Sampling Activity: Gettysburg's Address

Background

Sampling is often an cost-effective method to gain information about a population, compared to conducting a Census.

It is critical when sampling to utilize a sampling method that minimizes any potential bias.

- **Statistical Inference** is the process of using data from a sample to gain information about a population.

Background

Sampling is often an cost-effective method to gain information about a population, compared to conducting a Census.

It is critical when sampling to utilize a sampling method that minimizes any potential bias.

- **Statistical Inference** is the process of using data from a sample to gain information about a population.
- **Sampling Bias** occurs when the method for selecting a sample causes the sample's properties to not reflect the population's properties.

Background

Sampling is often an cost-effective method to gain information about a population, compared to conducting a Census.

It is critical when sampling to utilize a sampling method that minimizes any potential bias.

- **Statistical Inference** is the process of using data from a sample to gain information about a population.
- **Sampling Bias** occurs when the method for selecting a sample causes the sample's properties to not reflect the population's properties.
- If **Sampling Bias** is present and **Statistical Inferences** are potentially erroneous.

Activity

Your objective is to use a sample to estimate the **population mean** for number of letters (i.e., word length) in each word spoken in the Gettysburg Address.

- What is the **mean** word length for the following population of words?
  {Dog, Cat, Horse, Iguana, Parrot}

Activity

Your objective is to use a sample to estimate the **population mean** for number of letters (i.e., word length) in each word spoken in the Gettysburg Address.

- What is the **mean** word length for the following population of words?
  {Dog, Cat, Horse, Iguana, Parrot}
- $\frac{3+3+5+6+6}{5} = 4.6$

Gettysburg Address

Four score and seven years ago our fathers brought forth upon this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this.

But, in a larger sense, we can not dedicate – we can not consecrate – we can not hallow – this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before

Representative Sampling via Random Sampling

- A **representative sample** resembles the population, only in smaller numbers.

Representative Sampling via Random Sampling

- A **representative sample** resembles the population, only in smaller numbers.
- Random sampling tends to produce a representative sample.

Representative Sampling via Random Sampling

- A **representative sample** resembles the population, only in smaller numbers.
- Random sampling tends to produce a representative sample.
- Random sampling avoids sampling bias.

Representative Sampling via Random Sampling

- A **representative sample** resembles the population, only in smaller numbers.
- Random sampling tends to produce a representative sample.
- Random sampling avoids sampling bias.
- How do we take a random sample of words from the Gettysburg Address?

Random Sampling in R

```
speech_sample <- sample_n(speech, 10)
speech_sample
```

```
##      ID       Word Length
## 1   165        the      3
## 2   216      which      5
## 3   215        for      3
## 4   170 dedicated      9
## 5   110  dedicate      8
## 6    20        and      3
## 7   142       will      4
## 8   171       here      4
## 9    18         in      2
## 10 157       what      4
```

```
mean(speech_sample$Length)
```

```
## [1] 4.5
```