Lecture 3 - Key

Types of Probability Samples

A **simple random sample (SRS)** of size n is *a sample of n units selected from a population in such a way that every sample of size n has an equal chance of being selected.*

- A sample of twenty MSU instructors is selected based on the following scheme: *An alphabetical list of all the instructors is prepared, then a unique number is assigned in sequence (1,2,3,...) to each instructor, and finally, using a random number generator, twenty instructors are randomly chosen.*
- STAT 446 students are placed in groups for lab sessions. Students are ordered alphabetically and using a random number generator, the numbers are randomly selected to comprise the first group.

A **stratified random sample** is a sample selected by first dividing the population into non-overlapping groups called **strata** and then taking a simple random sample within each stratum. *Dividing the population in to strata should be based on some criterion so that units are similar with a stratum but are different between strata.*

- The career services staff wants to know if they are adequately meeting the needs of the students they serve. *They design a survey which addresses this question, and they send the survey to 50 students randomly chosen from each of the universities departments.*
- A biologist is interested in estimating a deer population total in a small geographic region. The region contains two habitat types which are known to influence deer abundance. From each habitat type, 10 plots are randomly selected to be surveyed.
- Note: stratum sample sizes do not have to be equal.

A **systematic sample** is a sample in which units are selected in a *"systematic" pattern in the population of interest. To take a systematic sample you will need to divide the sampling frame into groups of units, randomly choose a set of starting points in the first group, and then sample from every group using the same positions of the starting points.*

- You are a quality engineer at Gore and are testing the quality of newly-produced mittens. You need to take a sample of mittens and test their quality. *As the mittens roll of the production line, you decided to test every 50^{th} mitten starting with the third mitten (i.e. sample chips 3, 53,103,...)*

Suppose the observation units in a population are grouped into non-overlapping sets called **clusters.** *A **cluster sample** is a SRS of clusters.*

- You work for the Department of Agriculture and wish to determine the percentage of farmers in the United States who use organic farming techniques. It would be difficult and costly to collect a SRS of systematically sample because both of these sampling designs would require visiting many individual farms that are located far from each other. A convenience sample of farmers from a single county would be biased because farming practices vary from region to region. You decide to select several dozen counties from across the United States and then survey every farmer in each of these selected counties. Each county contains a cluster of farmers and data is collected on every farm within the randomly selected counties (clusters).

A **multistage sample** is a sample acquired by *successively selecting smaller groups within the population in stages. The selection process at any stage may employ any sampling design.*

- A city official is investigating rumors about the landlord of a large apartment building complex. To get an idea of the tenants' opinions about their landlord, the official takes a SRS of buildings in the complex followed by a SRS of apartments from each selected building. From each chosen apartment a resident is interviewed.

Probability Sampling Designs

Suppose N is the population size. That is, there are N units in the universe or finite population of interest. The N units in the universe are denoted by an index set of labels: $\mathcal{U} = \{1, 2, 3, ..., N\}$.

From this universe (or population) a sample of n units is to be taken. Let $\mathcal{S}$ represent a sample of n units from $\mathcal{U}$.

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, and the population of y-values is $\{y_1, .y_2, ..., y_n\}$.

Sampling designs that are based on planned randomness are called **probability samples**, and a probability $P(\mathcal{S})$ is assigned to every possible sample $\mathcal{S}$.

The probability that unit i will be included in a sample is denoted π_i and *is called the inclusion probability for unit i.*

Suppose you have a sampling frame with four units and you randomly select two units to sample with a SRS.

- What is the probability of unit # 1 being included in the sample?

4 choose 2, possibilities of which three have unit 1 or $1 - \left(\frac{3}{4} \times \frac{2}{3}\right) = \frac{1}{2}$ *or with simulation.*

Simulation or more specifically Monte Carlo procedures can also be used to estimate inclusion probabilities.

```
library(dplyr)
set.seed(09042019)
samples <- replicate(10, sample(4,2, replace = F))
samples
```

```
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    3    2    2    3    3    3    3    1    2     4
## [2,]    1    1    4    1    2    4    2    3    1     3
```

```
find_val <- function(samples, val){
  return(val %in% samples)
}
samples_large <- replicate(1000, sample(4,2, replace = F))

apply(samples_large, 2, find_val,val = 1 ) %>% mean()
```

```
## [1] 0.475
```

Inference from Samples

One goal of sampling is to draw conclusions about a population of interest based on the data collected. This process of drawing conclusions is called **statistical inference**.

A **parameter** is a value which describes some characteristic of a population (or possible describes the entire population).

A **statistic** is a value that can be computed from the sample data without making use of any unknown parameters.

Unless the statistic and parameter are explicitly stated, we will use $\hat{\theta}$ *and* θ *to represent the unspecified statistic and parameter of interest, respectively.*

In general, the value of the population parameter is unknown. Statistics computed from sampling data can provide information about the unknown parameter.

Common statistics of interest: Let $y_1, y_2, ..., y_n$ be a sample of y-values.

- The **sample mean** is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- The **sample variance** is $s^2 = \frac{1}{n-1}[(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + ... + (y_n - \bar{y})^2] = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{n-1}\left[\sum y_i^2 - \frac{1}{n}(\sum y_i)^2\right]$

Common parameters of interest:

Let t be the population total and $\bar{y}_U$ be the population mean from a population of finite size N. Thus, $t = \sum_{i=1}^{N} y_i$ and $\bar{y}_u = \frac{1}{N}\sum_{i=1}^{N} y_i$.

- The **population variance** parameter S^2 is defined as:

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y}_U)^2 \\ &= \left(\frac{1}{N-1}\right)\left(\sum_{i=1}^{N} y_i^2 - \frac{t^2}{N}\right) = \left(\frac{1}{N-1}\right)\left(\sum_{i=1}^{N} y_i^2 - N\bar{y}_U^2\right) \end{aligned}$$

In some textbooks, τ, μ, and σ^2 are used to represent the population total, mean, and variance.

Because only part of the population is sampled in any sampling plan, the value of $\hat{\theta}$ *will vary in repated random sampling. The inherent variability of* $\hat{\theta}$ *associated with sampling is called* ***sampling variability***.

The **sampling distribution** of a statistic $\hat{\theta}$ *is the probability distribution of the values that can be observed for the statistic over all possible samples for a given sampling scheme.*

The **expected value** of $\hat{\theta}$, denoted $E[\hat{\theta}]$ *is the mean of the sampling distribution of* $\hat{\theta}$: $E[\hat{\theta}] = \sum_{\mathcal{S}} \hat{\theta}_{\mathcal{S}} P(\mathcal{S})$

The **estimation bias** of the estimator $\hat{\theta}$ for estimating a parameter θ *is the numerical difference between* $E[\hat{\theta}]$ *and the parameter value* θ*. That is,* $Bias[\hat{\theta}] = E[\hat{\theta}] - \theta$.

An estimator $\hat{\theta}$ is unbiased to estimate a parameter θ if $Bias[\hat{\theta}] = 0$.

So far the focus has been on the expected value of a statistic to check for bias. *Another natural concern is the variability of the statistic. It is certainly possible for an unbiased statistic to have large variability.*

We will consider two measures of variability: the variance and the mean square error.

- The **variance** of the sampling distribution of $\hat{\theta}$ ($V(\hat{\theta})$) is defined to be

$$V(\hat{\theta}) = E\left[(\hat{\theta}_{\mathcal{S}} - E(\hat{\theta})^2\right] = \sum_{\mathcal{S}} P(\mathcal{S})\left[\hat{\theta}_{\mathcal{S}} - E(\hat{\theta})\right]^2$$

  where $\hat{\theta}_{\mathcal{S}}$ is the value of $\hat{\theta}$ calculated for sample $\mathcal{S}$.

- The **mean squared error** is defined to be :

$$MSE[\hat{\theta}] = E\left[(\hat{\theta} - \theta)^2\right] = \sum_{\mathcal{S}} P(\mathcal{S})\left[(\hat{\theta}_{\mathcal{S}} - \theta)\right]^2$$

- The MSE, however, can be rewritten as

$$MSE[\hat{\theta}] = V(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

  Derivation may be a homework problem.

The idea of a sampling distribution is foundational to statistics as a whole.

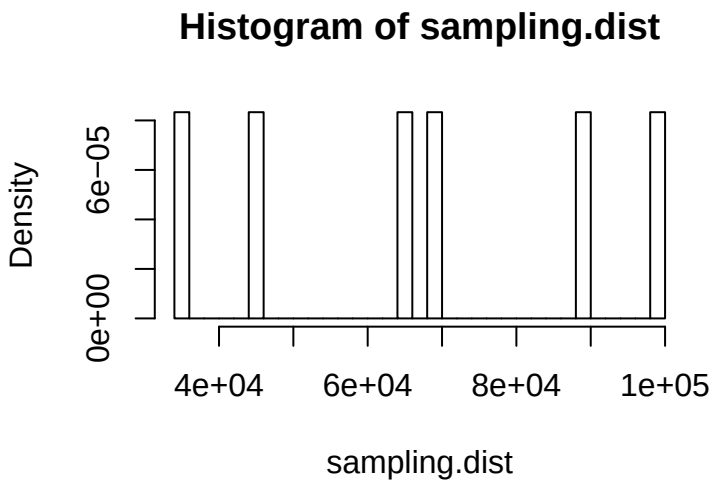
Consider the previous example of taking a SRS of size two from a population of four. Now assume the goal is to estimate the average income of the four units from a sample of size two. Let the salaries be $80,000, $120,000, $60,000, and $10,000.

Using the sample mean, what are the possible values for θ?

```
vals <- c(80000, 120000, 60000, 10000)
sampling.dist <- c(mean(vals[1:2]), mean(vals[c(1,3)]), mean(vals[c(1,4)]),
                   mean(vals[c(2,3)]), mean(vals[c(2,4)]), mean(vals[c(3,4)]))
```

Sketch out the sampling distribution, note: this will be a discrete distribution.

```
hist(sampling.dist, breaks = 25, prob = T)
```



Is the SRS unbiased?

```
sum(1/6 *sampling.dist) - mean(vals)
```

```
## [1] 0
```

What is the MSE of this estimator?

```
sum(1/6 * (sampling.dist - mean(sampling.dist))^2)
```

```
## [1] 522916667
```

Now consider a simulation approach

```
samples <- sample(sampling.dist, 10000, replace = T) # a  bit of a shortcut
mean(samples) - mean(vals)
```

```
## [1] 90
```

```
var(samples)
```

```
## [1] 522004100
```