Activity: Sampling Words (sampling bias, sampling distribution of sample mean)
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 above passage is, of course, Lincoln's Gettysburg Address. For this activity 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).

b) Record the word and the number of letters in each of the ten words in your sample:

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

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

d) Create a dotplot of your sample results (number of letters in each word). Also indicate what the observational units and variable are in this dotplot. Is the variable categorical or quantitative?

Dotplot:

Observational units: Variable: Type:

e) Determine the average (mean) number of letters in your ten words.

f) Combine your sample average with the rest of the class to produce a well-labeled dotplot.

g) Indicate what the "observational units" and "variable" are in the dotplot in (f). (*Hint*: To identify what the observational units are, ask yourself what each dot on the plot represents. The answer is different from above.)

One conceptual challenge here is realizing that the observational units are no longer the individual words but rather the *samples* of ten words. Each dot in this plot comes from a *sample* of ten words, not from an *individual* word.

h) The average number of letters per word in the population of all 268 words is 4.295. Mark this value on the dotplot in (d). How many students produced a sample average greater than the actual population average? What proportion of the students is this?

When the sampling method produces characteristics of the sample that systematically differ from those characteristics of the population, we say that the sampling method is **biased**.

i) Would you say that this sampling method (asking people to simply circle ten representative words) is biased? If so, in which direction? Explain how you can tell this from the dotplot.

j) Suggest some reasons why this sampling method turned out to be biased as it did.

k) Consider a different sampling method: close your eyes and point to the page ten times in order to select the words for your sample. Would this sampling method also be biased? Explain.

Most people tend to choose larger words, perhaps because they are more interesting or convey more information. Therefore, the first sampling method is *biased* toward *over*estimating the average number of letters per word in the population. Some samples may not overestimate this population value, but samples chosen with this method *tend to* overestimate the population mean. Closing your eyes does not eliminate the bias, because you are still more likely to select larger words, because they take up more space on the page. In general, human judgment is not very good at selecting representative samples from populations.

l) Would using this same sampling method but with a larger sample size (say, 20 words) eliminate the sampling bias? Explain.

m) Suggest how you might employ a different sampling method that would be unbiased.

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.

Although the principle of simple random sampling is probably clear, it is by no means simple to implement. The first step is to obtain a **sampling frame** where each member of the population can be assigned a number. Here we just need to number the words in the above passage. This sampling frame appears below.

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,	

Then use an applet (e.g., random.org or **Random Number Generator** applet) to generate **5** unique numbers between 1 and 268.

n) Record the ID numbers that you selected, the corresponding words, and the lengths of the words:

	1	2	3	4	5
ID number					
Word					
Word length					

o) Determine the average length in your sample of five words.

p) Again combine your sample mean with those of your classmates to produce a dotplot below. Be sure to label the horizontal axis appropriately.

q) Comment on how the distribution of sample averages from these random samples compares to that from your "circle ten words" samples.

r) Do the sample averages from the random samples tend to over- or under-estimate the population average, or are they roughly split evenly on both sides?

To really examine the long-term patterns of this sampling method, we will use technology to take many, many samples.

Open the **Sampling Words** applet. The first box shows you the variables for the Gettysburg Address (including length of words, whether or not the word contains at least one letter *e*, and whether or not the word is a noun). The histogram below shows the population distribution for the lengths of words.

- Check the **Show Sampling Options** box.
- Specify 5 as the **Sample Size** and press **Draw Samples**. Note the lengths of the words (which are displayed in blue in the dotplot below) and the average for the sample of 5 words (which is also added to the Statistic dotplot to the right).
- Then press **Draw Samples** again to add a second sample mean to the Statistic dotplot.
- Then change the **Number of Samples** from 1 to 98 and press the **Draw Samples** button. The applet now takes 98 more simple random samples from the population (for a total of 100 so far) and adds the sample averages to the Statistic graph in the lower right panel. The mean for these 100 sample means is reported in the top left corner of that graph.

s) What does this graph reveal? (*Hint*: If you select the Population Scale, it will make the scaling match the original population graph.)

t) Now change the **Sample Size** from 5 to 10 press **Draw Samples**. Does the sampling method still appear to be unbiased? How are you deciding? What has changed about the type of sample averages that we obtain? Why does this make sense?

Once we have a representative sampling method, we can improve the **precision** by increasing the sample size. With larger random samples, the results will tend to fall even closer to the population results.

Three caveats about random sampling are in order:

- One still gets the occasional "unlucky" sample whose results are not close to the population even with large sample sizes.
- Second, the sample size means little if the sampling method is not random. In 1936 the *Literary Digest* magazine had a huge sample of 2.4 million people, yet their predictions for the Presidential election did not come close to the truth about the population.
- Although the role of *sample size* is crucial in assessing how close the sample results will be to the population results, the size of the *population* does not affect this. As long as the population is large relative to the sample size (at least 10 times as large), the precision of a sample statistic depends on the sample size but not on the population size!

u) How could you convince yourself of the last bullet point?

Founded in 1890, the *Literary Digest* was a popular magazine in the early 1900s. The magazine had correctly predicted the outcomes of the 1916, 1920, 1924, 1928, and 1932 presidential elections by conducting polls. With the Great Depression in full swing, the magazine ventured forth in 1936 to predict another presidential election outcome. Questionnaires were mailed out to more than 10 million adult Americans whose names and addresses were obtained from subscribers to the magazine and also from phone books and vehicle registration lists. More than 2.4 million responses were received, the largest survey that had ever been undertaken at that time, with 57% indicating that they planned to vote for Republican challenger Alf Landon over Democrat incumbent Franklin Roosevelt.

v) Suggest some reasons why the *Literary Digest*'s sampling method was biased.

An up-and-coming pollster named George Gallup made his first election prediction in 1936, the same year as the infamous *Literary Digest* poll. Gallup surveyed a sample of only 50,000 adult Americans and was able to predict the election result correctly. The secret to Gallup's success was random sampling.

Possible extensions:

Other statistics options include percentile (like for the Flint water data) and sample standard deviation.

Activity: Sleepless Nights? (inference for single mean, *t*-test, 6-step process)

Step 1: Ask a Research Question
How much do students at your school sleep on a typical night? Is the average less than the recommended either hours? How can we estimate this average?

1) Based on these questions, what is the population of interest? What is the parameter? What symbol do we use for this parameter?

Step 2: Design a study and collect data
2) How much sleep (to the nearest quarter hour) did you get last night?

3) Are students in this class a random sample of all students at your school? Can you identify a potential source of sampling bias?

Step 3: Explore the data
4) In the **Sampling Words** applet, press **Clear** and copy in the sleep data from your class and press **Use Data**. Use the displayed graph to comment on the shape, center, and variability of this distribution. Are there any unusual observations that merit additional exploration? Is this distribution the sample or the population?

5) Is the sample mean lower than 8 hours? Could this have happened by random chance alone?

Step 4: Draw inferences beyond the data
6) If students at your school tended to get 8 hours of sleep on a typical night, what does that imply about the population mean? Write this as a null hypothesis (using appropriate symbols).

7) What is our conjecture about the population mean? Write this as an alternative hypothesis.

We want to apply the same 3S strategy to analyze the strength of evidence provided by our sample against the null hypothesis that the population mean is equal to eight.

<u>1. Statistic</u>
8) What do you suggest for the choice of a statistic? What was the observed value of the statistic in our sample?

2. Simulate

9) Describe a process for simulating values of the statistic assuming the null hypothesis is true.

Although we want to use the same logic as we did with categorical data, things are a bit more complicated with quantitative data. Namely, we don't have access to the entire population to sample from or a complete description of the process. We can make up a population to sample from, but we have to make some assumptions first. In addition to assuming $\mu = 8$, we also need to assume the population shape and the population standard deviation.

Approach 2: Sampling from a finite population

a) Select **Sleep1** from the list of populations to display a distribution of a population of 18,000 students, assuming the null hypothesis is true. What are the mean and standard deviation of this population of sleep times?

b) Check the **Show Sampling Options** box, specify 1000 random samples, and use the sample size from our class. Describe the shape, center, and variability of the distribution of sample means. How do these compare to the shape, center, and variability of the population?

c) Use the **Count Samples** box to estimate the probability of obtaining a sample mean as small or smaller than the one we found in class.

d) Repeat (a)-(c) using **Sleep2**. Summarize what you learn about how the distribution of sample means changes when we change the shape of the population.

Approach 1: Sampling from a theoretical population

Follow the Population model link.

Approach 3: Bootstrapping

Follow the Bootstrapping link.

3. Strength of evidence

e) Based on the p-values you simulated, is there strong evidence that the population mean sleep time of students at your school is less than 8 hours? Explain the reasoning behind your conclusion.

Approach 4. Theory-based approach

Just like we had a Central Limit Theorem to predict the distribution of sample proportions, the sample mean statistic has some very specify properties:

- It is an unbiased estimator of the population mean. In other words, one could show mathematically that the average of all possible sample means from a population will be equal to the population mean
- The standard deviation decreases with the sample size. In fact, if we know the population standard deviation σ, then one could show mathematically that the standard deviation of the sample means will equal $\sigma/\sqrt{n}$.
- The shape of the distribution of sample means will either be normal, if the population distribution itself is normal, or approximately normal if the sample size is large, regardless of the shape of the population!

This last statement is the "central limit theorem" for sample means. It allows us to assume the distribution of sample means is normal, even if the population distribution is not, as long as the sample size is large. How large the sample size needs to be depends on how nonnormal the original population is (which we often don't know!) but many consider 20 as a reasonable guideline, as long as the sample doesn't exhibit way severe skewness or outliers.

f) Discuss how the above simulation results agree or disagree with this theory.

g) Because we don't typically know σ, it is common to use *s* instead and calculate the *standard error* of the sample mean as $s/\sqrt{n}$. Calculate this value for our observed sample.

h) Using your answer to (g) as the estimate of the sample to sample variation in the sample mean, construct a rough 95% confidence interval for the population mean. Interpret this interval in context.

i) Does the interval in (h) include the value of 8 hours? Did you expect this based on your earlier analyses? What additional information is provided by this confidence interval?

Step 5. Formulate conclusions

j) Think again about how the sample was selected from the population. Do you feel comfortable generalizing the results of your analysis to the population of all students at your school? Explain.

Step 6. Look back and ahead

k) What are some variables that could be useful in helping us predict the amount of sleep by students in this population?