Activity 1: Sampling Words

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 this dotplot. [*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 (f). How many students produced a sample average greater than the actual population average? What proportion of the students is this?

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. One approach is to use a computer-generated **table of random digits**. Such a table is constructed so that each position is equally likely to be occupied by any one of the digits 0-9, and so that the value of any one position has no impact on the value of any other position.

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 on the next page, and a table of random digits appears on the page after that.

You will now use the table to random digits to select a simple random sample of *five* words from the Gettysburg address. Do this by entering the table at any point (it does not have to be at the beginning of a line) and reading off three-digit numbers between 001 and 268. (Disregard any numbers not in this range. If you happen to get repeats, keep going until you have five different two-digit numbers. If you finish a line without obtaining five words, just continue on to the next line.) Continue until you have five numbers corresponding to words in this population.

(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.

Sampling frame:

00	Four	03	In	06	dedicate	10	But	13	add	17	here	20	These	23	that
01	Score	05	A	07	A	10	in	13	or	17	to	20	Honored	23	this
02	And	06	great	07	portion	10	a	13	detract	17	the	20	Dead	23	nation
03	Seven	07	Civil	07	Of	10	larger	13	The	17	unfinishe	20	We	24	under
04	Years	08	War	07	That	10	sense	13	world	17	work	20	Take	24	God
06	Ago	09	testing	07	Field	10	we	14	will	17	which	20	increase	24	shall
06	Our	09	whether	07	As	10	cannot	14	little	17	they	20	Devotio	24	have
07	Fathers	04	That	07	A	10	dedicate	14	note	17	who	21	To	24	a
08	brought	02	nation	07	final	10	we	14	nor	17	fought	21	That	24	new
09	Forth	03	Or	07	resting	11	cannot	14	long	19	here	23	Cause	24	birth
00	Upon	04	Any	07	place	12	consecrat	14	rememb	18	have	24	For	24	of
01	This	05	nation	09	For	13	we	14	what	18	thus	25	Which	25	freedom
02	continent	06	So	08	those	14	cannot	14	we	18	far	26	They	25	and
03	A	07	conceive	08	who	15	hallow	15	say	18	so	27	Gave	25	that
04	New	08	And	08	here	16	this	15	here	18	nobly	28	The	22	governme
05	Nation	09	So	08	gave	17	ground	15	but	16	advance	29	Last	25	of
06	conceive	05	dedicate	08	their	18	The	15	it	16	d	20	Full	24	The
07	d	05	Can	08	lives	19	brave	15	can	17	is	21	Measur	25	People
08	Liberty	05	Long	06	That	10	men	15	never	18	rather	22	Of	26	By
09	And	05	endure	07	That	11	living	15	forget	19	for	23	Devotio	27	The
00	dedicated	05	We	08	nation	12	and	15	what	19	us	24	That	25	People
01	To	05	Are	09	might	13	dead	15	they	19	to	25	We	29	For
02	The	06	Met	09	Live	14	who	15	did	19	be	26	Here	26	The
03	propositio	07	On	09	It	15	struggled	10	here	19	here	27	Highly	26	People
04	that	08	A	02	Is	16	here	10	It	19	dedicate	28	Resolve	22	Shall
05	All	09	great	09	altogeth	17	have	16	is	19	to	29	That	26	Not
06	Men	06	battlefiel	04	fitting	18	consecrat	12	for	19	the	20	These	26	Perish
02	Are	06	Of	09	And	19	ed	16	us	17	great	23	Dead	26	From
08	created	02	That	09	proper	10	far	16	the	19	task	22	Shall	26	The
09	Equal	06	War	07	That	13	above	16	living	20	remainin	23	Not	27	Earth
00	Now	06	We	09	We	12	our	16	rather	20	before	23	Have	8	
03	We	05	have	19	should	13	poor	16	to	20	us	25	Died		
02	Are	06	come	10	Do	14	power	16	be	20	that	26	In		
03	engaged	07	To	10	This	15	to	17	dedicate	20	from	27	Vain		
4		8		2		6		0	d	4		8			

A SAMPLE PAGE FROM THE TABLE OF RANDOM DIGITS

00000	10097	32533	76520	13586	34673	54876	80959	09117	39292	74945
00001	37542	04805	64894	74296	24805	24037	20636	10402	00822	91665
00002	08422	68953	19645	09303	23209	02560	15953	34764	35080	33606
00003	99019	02529	09376	70715	38311	31165	88676	74397	04436	27659
00004	12807	99970	80157	36147	64032	36653	98951	16877	12171	76833
00005	66065	74717	34072	76850	36697	36170	65813	39885	11199	29170
00006	31060	10805	45571	82406	35303	42614	86799	07439	23403	09732
00007	85269	77602	02051	65692	68665	74818	73053	85247	18623	88579
00008	63573	32135	05325	47048	90553	57548	28468	28709	83491	25624
00009	73796	45753	03529	64778	35808	34282	60935	20344	35273	88435
00010	98520	17767	14905	68607	22109	40558	60970	93433	50500	73998
00011	11805	05431	39808	27732	50725	68248	29405	24201	52775	67851
00012	83452	99634	06288	98083	13746	70078	18475	40610	68711	77817
00013	88685	40200	86507	58401	36766	67951	90364	76493	29609	11062
00014	99594	67348	87517	64969	91826	08928	93785	61368	23478	34113
00015	65481	17674	17468	50950	58047	76974	73039	57186	40218	16544
00016	80124	35635	17727	08015	45318	22374	21115	78253	14385	53763
00017	74350	99817	77402	77214	43236	00210	45521	64237	96286	02655
00018	69916	26803	66252	29148	36936	87203	76621	13990	94400	56418
00019	09893	20505	14225	68514	46427	56788	96297	78822	54382	14598
00020	91499	14523	68479	27686	46162	83554	94750	89923	37089	20048
00021	80336	94598	26940	36858	70297	34135	53140	33340	42050	82341
00022	44104	81949	85157	47954	32979	26575	57600	40881	22222	06413
00023	12550	73742	11100	02040	12860	74697	96644	89439	28707	25815
00024	63606	49329	16505	34484	40219	52563	43651	77082	07207	31790
00025	61196	90446	26457	47774	51924	33729	65394	59593	42582	60527
00026	15474	45266	95270	79953	59367	83848	82396	10118	33211	59466
00027	94557	28573	67897	54387	54622	44431	91190	42592	92927	45973
00028	42481	16213	97344	08721	16868	48767	03071	12059	25701	46670
00029	23523	78317	73208	89837	68935	91416	26252	29663	05522	82562
00030	04493	52494	75246	33824	45862	51025	61962	79335	65337	12472
00031	00549	97654	64051	88159	96119	63896	54692	82391	23287	29529
00032	35963	15307	26898	09354	33351	35462	77974	50024	90103	39333
00033	59808	08391	45427	26842	83609	49700	13021	24892	78565	20106
00034	46058	85236	01390	92286	77281	44077	93910	83647	70617	42941
00035	32179	00597	87379	25241	05567	07007	86743	17157	85394	11838
00036	69234	61406	20117	45204	15956	60000	18743	92423	97118	96338
00037	19565	41430	01758	75379	40419	21585	66674	36806	84962	85207
00038	45155	14938	19476	07246	43667	94543	59047	90033	20826	69541
00039	94864	31994	36168	10851	34888	81553	01540	35456	05014	51176
00040	98086	24826	45240	28404	44999	08896	39094	73407	35441	31880
00041	33185	16232	41941	50949	89435	48581	88695	41994	37548	73043
00042	80951	00406	96382	70774	20151	23387	25016	25298	94624	61171
00043	79752	49140	71961	28296	69861	02591	74852	20539	00387	59579
00044	18633	32537	98145	06571	31010	24674	05455	61427	77938	91936

Activity 2: Dolphin Therapy

Swimming with dolphins can certainly be fun, but is it also therapeutic for patients suffering from clinical depression? To investigate this possibility, researchers recruited 30 subjects aged 18-65 with a clinical diagnosis of mild to moderate depression. Subjects were required to discontinue use of any antidepressant drugs or psychotherapy four weeks prior to the experiment, and throughout the experiment. These 30 subjects went to an island off the coast of Honduras, where they were randomly assigned to one of two treatment groups. Both groups engaged in the same amount of swimming and snorkeling each day, but one group (the animal care program) did so in the presence of bottlenose dolphins and the other group (outdoor nature program) did not. At the end of two weeks, each subjects' level of depression was evaluated, as it had been at the beginning of the study, and it was determined whether they showed "substantial improvement" (reducing their level of depression) by the end of the study (Antonioli and Reveley, 2005).

Before the data are collected, you should anticipate outcomes/state the research hypothesis.
(a) What were the researchers hoping to show in this study?

(b) Based on the above description of the study, identify the following terms:

Observational units
Explanatory variable
Response variable
Type of study (anecdotal, observational, experimental)
How was randomness used in the study (sampling or assignment or both)

The researchers found that 10 of 15 subjects in the dolphin therapy group showed substantial improvement, compared to 3 of 15 subjects in the control group.

(c) Organize these results in a 2 × 2 table:

	Dolphin therapy	Control group	Total
Showed substantial improvement			
Did not show substantial improvement			
Total			

(d) Calculate the *conditional proportion* who improved in each group and the observed difference in these two proportions (dolphin group – control group).

Proportion in Dolphin group that substantially improved:

Proportion of Control group that substantially improved:

Difference (dolphin- control):

Construct the 2×2 table to show the number of improves and non-improvers in each group (where clearly nothing different happened to those in "group A" and those in "group B" – any differences that arise are due purely to the random assignment process – a "could have been" distribution).

(g) Report your resulting table and calculate the conditional proportions that improved in each group and the difference (dolphin-control) between them. (If working with a partner, repeat this process a second time.)

Simulated table:

Difference in conditional proportions (dolphin – control):

(h) Is the result of this simulated random assignment *as extreme* as the actual results that the researchers obtained? That is, did 10 or more of the subjects in the dolphin group improve in this simulated table?

(n) Would you say that the results that the researchers obtained provide strong evidence that dolphin therapy is more effective (i.e., that the null model is not correct)? Explain your reasoning, based on your simulation results, including a discussion of the purpose of the simulation process and what information it revealed to help you answer this research question.

The three S's are used again here.

- *Statistic*: We can use either the number of successes in group A or the difference in the conditional proportions
- *Simulation*: We assumed there was no treatment effect, that the number of successes and failures was not influenced by which group individuals were assigned to and we simulated the random assignment of the subjects to the treatment groups.
- *Strength of evidence*: Again our observed statistic (.467) was in the tail of the "what if" distribution, providing strong evidence (p-value < .05) against the null model.

Summarize:

(o) Are you willing to draw a cause-and-effect conclusion about dolphin therapy and depression based on these results? Justify your answer based on the design of the study.

(p) Are you willing to generalize these conclusions to all people who suffer from depression? How about to all people with mild to moderate depression in this age range? Justify your answer based on the design of the study

Scope of Conclusions permitted depending on study design (adapted from Ramsey and Schafer's *The Statistical Sleuth*)

		Allocation of units to groups		
		By random assignment	**No random assignment**	
Selection of units	**Random sampling**	A random sample is selected from one population; units are then randomly assigned to different treatment groups	Random samples are selected from existing distinct populations	⇒ *Inferences to populations can be drawn*
	Not random sampling	A groups of study units is found; units are then randomly assigned to treatment groups	Collections of available units from distinct groups are examined	

⇓

Can draw cause and effect conclusions

Activity 3: German Tank Problem

In World War II, Allies used serial numbers on captured axis tanks to estimate the total number of tanks in the German army. Through intelligence, it was determined that the serial numbers on gearboxes as well as those on chassis and engine were unbroken sequences of numbers. The statistical analysis is addressed in Ruggles and Brodie (1947). The task at hand was to estimate N from a sample of size n taken from a sequential list of numbers: 1, 2, … N.

How can a random sample of integers between 1 and N be used to estimate N?

1. The tanks are numbered from 1 to N. Working with your group, randomly select five tanks, without replacement, from the bowl. The tanks are numbered:

2. Think about how you would use your data to estimate N. Come up with at least 3 estimators. Our estimates of N are:

Our rules or formulas for the estimators of N based on a sample of size n (in your case 5) integers are:

3. What makes a good estimator? Come up with at least 2 qualities that you hope your estimator satisfies.