# 3.1 SELECTING SAMPLES FROM POPULATIONS I

## INVESTIGATION 3.1.1 SAMPLING WORDS

**a.** Circle 10 representative 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.

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 is of course Abraham Lincoln's Gettysburg Address, given November 19, 1863 on the battlefield near Gettysburg, Pennsylvania. In characterizing this passage, we would ideally examine every word. However, often it is much more convenient and even more efficient to examine only 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 selected are therefore a *sample* from this population.

In most statistical studies, we do not have access to the entire population and can only consider data for a sample from that population. For example in Chapter 1, we examined only 116 current popcorn plant workers—not all previous workers, and not all plants; in Chapter 2, the sleep researchers performed their experiment on only 21 students rather than all students at their university. Our ultimate goal is to make conclusions about a larger population or overall process, even when we have access to only a sample from the population.

**b.** Consider the following variables:

Length of word (number of letters) Type: _____

Whether or not the word contains more than four letters Type: _____

Identify each variable as quantitative or categorical.

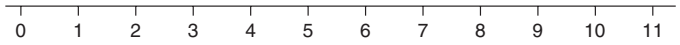
Record the data from your sample for the these variables:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| Word | | | | | | | | | | |
| Length<br>> 4 characters? | | | | | | | | | | |

Ideally, we want our sample to be *representative* of the population, that is, having the same characteristics.

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

**d.** Construct a dotplot of the distribution of the word lengths in your sample. Also calculate the sample mean length and describe the characteristics of this distribution. (Remember to label your plot and to relate your comments to the context.) What are the observational units in this graph?




To display the distribution of a categorical variable for just one sample, we can construct a *bar graph*. This graph has a separate bar for each category, with heights corresponding to the proportion of observational units in that category.

**e.** Construct a bar graph for the categorical variable (whether or not the word is "long"—more than four characters) for your sample and describe its distribution. (Remember to label your plot and to relate your comments to the context.) What are the observational units in this graph?

**DEFINITION** A *parameter* is a numerical characteristic of the population. A *statistic* is a numerical characteristic of the sample. We usually denote population parameters with Greek letters, for example, $\pi$ or $\mu$ for a proportion or mean, respectively. We denote the statistics for a sample proportion and a sample mean by $\hat{p}$ and $\bar{x}$, respectively.

**f.** Is the average length you calculated in (d) a parameter or a statistic? Explain. What symbol do we use to denote this value?

**g.** Is the proportion of long words you calculated in (e) a parameter or a statistic? Explain. What symbol do we use to denote this value?

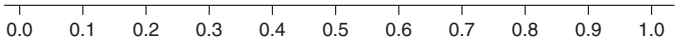
**h.** The average length of all 268 words in this population is 4.29 letters. Is this number a parameter or a statistic? What symbol do we use to denote this value?

**i.** There are 99 "long" words in this population of 268 words. What proportion of the words in the population are "long"? Is this number a parameter or a statistic? What symbol do we use to denote this value?

**j.** Did everyone in your class obtain the same value for the sample mean? The same sample proportion?

**k.** Construct a dotplot or histogram combining the average length of words in your sample with those of your classmates. Be sure to label the horizontal axis, and also indicate where the population mean falls. What are the observational units in this graph? Describe the distribution of these sample means, particularly with regard to where the population average falls.


**l.** Construct a dotplot combining the proportion of long words in your sample with those of your classmates. Be sure to label the horizontal axis, and also indicate where the population proportion falls. What are the observational units in this graph? Describe the distribution of these sample proportions, particularly with regard to where the population proportion falls.




You have witnessed the fundamental principle of *sampling variability*: Values of sample statistics *vary* when one repeatedly takes random samples from a population. A key point in analyzing these results is that now we are treating the samples as the observational units and the sample statistics as the variable of interest (so the first graph label should be something like "sample means" and the second label should be something like "sample proportions"). Many statistical methods are based on describing the pattern of variation of the sample results. Looking at the resulting pattern of variation in sample statistics is also one way to determine if a sampling method is reasonable.

**m.** Was your sample average word length, $\bar{x}$, above or below the population mean $\mu$? How many and what proportion of students in your class found a sample mean word length that exceeded the population mean?

**n.** Was your sample proportion of long words, $\hat{p}$, above or below the population proportion $\pi$? How many and what proportion of students in your class found a sample proportion of long words that exceeded the population proportion?

**o.** Based on your answers to (k)–(n), would you say that this sampling method (asking people to circle 10 words) is likely to produce a sample that is truly representative of the population with respect to the length of the words? Explain.

**DEFINITION** When characteristics of the resulting samples are systematically different from characteristics of the population, we say that the sampling method is *biased*. When the distribution of the sample statistics is centered at the value of the population parameter, the sampling method is said to be *unbiased*.

For example, we suspect that your class repeatedly and consistently overestimated the average length of a word and the proportion of long words. Not everyone has to overestimate, but if there is a *tendency* to err in the same direction time and time again, then the sampling method is biased. In other words, sampling bias is evident if we repeatedly draw samples from the population and the distribution of the sample statistics is not centered at the population parameter of interest. Note that bias is a property of a sampling *method*, not of a single sample. Studies have shown that human judgment is not a good basis for selecting representative samples, so we will rely on other techniques to do the sampling for us.

**p.** Consider another sampling method: You close your eyes and point at the passage on page 00 and select whatever word your pen lands on. Would this sampling method be biased? If so, in which direction? Explain.

**Discussion** Even though the method in (p) sounds "random," it is also going to be biased because longer words take up more space on the page and so will be more likely to be selected than shorter words. Once again, there will be a tendency to oversample the longer words and thus to repeatedly overestimate the average length and the proportion of long words.

**q.** Suggest a better method for selecting a sample of 10 words from this population.

**DEFINITION** A *simple random sample* 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. So any set of 10 words is equally likely to end up as our sample.

"Low-tech" methods for obtaining a simple random sample from a population include using a *random number table* or a calculator, but we will use Minitab. A random number table is constructed so that each position is equally likely to be filled by any digit from 0 to 9, and the digit in one position is unaffected by the digit in any other position. The first step is to obtain a list of every member of your population (this list is called a *sampling frame*). Then, give each observational unit on the list a unique ID number.

The following is a sampling frame for the Gettysburg address, with each word in the population numbered.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 Four | 35 in | 69 dedicate | 103 But, | 137 add | 171 here | 205 these | 239 that |
| 2 score | 36 a | 70 a | 104 in | 138 or | 172 to | 206 honored | 240 this |
| 3 and | 37 great | 71 portion | 105 a | 139 detract. | 173 the | 207 dead | 241 nation, |
| 4 seven | 38 civil | 72 of | 106 larger | 140 The | 174 unfinished | 208 we | 242 under |
| 5 years | 39 war, | 73 that | 107 sense, | 141 world | 175 work | 209 take | 243 God, |
| 6 ago, | 40 testing | 74 field | 108 we | 142 will | 176 which | 210 increased | 244 shall |
| 7 our | 41 whether | 75 as | 109 cannot | 143 little | 177 they | 211 devotion | 245 have |
| 8 fathers | 42 that | 76 a | 110 dedicate, | 144 note, | 178 who | 212 to | 246 a |
| 9 brought | 43 nation, | 77 final | 111 we | 145 nor | 179 fought | 213 that | 247 new |
| 10 forth | 44 or | 78 resting | 112 cannot | 146 long | 180 here | 214 cause | 248 birth |
| 11 upon | 45 any | 79 place | 113 consecrate, | 147 remember, | 181 have | 215 for | 249 of |
| 12 this | 46 nation | 80 for | 114 we | 148 what | 182 thus | 216 which | 250 freedom, |
| 13 continent | 47 so | 81 those | 115 cannot | 149 we | 183 far | 217 they | 251 and |
| 14 a | 48 conceived | 82 who | 116 hallow | 150 say | 184 so | 218 gave | 252 that |
| 15 new | 49 and | 83 here | 117 this | 151 here, | 185 nobly | 219 the | 253 government |
| 16 nation: | 50 so | 84 gave | 118 ground. | 152 but | 186 advanced. | 220 last | 254 of |
| 17 conceived | 51 dedicated, | 85 their | 119 The | 153 it | 187 It | 221 full | 255 the |
| 18 in | 52 can | 86 lives | 120 brave | 154 can | 188 is | 222 measure | 256 people, |
| 19 liberty, | 53 long | 87 that | 121 men, | 155 never | 189 rather | 223 of | 257 by |
| 20 and | 54 endure. | 88 that | 122 living | 156 forget | 190 for | 224 devotion, | 258 the |
| 21 dedicated | 55 We | 89 nation | 123 and | 157 what | 191 us | 225 that | 259 people, |
| 22 to | 56 are | 90 might | 124 dead, | 158 they | 192 to | 226 we | 260 for |
| 23 the | 57 met | 91 live. | 125 who | 159 did | 193 be | 227 here | 261 the |
| 24 proposition | 58 on | 92 It | 126 struggled | 160 here. | 194 here | 228 highly | 262 people, |
| 25 that | 59 a | 93 is | 127 here | 161 It | 195 dedicated | 229 resolve | 263 shall |
| 26 all | 60 great | 94 altogether | 128 have | 162 is | 196 to | 230 that | 264 not |
| 27 men | 61 battlefield | 95 fitting | 129 consecrated | 163 for | 197 the | 231 these | 265 perish |
| 28 are | 62 of | 96 and | 130 it, | 164 us | 198 great | 232 dead | 266 from |
| 29 created | 63 that | 97 proper | 131 far | 165 the | 199 task | 233 shall | 267 the |
| 30 equal. | 64 war. | 98 that | 132 above | 166 living, | 200 remaining | 234 not | 268 earth. |
| 31 Now | 65 We | 99 we | 133 our | 167 rather, | 201 before | 235 have | |
| 32 we | 66 have | 100 should | 134 poor | 168 to | 202 us, | 236 died | |
| 33 are | 67 come | 101 do | 135 power | 169 be | 203 that | 237 in | |
| 34 engaged | 68 to | 102 this. | 136 to | 170 dedicated | 204 from | 238 vain, | |

**r.** Open an empty Minitab worksheet. Since the largest ID number is 268, we will place the integers 1–268 in column 1:

```
MTB> set c1                    places the integers 1–268 into column 1
DATA> 1:268
DATA> end
```

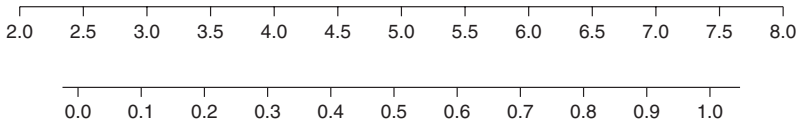
Then take a random sample of five ID numbers and store them in C2:

```
MTB> sample 5 c1 c2            selects a simple random sample of 5 items from column 1
                               and places them into column 2
```

**s.** Match these ID numbers with the corresponding words in the sampling frame. Write down resulting ID numbers, the corresponding words, the length of each word, and whether or not it is long.

| | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| ID number | | | | | |
| Word | | | | | |
| Length | | | | | |
| > 4 characters? | | | | | |

**t.** Calculate the average word length and the proportion of "long" words in this sample of 5 words. Again combine your results with your classmates to produce a dotplot of the sample means and a dotplot on the sample proportions. Comment on each distribution. How do these distributions compare to the ones in (k) and (l)?




**u.** When taking random samples, did everyone obtain a sample mean equal to the population mean? Did everyone in your class obtain the same sample mean in the random samples? In what way would you say these random samples produce "better" sample results?

**v.** What proportion of your class obtained a sample average that was larger than the population mean? What proportion of your class obtained a sample proportion that was larger than the population proportion? Does this sampling method appear to be biased? Explain.

**w.** Is it plausible that the sampling method of randomly selecting just 5 words will have less bias than the sampling method of selecting 10 words you used in (a)? Explain.

## Summary

Even with randomly drawn samples, sampling variability still exists (*random sampling errors*). Although not every random sample produces the same characteristics as the population, the goal is for the sample statistics to not consistently overestimate the value of the parameter or consistently underestimate the value of the parameter. You should have found that when the samples are randomly selected, you will not have the same pattern of consistently overestimating the value of the population parameter. A *simple random sample* eliminates bias by giving every sample of size $n$ the same probability—$1/C(N, n)$—of being selected. When the distribution of the sample statistics is centered at the value of the population parameter, the sampling method is said to be *unbiased*. When the sampling method is unbiased, we expect the individual samples to have characteristics similar to those of the population. Thus, when we select just one sample, we will feel comfortable generalizing the results from that one sample to the larger population. If the sampling method is biased, we can make no claims about the population parameter.

Furthermore, a second virtue of random samples is that we will see that sampling variability follows a predictable, long-term pattern. In particular, with random samples we will be able to estimate how far the sample statistic is likely to fall from the population parameter—that is, the size of the *random sampling error*—and what factors affect the size of this random sampling error.

In this example, we knew the value of the population parameter, but that is not usually the case. Thus, it is very important to determine whether or not the sample was selected at random before you believe that the sample results are representative of the population.

Keep in mind that *random sampling*, as you have done here, and *randomization*, as you studied in Chapters 1 and 2, are different uses of randomness. Now we are focusing on selecting a subset from a larger population and making inferences back to that population instead of trying to compare two groups as you did earlier.

Notice that the ideas of a parameter and a statistic are closely related; one describes the population and one describes the sample. For example, the average length of words in the population is a parameter, but the average length of words in the sample is a statistic. A crucial observation is that the value of a statistic will vary from sample to sample (we are now thinking of the statistic as a variable but with the samples as the observational units!). But as we saw, when we take random samples, this variation follows a very predictable pattern. Keep in mind that although the value of a statistic varies from sample to sample, a parameter does not vary.

## Practice Problems

### 3.1.1 Literary Digest

In 1936, *Literary Digest* magazine conducted the most extensive (to that date) public opinion poll in history. The editors mailed out questionnaires to over 10 million people, whose names and addresses they obtained from telephone books and vehicle registration lists. More than 2.4 million responded, with 57% indicating that they would vote for Republican Alf Landon in the upcoming presidential election.