cdot 5808, 8.23 \cdot 5808) = \underline{(16655, 47813)}.$$

Estimating Proportions (Ch. 5)

Let $y_i = \left\{ \begin{array}{l} 1 \text{ if the } i^{\text{th}} \text{ unit has some characteristic} \\ 0 \text{ otherwise} \end{array} \right\}$. Define:

$$p = \frac{1}{N} \sum_{i=1}^N y_i = \text{the population proportion of units with some characteristic,}$$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n y_i = \text{the sample proportion, with:}$$

$$\sigma^2 = \frac{N}{N-1} p(1-p) = \text{the population variance,}$$

$$s^2 = \frac{n}{n-1} \hat{p}(1-\hat{p}) = \text{the sample variance.}$$

- Note that p & \hat{p} are defined exactly as μ & \bar{y} were for these y_i 's. So, defined as above, proportions can be represented as means.

The variance and estimated variance of \hat{p} follow immediately then from the forms of $\text{Var}(\bar{y})$ and $\widehat{\text{Var}}(\bar{y})$ given earlier in the handout:

$$\underline{\text{Var}(\hat{p})} = \left(\frac{N-n}{N}\right) \frac{\sigma^2}{n} = \left(\frac{N-n}{N}\right) \frac{N}{n(N-1)} p(1-p) = \underline{\left(\frac{N-n}{N-1}\right) \frac{p(1-p)}{n}},$$

$$\underline{\widehat{\text{Var}}(\hat{p})} = \left(\frac{N-n}{N}\right) \frac{s^2}{n} = \underline{\left(\frac{N-n}{N}\right) \frac{\hat{p}(1-\hat{p})}{n-1}}.$$

This estimated variance allows us to construct a confidence interval for p of the form:

$$\hat{p} \pm z \sqrt{\frac{N-n}{N} \left(\frac{\hat{p}(1-\hat{p})}{n-1} \right)}.$$

- Is it valid to use such an interval? When?
- Exact confidence limits based on the hypergeometric distribution are given on page 41 of the text.

Design-Based vs. Model-Based Sampling: In everything we have looked at thus far, we have taken a design-based or fixed-population approach to sampling. In other words, *no distributional assumptions were placed on the population* in finding the form of the estimators and their variances. In the model-based approach to sampling, the “population” y_1, \dots, y_N are considered to be random variables, that is, one possible realization of all possible realizations that could have occurred under some model for the population.

For example, suppose I flip a (possibly biased) coin 100 times and record the result of each toss on a slip of paper and place the slips in a box. I ask you, who don’t know the results, to estimate the proportion of the 100 flips that came out heads. You are allowed to take a sample of slips from which to make your estimate. This is a design-based approach – the 100 flips is the population and the parameter of interest is the proportion of heads, say θ , in these 100 flips. If you base your estimate on a random sample of slips of paper, the only uncertainty in your estimate is due to sampling variability. If you could do a census, there there would be no uncertainty in your estimate. If you use the the sample proportion of heads as your estimator, then this estimator will be unbiased for θ and you will be able to generate an unbiased estimate of its variance. These properties depend only on the sample design – they don’t depend on whether I really flipped a coin 100 times or simply wrote down 100 0’s and 1’s in any combination or order I wanted.

In the model based approach, we view these 100 flips as 100 realizations from a random process, and the parameter we are really interested in is not the proportion of heads in these 100 flips, but the true long-term probability of heads. Our model might then be that the 100 flips are 100 independent Bernoulli trials with probability of heads p . Even if I observed all 100 flips, I would still not know p . The difference between the two approaches doesn’t really make a difference in how I would estimate θ or p : I would use the proportion of heads in whatever sample I observed. It would make a difference in the standard error of the estimate as the population is considered infinite in the model based approach. It also has implications for how I sample. If my model is correct – that these 100 flips are 100 independent Bernoulli trials – then it doesn’t matter if I observe a random sample of flips or not. If I get to observe 10 flips, then the first 10 are just as good as any other set of 10. I simply view these 10

flips as 10 independent Bernoulli random variables; the randomness (and unbiasedness) in my estimator comes from the model, not the sampling scheme. However, if my model is not correct and these are not independent Bernoulli trials with constant probability of heads (from example, there is dependence between trials, or the probability of heads decreases as I go along), then my estimator based on the first 10 flips might not be very good (biased, large variance) and I won't realize it. In fact, the parameter p might not even make sense.

Design-based approaches are valid regardless of how the data were generated, but the scope of inference is confined to the fixed population. The model-based approach allows for inference to a larger population or model, but depends crucially on the appropriateness of the model.