The authorship of literary works is often a topic for debate. Were some of the works attributed to Shakespeare actually written by Bacon or Marlowe? Which of the anonymously published *Federalist Papers* were written by Hamilton, which by Madison, which by Jay? Who were the authors of the writings contained in the Bible? The field of “literary computing” examines ways of numerically analyzing authors’ works, looking at variables such as sentence length and rates of occurrence of specific words.

The passage below is, of course, Abraham Lincoln’s Gettysburg Address, given November 19, 1863 on the battlefield near Gettysburg, PA. In characterizing this passage, we would ideally examine every word. However, often it is much more convenient and even more efficient to only examine a subset of words. In this case, you will examine data for just 10 of the words. We are considering this passage to be a *population* of 268 words, and the 10 words you select are a *sample* from this population.

1. Circle 10 representative words in the following passage.
2. Consider the following variables:
   length of word (number of letters)  Type of variable:
   whether or not the word contains more than 4 letters  Type of variable:
   whether or not the word is a noun  Type of variable:

   Record the data from your sample for the above variables:

<table>
<thead>
<tr>
<th>Word</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&gt; 4 characters?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>noun?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

   Calculate the average length of your words and your sample proportions (of long words and of nouns.)

3. Do you think the words you selected are representative of the population of 268 words in this passage? Explain.

   Recall, a parameter is a numerical characteristic of the population. A statistic is a numerical characteristic of the sample.

4. Is the average length you calculated above a parameter or a statistic? Explain. What symbol do we use to denote this value?

5. The average length of all 268 words in this population is 4.29 letters. Is this number a parameter or a statistic? What symbol do we use to denote this value?

6. There are 99 “long” words in this population of 268 words. What proportion of the words in the population are “long”? Is this a parameter or a statistic?

7. Did everyone in the class obtain the same value for the sample mean? What about the sample proportion of long words?

8. Were our estimates consistently above or below the parameter values?
9. Based on the class results would you say that this sampling method (asking people to find 10 words) is likely to produce a sample that is truly representative of the population with respect to the length of the words? Explain.

10. Repeat #2 for 10 randomly selected words:

Record the data from your sample for the above variables:

<table>
<thead>
<tr>
<th>Word</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length &gt; 4 characters?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>noun?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Calculate the average length of your words and your sample proportions (of long words and of nouns.)

11. After random sampling, did everyone in the class obtain the same values for the sample means and sample proportions?

12. After random sampling, were our estimates consistently above or below the parameter values?
Write a short paper (3-5 pages, double spaced) on sampling techniques. Use the random words activity and the Literary Digest articles to support your claims. Address the following:

- Discuss the concept of selection bias. Convince me that you understand why it is important to have a sample that is representative of the population (and how bad sampling methods can give us a sample that is not representative of the population.)

- Read the articles on the Literary Digest poll. Alf Landon was predicted to beat FDR by a 3-to-2 margin. Instead, FDR won with 62% of the votes. The articles discuss two different poor sampling methods. Describe each and comment on which you think was more influential in the regretful events.

- Think of a hypothesis (question) you might want to test (ask.) Who would you ask and how would you go about getting a representative sample from your population? (Be sure to define who your population is and who your sample is.) What might some limitations of the sampling be? What would be the best type of sampling procedure? Would this be possible? Difficult? (You can answer this question either from the perspective of a student or from the perspective of a researcher who has grant funding to use on sampling.)

For your sample, list as many possible biases as you can think of, and label them as selection, response, or non-response bias.

Your essay will be graded on the clarity of your explanations and the soundness of your arguments as well as the organization and flow of the paper. This should not be a draft, but rather the result of at least one re-write.