Teaching Introductory Statistics: Implementing GAISE Recommendations

Allan Rossman
Department of Statistics, Cal Poly – San Luis Obispo
arossman@calpoly.edu https://askgoodquestions.blog @allanjrossman
presented at CSU – Monterey Bay on October 8, 2019

GAISE recommendations (www.amstat.org/education/gaise/):

1. Teach statistical thinking.
    - Teach statistics as an investigative process of problem-solving and decision-making.
    - Give students experience with multivariable thinking.
2. Focus on conceptual understanding.
3. Integrate real data with a context and a purpose.
4. Foster active learning.
5. Use technology to explore concepts and analyze data.
6. Use assessments to improve and evaluate student learning.

Workshop schedule:

9:00 – 9:10 Overview, introductions

9:10 – 9:40 Statistical thinking
Activity 1: Graduate admissions
Activity 2: Cancer pamphlets
Activity 3: Draft lottery

9:40 – 10:20 Data collection
Activity 4: Lincoln's words
Activity 5: Home court disadvantage
Activity 6: Mandela's age

10:20 – 10:30 Break

10:30 – 11:30 Simulation-based inference
Activity 7: Toy preferences
Activity 8: Facial prototyping
Activity 9: Tagging penguins

11:30 – 11:50 Cautions about inference
Activity 10: Cat households
Activity 11: Female senators
Activity 12: Remembering presidents

11:50 – 12:00 Wrap-up

Contact me (arossman@calpoly.edu) for an electronic file of this handout.
Follow my blog (askgoodquestions.blog) and follow me on twitter (@allanjrossman).

Activity 1: Graduate admissions

The University of California at Berkeley was charged with having discriminated against women in their graduate admissions process for the fall quarter of 1973. The table below shows the number of men accepted and denied and the number of women accepted and denied for two of the university's graduate programs:

	Men	Women
Accepted	533	113
Denied	665	336
Total	1198	449

a) Calculate the proportion of men applicants who were accepted and the proportion of women applicants who were accepted. Is there evidence that men were accepted at a much higher rate than women? Explain.

Men: Women:

The table below further classifies this information for the two graduate programs:

	Men		Women	
	Accepted	Denied	Accepted	Denied
Program A	511	314	89	19
Program F	22	351	24	317
Total	533	665	113	336

b) Verify that the column totals of the two programs match the counts in the table above.

c) *Within each program*, calculate the proportion of men who were accepted and the proportion of women who were accepted. Did men have the higher rate of acceptance in both programs? Does this seem consistent with your results in (a)? Explain.

Program A: Program F:

Men: Men:

Women: Women:

d) Describe the oddity that is revealed by your answers to (a) and (c).

e) Using the data provided in the tables, explain how this oddity happened in this study. Also describe what this means about the issue of whether the university was guilty of sex discrimination.

Activity 2: Cancer pamphlets

Researchers investigated whether pamphlets containing information for cancer patients are written at a level that the cancer patients can comprehend. They applied tests to measure the reading levels of 63 cancer patients and also the readability levels of 30 cancer pamphlets. These numbers correspond to grade levels, but patient reading levels of under grade 3 and above grade 12 are not determined exactly. The following tables indicate the number of patients at each reading level and the number of pamphlets at each readability level:

Patients' Reading Level	Below 3	3	4	5	6	7	8	9	10	11	12	Above 12	Total
Tally (count)	6	4	4	3	3	2	6	5	4	7	2	17	63

Pamphlets' Readability Level	6	7	8	9	10	11	12	13	14	15	16	Total
Tally (count)	3	3	8	4	1	1	4	2	1	2	1	30

a) Explain why the form of the data do not allow one to calculate the *mean* reading skill level of a patient.

b) Determine the *median* reading level of a patient.

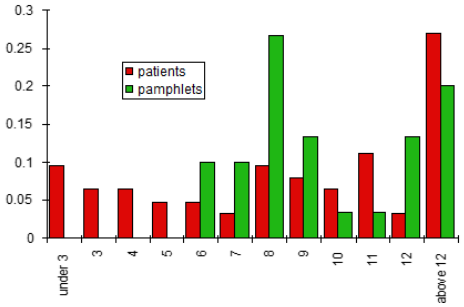
c) Determine the median readability level of a pamphlet.

d) Does the closeness of these medians indicate that the pamphlets are well matched to the patients' reading levels? Explain.

e) What proportion of the *patients* do not have the reading skill level necessary to read even the *simplest* pamphlet in the study?

f) Do you want to reconsider your answer to (d) in light of (e)?

Consider the following graph of these data:



g) Do you think the graph makes the key point more clear than the table of counts? Explain.

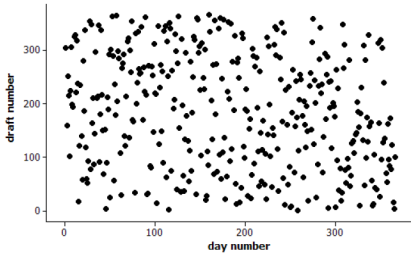
Activity 3: Draft lottery

In the year 1970, the United States instituted a draft of young men into military service. In an effort to be as fair as possible, men were assigned draft numbers based on their birthdays. These numbers were determined by a random mechanism involving ping-pong balls. Men born on the date assigned a draft number of 1 were the first to be drafted, followed by those born on the date assigned draft number 2, and so on. The following table lists the draft numbers (1-366) assigned to birthdates in the 1970 draft lottery:

date	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	305	86	108	32	330	249	93	111	225	359	19	129
2	159	144	29	271	298	228	350	45	161	125	34	328
3	251	297	267	83	40	301	115	261	49	244	348	157
4	215	210	275	81	276	20	279	145	232	202	266	165
5	101	214	293	269	364	28	188	54	82	24	310	56
6	224	347	139	253	155	110	327	114	6	87	76	10
7	306	91	122	147	35	85	50	168	8	234	51	12
8	199	181	213	312	321	366	13	48	184	283	97	105
9	194	338	317	219	197	335	277	106	263	342	80	43
10	325	216	323	218	65	206	284	21	71	220	282	41
11	329	150	136	14	37	134	248	324	158	237	46	39
12	221	68	300	346	133	272	15	142	242	72	66	314
13	318	152	259	124	295	69	42	307	175	138	126	163
14	238	4	354	231	178	356	331	198	1	294	127	26
15	17	89	169	273	130	180	322	102	113	171	131	320
16	121	212	166	148	55	274	120	44	207	254	107	96
17	235	189	33	260	112	73	98	154	255	288	143	304
18	140	292	332	90	278	341	190	141	246	5	146	128
19	58	25	200	336	75	104	227	311	177	241	203	240
20	280	302	239	345	183	360	187	344	63	192	185	135
21	186	363	334	62	250	60	27	291	204	243	156	70
22	337	290	265	316	326	247	153	339	160	117	9	53
23	118	57	256	252	319	109	172	116	119	201	182	162
24	59	236	258	2	31	358	23	36	195	196	230	95
25	52	179	343	351	361	137	67	286	149	176	132	84
26	92	365	170	340	357	22	303	245	18	7	309	173
27	355	205	268	74	296	64	289	352	233	264	47	78
28	77	299	223	262	308	222	88	167	257	94	281	123
29	349	285	362	191	226	353	270	61	151	229	99	16
30	164		217	208	103	209	287	333	315	38	174	3
31	211		30		313		193	11		79		100

a) What draft number was assigned to *your* birthday? Is this draft number in the top third, middle third, or last third of the draft order?

b) Does the following graph reveal any association (i.e., relationship) between draft number and birthdate (where January 1 is 1, January 31 is 31, February 1 is 32, and so on through December 31 as 366)? Or does the graph appear to display purely random variation? Based on the graph, is there any reason to doubt that the assignment of draft numbers to birthdays was a fair, random process? Explain.



c) Determine the *median* draft number for your birth month. (The median is the middle value, once the data are sorted from smallest to largest. The table on the next page should be helpful, because it arranges the draft numbers for each month *in order*.)

d) Combine the findings of the class to report the median draft number for each month:

Month	Median draft number	Month	Median draft number
January		July	
February		August	
March		September	
April		October	
May		November	
June		December	

e) Do you notice a pattern in the median draft numbers as the months progress? If so, describe the pattern.

f) What does this suggest about whether the assignment of draft numbers to birthdays was a truly fair, random process? Explain.

rank	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1	17	4	29	2	31	20	13	11	1	5	9	3
2	52	25	30	14	35	22	15	21	6	7	19	10
3	58	57	33	32	37	28	23	36	8	24	34	12
4	59	68	108	62	40	60	27	44	18	38	46	16
5	77	86	122	74	55	64	42	45	49	72	47	26
6	92	89	136	81	65	69	50	48	63	79	51	39
7	101	91	139	83	75	73	67	54	71	87	66	41
8	118	144	166	90	103	85	88	61	82	94	76	43
9	121	150	169	124	112	104	93	102	113	117	80	53
10	140	152	170	147	130	109	98	106	119	125	97	56
11	159	179	200	148	133	110	115	111	149	138	99	70
12	164	181	213	191	155	134	120	114	151	171	107	78
13	186	189	217	208	178	137	153	116	158	176	126	84
14	194	205	223	218	183	180	172	141	160	192	127	95
15	199	210	239	219	197	206	187	142	161	196	131	96
16	211	212	256	231	226	209	188	145	175	201	132	100
17	215	214	258	252	250	222	190	154	177	202	143	105
18	221	216	259	253	276	228	193	167	184	220	146	123
19	224	236	265	260	278	247	227	168	195	229	156	128
20	235	285	267	262	295	249	248	198	204	234	174	129
21	238	290	268	269	296	272	270	245	207	237	182	135
22	251	292	275	271	298	274	277	261	225	241	185	157
23	280	297	293	273	308	301	279	286	232	243	203	162
24	305	299	300	312	313	335	284	291	233	244	230	163
25	306	302	317	316	319	341	287	307	242	254	266	165
26	318	338	323	336	321	353	289	311	246	264	281	173
27	325	347	332	340	326	356	303	324	255	283	282	240
28	329	363	334	345	330	358	322	333	257	288	309	304
29	337	365	343	346	357	360	327	339	263	294	310	314
30	349		354	351	361	366	331	344	315	342	348	320
31	355		362		364		350	352		359		328

Activity 4: Lincoln's words
Consider the population of 268 words in the passage below.

a) Select a sample by circling ten words that you believe to be representative of the population.

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.

b) For each word in your sample, record how many letters are in the word.

c) What the observational units and variable are in your sample? Is the variable categorical or numerical?

Observational units: Variable: Type:

d) Calculate the average (mean) number of letters per word in your sample. [*Hint*: Add up the number of letters in each word and divide by ten.]

e) Combine your sample average with those of your classmates to produce a dotplot of sample averages below. Be sure to label the horizontal axis appropriately.

f) Indicate what the observational units and variable are in this dotplot in (e). [*Hint*: To identify what the observational units are, ask yourself what each dot on the plot represents. The answer is different than in (c).]

g) The average number of letters per word in the population of all 268 words is 4.3. Mark this value on the dotplot in (e). Were most sample averages greater or less than this population average?

h) 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.

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

j) Consider a different sampling method: close your eyes and point to the page ten times in order to select the words for your sample. Explain why this method would also be biased toward overestimating the average number of letters per word in the population.

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

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

To examine the long-term patterns of ***random*** sampling, we will use the **Sampling Words** applet (http://www.rossmanchance.com/applets.html) to select a large number of random samples from this population. The information in the top right panel shows the population distribution and tells you the average number of letters per word in the population.

m) Specify 5 as the sample size and click Draw samples. Record the lengths of the words and the average length for the sample of 5 words.

n) Click Draw Samples again. Did you obtain the same sample of words this time? Did you obtain the same average length?

o) Change the Number of samples from 1 to 998. Click off the Animate button. Then click on the Draw Samples button. The applet now takes 998 more simple random samples from the population (for a total of 1000) and adds the sample results to the graph in the lower right panel. The red arrow indicates the average of the 1000 sample averages. Record this value below.

p) If the sampling method is unbiased, the sample averages should be centered near the population average of 4.3 words. Does this appear to be the case?

q) What do you suspect would happen to the distribution of sample averages if you took many random samples of 20 words rather than 5? Explain briefly. [*Hint*: Comment on center/tendency and variability/consistency.]

r) Change the sample size in the applet to 20, and take 1000 random samples of 20 words each. Summarize how the distributions of average word lengths compare to when the sample size was 5 words per sample.

s) Which of the two distributions (sample size 5 or sample size 20) has less variability in the values of the sample average word length? In which case (sample size 5 or sample size 20) is the result of a *single* sample more likely to be close to the truth about the population?

t) Which is preferable in this situation: more variability or less variability? Explain.

u) Is taking a large sample *always* preferable to taking a smaller sample? Explain. [*Hint*: Think about the *Literary Digest* poll.]

Activity 5: Home court disadvantage
During the 2008-09 basketball season, the Oklahoma City Thunder won 3 home games and lost 15 when they had a sell-out crowd, compared to 12 home wins and 11 losses when they had a smaller crowd.

a) Identify the observational units in this study, and also the explanatory and response variables.

Observational units:

Explanatory:

Response:

b) Organize the data into a table of counts, with the explanatory variable groups in columns:

	Smaller crowd	Sell-out crowd	Total
Win			
Loss			
Total			

c) Calculate the proportion of wins for each group. Do these proportions suggest an *association* (relationship) between the two variables? Explain.

d) Is it reasonable to conclude that a sell-out crowd *caused* the team to play worse? If not, provide an alternative explanation that plausibly explains the observed association.

- A **confounding variable** is one whose potential effects on a response variable cannot be distinguished from those of the explanatory variable.
  - A confounding variable is related to both the explanatory and response variable.
  - Because of the potential for confounding variables, one *cannot* legitimately draw **cause-and-effect** conclusions from observational studies.

e) Identify a confounding variable in this study, and explain how this confounding variable is related to both the explanatory and response variable.

Activity 6: Mandela's age
We will gather data from the class on the question of how old Nelson Mandela, the first president of South Africa following apartheid, was at his death.

a) Identify the observational units and variables.

b) Was this an observational study or a randomized experiment?

c) Did this study make use of random sampling, random assignment, both, or neither? Also explain the purpose of any randomness that was used.

d) Examine graphical displays of the data. Comment on what they reveal about the research question that motivated the data collection.

Activity 7: Toy preferences

Do children less than a year old recognize the difference between nice, friendly behavior as opposed to mean, unhelpful behavior? Do they make choices based on such behavior? In a study reported in the November 2007 issue of *Nature*, researchers investigated whether infants take into account an individual's actions towards others in evaluating that individual as appealing or aversive (Hamlin, Wynn, and Bloom, 2007). In one component of the study, 10-month-old infants were shown a "climber" character (a piece of wood with "google" eyes glued onto it) that could not make it up a hill in two tries. Then they were alternately shown two scenarios for the climber's next try, one where the climber was pushed to the top of the hill by another character (the nice toy) and one where the climber was pushed back down the hill by another character (the mean toy). The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (representing the nice toy and the mean toy) and asked to pick one to play with. Videos of this study are available at: www.youtube.com/watch?v=anCaGBsBOxM

a) Identify the observational units and variable in this study. Is the variable categorical or numerical?

Researchers found that 14 of the 16 infants in the study selected the nice toy.

b) Determine the sample proportion of infants who selected the nice toy.

c) Suggest two possible explanations for this result that the researchers observed.

d) Suppose for the moment that the researchers' conjecture is wrong, and infants *actually have no preference* for either type of toy. Would it be *possible* to have obtained a result as extreme as the researchers found?

e) Again suppose that infants have no preference. How many of the 16 would you expect to select the nice toy? Would you *always* expect to see that many of the 16 infants select the nice toy? Explain briefly.

The key question here is to determine what results would occur in the long run under the assumption that infants actually have no preference. (We will call this assumption of no genuine preference the **null hypothesis**.) We will answer this question with by **simulating** (artificially re-creating) the selection process of 16 infants over and over, *assuming that infants actually have no genuine preference*.

f) Describe how we could use a common device to simulate the infants' selection process.

g) Flip a coin 16 times. Record the number of heads that you obtain, which represents the number of your 16 hypothetical infants who choose the nice toy.

h) Combine your simulation results with your classmates. Produce a well-labeled dotplot. Identify the observational units and variable in the resulting dotplot.

Observational units: Variable:

i) Where is the distribution of number of heads in 16 tosses centered? Explain why this makes sense.

j) Looking at this dotplot, does it seem that the result obtained by the researchers would have been *surprising* if in fact the infants had no preference? What does this suggest about whether the researchers' result provides much evidence that the infants do genuinely prefer the nice toy? Explain.

We really need to simulate this random selection process hundreds, preferably thousands of times. This would be very tedious and time-consuming with coins, so we'll turn to technology.

k) Use the One Proportion Inference applet (www.rossmanchance.com/ISIapplets.html) to simulate the random process of 16 infants making this toy choice, still assuming the null model that infants have no real preference and so are equally likely to choose either toy. (Start with 1 repetition, and click the "16 tosses" button. Repeat this a total of 5 times. Then ask for 20 repetitions and watch the results. Finally, change the number of repetitions to 9975, which will produce a total of 10,000 repetitions.) Describe the shape of the resulting dotplot, and comment on whether it is centered where you expected.

l) Based on your simulation results, would you say that it would be *very surprising*, if infants actually have no genuine preference, that 14 out of 16 infants in the study would have chosen the nice toy just by chance? Explain.

m) Report how many of your 10,000 repetitions produced 14 or more infants choosing the friend toy. Also determine the *proportion* of these 10,000 repetitions that produced such an extreme result.

- This proportion is called an *approximate* **p-value**. A p-value is the probability of obtaining a result at least as extreme as the one observed, assuming that there is no genuine preference/difference.
- A *small* p-value provides evidence against (i.e., casts doubt on) the null model/hypothesis used to perform the simulation (in this case, that infants have no genuine preference).
    - A p-value of .10 or less is generally considered to be *some* evidence against the null model/hypothesis.
    - A p-value of .05 or less is generally considered to be *fairly strong* evidence against the null model/hypothesis.
    - A p-value of .01 or less is generally considered to be *very strong* evidence against the null model/hypothesis.
    - A p-value of .001 or less is generally considered to be *extremely strong* evidence against the null model/hypothesis.

n) Would you say that the data obtained by the researchers provide strong evidence that infants in general have a genuine preference for the nice toy over the mean toy? Explain the reasoning process behind your answer.

o) Now suppose that 10 of the 16 infants had chosen the nice toy. Use the applet to report the approximate p-value, and summarize the conclusion that you would have drawn from this result about the strength of evidence that infants have a genuine preference for the nice toy.

Activity 8: Facial Prototyping

A study in *Psychonomic Bulletin and Review* (Lea, Thomas, Lamkin, & Bell, 2007) presented evidence that "people use facial prototypes when they encounter different names." Similar to one of the experiments they conducted, you will be asked to match photos of two faces to the names Tim and Bob:

a) Which name do you associate with the face on the LEFT: Bob or Tim?

The researchers wrote that their participants "overwhelmingly agreed" on which face belonged to which name. We will repeat this study by collecting data from students in our class, and then investigate whether our class data provide strong evidence that people have a natural tendency to associate names with faces as the researchers proposed.

b) Describe (in words) the null model/hypothesis to be investigated with this study.

c) Record the number of students in our class who matched the name that the researchers anticipated with the face on the left, and also the number who did not. What proportion of students in our class put the "correct" names with faces?

d) Describe how you could (in principle) use a coin to produce a simulation analysis of whether these data provide strong evidence that all students at your school/college would correctly match the names to faces more than half the time.

e) Use the One-Proportion Inference applet to conduct a simulation analysis, addressing the question of whether our class results provide strong evidence in support of the conjecture that Cal Poly students would correctly match the names to faces more than half the time. Where does our observed class result falls in that distribution?

f) Report the approximate p-value from this simulation analysis. Also write a sentence or two describing what this approximate p-value means.

g) Summarize what your simulation analysis reveals about whether our class data provide strong evidence in support of the conjecture that all students would correctly match the names to faces more than half the time. Also explain the reasoning process behind your conclusion.

Activity 9: Tagging penguins

Are metal bands used for tagging penguin actually harmful to them? Researchers (Saraux et al., 2011) investigated this question with a sample of 20 penguins near Antarctica. All of these penguins had already been tagged with RFID chips, and the researchers randomly assigned 10 of them to receive a metal band on their flippers in addition to the RFID chip. The other 10 penguins did not receive a metal band. Researchers then kept track of which penguins survived for the 4.5-year study and which did not.

a) Identify the observational units and classify the explanatory and response variables in this study.

Observational units:

Explanatory:

Response:

b) Why did the researchers include a control group in this study? Why didn't they just see how many penguins survived while wearing a metal band?

c) Is this an observational study or an experiment? Explain how you know.

d) State the appropriate null and alternative hypotheses (in words).

The researchers found that 9 of 20 penguins survived, of whom 3 had a metal band and 6 did not.

e) Organize these results into the following 2×2 table:

	No metal band (control)	**Metal band**	**Total**
Survived			
Did not survive			
Total	10	10	20

f) Calculate the *proportion* who survived in each group. Also calculate the *difference* between these proportions (subtracting the "metal band" proportion from the "control" proportion). Did the "metal band" group have a smaller proportion who survived, compared to the control group?

g) Is it *possible* that this difference could have happened even if the metal band had no effect; i.e., simply due simply to the random nature of assigning penguins to groups (i.e., the luck of the draw)?

Sure, this is *possible*. But how *unlikely* is it? In order to answer this question we will use *simulation*. We will perform our simulation assuming that the metal band has *no effect* on penguin survival. More specifically, we will assume that the 9 penguins who survived would have done so with the metal band or not, and the 11 penguins who did not survive also would have not survived with the metal band or not. We could use cards to conduct this simulation.

h) How many cards do we need? What will each card represent?

i) You will be asked to select a group of cards to represent the penguins in this study. Some cards will be red, and others will be black. What do you think these colors represent? How many cards will you select of each color?

j) You will shuffle the cards well and then deal them into two stacks. What will these stacks represent, and how many cards will you deal into each stack?

k) Perform this shuffling and dealing. Then fill in the table of simulated results below:

Simulation repetition #1	**No metal band (control)**	**Metal band**	**Total**
Survived			
Did not survive			
Total			

Difference in proportions who survived (control minus metal band group):

l) Repeat this shuffling and dealing a second time. Again fill in the table of simulated results:

Simulation repetition #2	**No metal band (control)**	**Metal band**	**Total**
Survived			
Did not survive			
Total			

Difference in proportions who survived (control minus metal band group):

We need to carry out this simulated random assignment process a very large number of times. This would be very tedious and time-consuming by hand, so let's turn to the **Two Proportions** applet (http://www.rossmanchance.com/ISIapplets.html).

m) Enter the observed counts in the 2×2 table for this study into the table in the upper left cell of the applet. Then press the **Use Table** button. Notice that the applet produces a segmented bar graph and calculates the observed difference in success proportions.

n) Now check the **Show Shuffle Options** and then press the **Shuffle** button. Notice that the applet does what you have done: shuffle the 20 cards and deal them into two groups. The applet also determines the 2×2 table for the simulated results (check the **Show table** box on the left) and plots the shuffled difference in conditional proportions on the graph to the right. Now press the **Shuffle** button four more times. Do the simulated differences in success proportions vary?

o) Now increase the **Number of Shuffles** to 95 and press the **Shuffle** button. This produces a total of 100 repetitions of the simulated random assignment process. Then use the applet to generate a total of 1000 random assignments. Describe the resulting distribution of the differences in sample proportions of survival between the two groups. Comment on its shape (does this look familiar?), center (why does the center make sense?), and variability.

p) Determine the proportion of your 1000 simulated random assignments for which the results are at least as extreme as the actual study (which, you'll recall, saw a 0.30 difference in success proportions between the groups). Report and interpret this (approximate) *p-value*, including what is meant by the phrase "by chance alone" in this context.

q) Is your simulated p-value small enough so that you would consider a difference in the observed proportion of successes of 0.30 or more to be *surprising* under the null hypothesis that the metal band has no effect on penguins' survival?

r) Summarize your conclusion from this study and your simulation analysis. Be sure to relate your conclusion to the context, and also explain the reasoning process that supports this conclusion. (Make use of your answer to the previous question.)

I have a confession to make: The results described above were actually just a small part of the study. We chose to start with partial results to make the by-hand simulation with cards fairly manageable. But the actual study involved 100 penguins, with 50 randomly assigned to each group. The researchers found that 16 of 50 survived in the metal band group, and 31 of 50 survived in the control group.

s) Record the actual results from the full study in the 2×2 table:

	No metal band (control)	Metal band	Total
Survived			
Did not survive			
Total			

t) Determine the conditional proportions who survived in each group. Also calculate the difference in these proportions (control – metal band). Is the value of this difference similar to what you calculated earlier for the partial study?

u) Use the applet to conduct a simulation analysis based on data from the full study. Conduct at least 1000 repetitions of the random assignment. Again describe the (null) distribution of the difference in conditional proportions of survival between the two groups (shape, center, variability).

v) Use the applet to determine the (approximate) p-value from your simulation analysis. Interpret what this p-value means in this context.

w) Based on the (approximate) p-value from your simulation analysis, do the observed results from the actual study (the *full* study, this time) provide convincing evidence that metal bands have a harmful effect on penguins' survival? Explain the reasoning process behind your conclusion, as if to someone with no training in statistics.

x) Is drawing a cause-and-effect conclusion (between the metal bands and a lower chance of penguin survival) justified from this study? Explain why or why not. (Be sure to identify both reasons for the correct conclusion.)

Activity 10: Cat households

A national survey of 50,347 households in December of 2011 found that 30.4% of American households own a pet cat (*2012 U.S. Pet Ownership & Demographics Sourcebook*, American Veterinary Medical Association).

a) Is this number a parameter or a statistic? Explain, and indicate the symbol used to represent it.

b) Conduct a significance test of whether the sample data provide evidence that the population proportion who own a pet cat differs from one-third. (Feel free to use the Theory-Based Inference applet.) State the hypotheses, and report the test statistic and p-value. Draw a conclusion in the context of this study.

c) Produce a 99% confidence interval (CI) for the population proportion who own a pet cat. (Again feel free to use technology.) Interpret this interval.

d) Are the test decision and confidence interval *consistent* with each other? Explain how you can tell.

e) Do the sample data provide *very* strong evidence that the population proportion who own a pet cat is not one-third? Explain whether the p-value or the CI helps you to decide.

f) Do the sample data provide strong evidence that the population proportion who own a pet cat is *very* different from one-third? Explain whether the p-value or the CI helps you to decide.

This activity illustrates the distinction between *statistical* significance and *practical* importance. Especially with large sample sizes, a small difference that is of little practical importance can still be statistically significant (unlikely to have happened by chance alone). Confidence intervals should accompany significance tests in order to estimate the size of an effect/difference.

Activity 11: Female senators

Suppose that an alien lands on Earth, notices that there are two different sexes of the human species, and sets out to estimate the proportion of humans who are female. Fortunately, the alien had a good statistics course on its home planet, so it knows to take a sample of human beings and produce a confidence interval. Suppose that the alien happened upon the members of the 2019 U.S. Senate as its sample of human beings, so it finds 25 women (the most ever to serve in the U.S. Senate!) and 75 men in its sample.

a) Use this sample information to produce a 95% confidence interval for the population proportion of all humans who are female.

b) Based on your experience living on earth for many years, do you think this confidence interval provides a reasonable estimate of the population proportion of all humans who are female?

c) Explain why the confidence interval procedure fails to produce an accurate estimate of the population parameter in this situation.

d) It clearly does not make sense to use the confidence interval in (a) to estimate the proportion of women on Earth, but does the interval make sense for estimating the proportion of women in the 2019 U.S. Senate? Explain your answer.

This activity illustrates some important limitations of inference procedures.

- First, they do not compensate for the problems of a biased sampling procedure. If the sample is collected from the population in a biased manner, the ensuing confidence interval will be a biased estimate of the population parameter of interest.
- A second point to remember is that confidence intervals and significance tests use sample statistics to estimate population parameters. If the data at hand constitute the entire population of interest, then constructing a confidence interval from these data is meaningless. In this case, you know precisely that the proportion of women in the population of the 2019 U.S. Senators is 0.25 (exactly!), so it is senseless to construct a confidence interval from these data.

Activity 12: Remembering presidents

a) List as many U.S. Presidents as you can in reverse chronological order. In other words, start with the current president, then list that president's predecessor, and so on as far as you can. (Do not cheat by using the internet or any other resource!)

b) As we present the list of U.S. presidents, calculate your score for this exercise, which we will define to be the number you identified correctly before getting one wrong or out of order.

c) We will collect the scores of the class and then examine visual displays of the distribution. Write a paragraph summarizing its key features.

d) Use the sample data to calculate a 95% confidence interval for the population mean.

e) Write a sentence or two interpreting this confidence interval.

f) How many and what proportion of the sample values fall within this interval? Is this close to 95%? Should you expect this to be close to 95%? Explain why or why not.

This last question is meant to remind you that confidence intervals of this type estimate the value of a population *mean*. They do not estimate the values of *individual* observations in the population or in the sample.