Table Number:_____   Group Name: _____

Group Members:_____   _____   _____

Gettysburg Address

One of the most important ideas in statistics is that we can learn a lot about a large group (called a population) by studying a small piece of it (called a sample). Consider the population of 268 words in the following passage:

> *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 battlefield 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 cannot dedicate, we cannot consecrate, we cannot 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 us, that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion, that we here highly resolve that these dead shall not have died in vain, that this nation, under God, shall have a new birth of freedom, and that government of the people, by the people, for the people, shall not perish from the earth.*

(a) Select a sample of ten *representative* words from this population by circling them in the passage above.

The authorship of several literary works is often a topic for debate. Were some of the works attributed to William Shakespeare actually written by Francis Bacon or Christopher Marlowe? Which of the anonymously published Federalist Papers were written by Alexander Hamilton, which by James Madison, which by John Jay? Who were the authors of the writings contained in the Bible? The field of "literary computing" began to find ways of numerically analyzing authors' works, looking at variables such as sentence length and rates of occurrence of specific words.

The above passage is, of course, Lincoln's Gettysburg Address, given November 19, 1863 on the battlefield near Gettysburg, PA. In characterizing this passage, we could have asked you to examine every word. Instead, we asked you to look at a sample of the words of the passage. We are considering this passage a **population** of words, and the 10 words you selected are considered a **sample** from this population. In most studies, we do not have access to the entire population and can only consider results for a sample from that population. The goal is to learn something about a very large population (e.g., all American adults, all American registered voters) by studying a sample. The key is in carefully selecting the sample so that the results in the sample are **representative** of the larger population (i.e., has the same characteristics).

> The **population** is the entire collection of observational units that we are interested in examining. A **sample** is a subset of observational units from the population. Keep in mind that these are objects or people, and then we need to determine what variable we want to measure about these entities.

(b) Do you think the ten words in your sample are representative of the 268 words in the population? Explain briefly.

(c) Record the length for each of the ten words in your sample:

Word	1	2	3	4	5	6	7	8	9	10
# letters										

(d) Determine the average (mean) number of letters in your sample (the ten words):_____

(e) List the average (mean) number of letters of the samples of each of the other two groups at your table and also at one neighboring table. List those averages (means) here:

_____ _____ _____ _____ _____ _____

(f) The population average number of letters for all 268 words is 4.295 letters. Were most of the samples' means *near* the population mean? Explain.

(g) For how many groups did the sample average exceed the population average? What proportion of the six groups is this?

(h) Explain why this sampling method (asking people to choose ten "representative "words) is **biased** and how this bias is exhibited. Also identify the *direction* of the bias. In other words, does the sampling method tend to overestimate or underestimate the average length of the words in the passage?

A **simple random sample** (SRS) gives every observational unit in the population the same chance of being selected. In fact, it gives every sample of size *n* the same chance of being selected. In this example we want every set of ten words to be equally likely to be the sample selected. While the principle of simple random sampling is probably clear, it is by no means simple to implement.

(i)   Use the app in the link below to generate 6 random samples of size 10, one at a time, and write the mean of each sample here:

_____ _____ _____ _____ _____ _____

   You'll need to click on "Show Sampling Options" ☑   in order to choose the number of samples and the sample size.   Choose one at a time and note the mean of each sample.

Show Sampling Options: ☑

Number of Samples: | 1

Sample Size: | 10

Click on "Draw Sample." The mean will be listed under **Most Recent Sample.** (You can ignore the **Statistic** histogram for now.)

http://www.rossmanchance.com/applets/OneSample.html?population=gettysburg

(j)   Compare the results of part (j) to part (e). Comment on the randomness and reliability of your samples in part (e) to those in part (j).

(k)   Compare the means you obtained from the 6 samples from part (e) and part (j) to the population mean. Justify and explain the differences you observe from part (e) and part (j).

Other Sampling Methods

(l)   A **systematic sample** is obtained by selecting every k^{th} individual from the population. The first individual selected corresponds to a random number between 1 and *k.*

   i.   Choose an arbitrary value for *k* and write it here:_____

   ii.   Use the random number generator on your calculator to choose a value between 1 and *k.* Suppose this value is 4. You will then start with the 4th word in the Gettysburg address and choose every k^{th} word for your sample.

   iii.   Underline each word in your systematic sample.

   iv.   Find the mean of your sample and write it here: _____

   v.   How does this mean compare to the population mean?

(m)   A **stratified sample** is obtained by separating the population into non-overlapping groups called *strata* and then obtaining a simple random sample from each stratum. The individuals within each stratum should be similar in some way.
How might you collect a stratified sample of ten words from the Gettysburg address?
Write your plan here:

(n)   A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.
How might you collect a cluster sample of ten words from the Gettysburg address? Write your plan here: