

Teaching Introductory Statistics: Implementing GAISE Recommendations

Allan Rossman

Department of Statistics, Cal Poly – San Luis Obispo

arossman@calpoly.edu

presented at College of San Mateo on February 29, 2020

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 – 9:30 Introductions, Overview of GAISE

9:30 – 10:30 Statistical thinking

Activity 1: Graduate admissions

Activity 2: Draft lottery

Activity 3: Smoking and lung capacity

10:30 – 10:45 Break

10:45 – 11:30 Data collection

Activity 4: Lincoln's words

Activity 5: Home court disadvantage

Activity 6: Mandela's age

11:30 – 12:15 Probability

Activity 7: Random babies

Activity 8: Senators

12:15 – 1:15 Lunch

1:15 – 2:30 Simulation-based inference

Activity 9: Toy preferences

Activity 10: Facial prototyping

Activity 11: Halloween treats

Activity 12: Tagging penguins

2:30 – 2:45 Break

2:45 – 4:00 Confidence intervals

Activity 13: Tattoos

Activity 14: Remembering presidents

Activity 15: Cat households

Activity 16: Senators (revisited)

4:00 – 4:30 Wrap-up, Q&A

My teaching philosophy: **Ask good questions!**

Blog: <https://askgoodquestions.blog>

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.

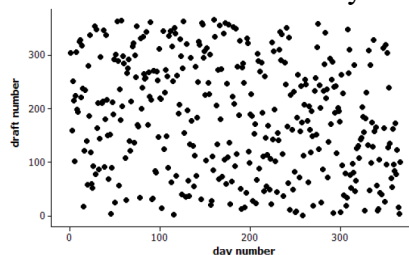
Activity 2: Draft lottery

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	

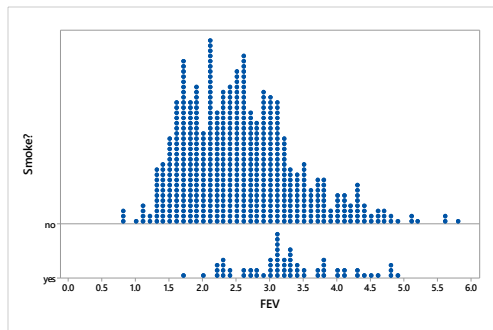
e) Do you notice a pattern in the median draft numbers as the months progress? If so, describe the pattern.

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 3: Smoking and lung capacity

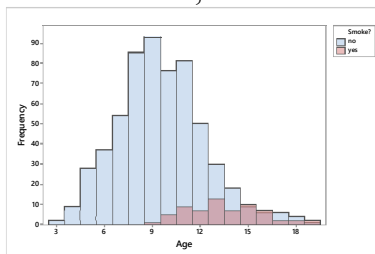
We will examine data on lung capacity, as measured by a quantity called forced expiratory volume (FEV), the amount of air an individual can exhale in the first second of forceful breath (in liters), for 654 people who participated in a research study. The following graph displays the distributions of FEV values for smokers and non-smokers:



a) The average FEV values for these two groups are 2.57 liters and 3.28 liters. Which average do you think goes with which group – smokers and non-smokers?

b) Do you consider these data to be surprising? Explain.

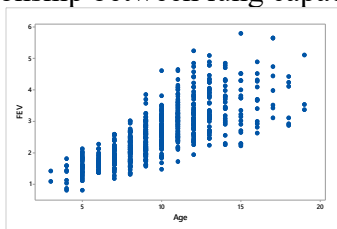
I neglected to tell you that the 654 people were all children and teen-agers, between 3 and 19 years old. Their ages, separated into smokers and non-smokers, are shown in the graph below:



c) Describe how the distributions of ages compare between smokers and non-smokers. Is this surprising or what you would have expected?

d) What do you expect to see in a graph displaying the relationship between lung capacity and age?

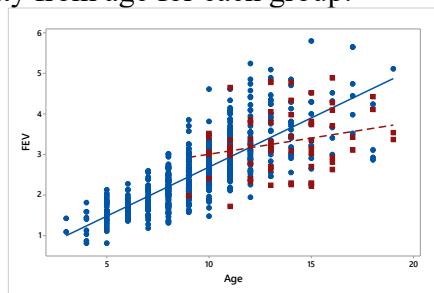
The following graph displays the relationship between lung capacity and age:



e) Describe what this graph reveals. Is this what you expected?

f) What do you expect to see when information on smoking status is added to the graph above?

The following graph includes smoking status along with lung capacity and age. It also displays “best fit” lines for predicting lung capacity from age for each group:



g) Does this graph enable you to make sense of the surprising finding that smokers in this study tend to have larger lung capacities than non-smokers? Explain.

h) Summarize what this analysis reveals about the relationship between smoking and lung capacity.

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 2018-19 basketball season, the Sacramento Kings won 13 home games and lost 16 when they had a sell-out crowd, compared to 11 home wins and 1 loss when they had a smaller crowd.

a) Identify the observational units, explanatory variable, and response variable in this study. Also classify each variable as categorical or numerical.

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.

f) Convert the proportions of wins in each group into percentages. Then write a sentence that uses/interprets the percentages.

g) Calculate the *difference* between the proportions of games that resulted in wins for each group.

h) Is it correct to say that the Kings were about 47% more likely to win a home game with a smaller crowd than with a sell-out crowd? Explain.

- The **percentage difference** between two proportions p_1 and p_2 , where p_1 is treated as the *reference/baseline*, is calculated as: $(p_2 - p_1) / p_1 \times 100\%$.

i) Calculate the percentage difference in the proportions of wins between the two groups, using the sold-out games as the reference group. Then write a sentence using/interpreting this value.

j) Calculate the *ratio* of the two proportions, again using sold-out games as the reference group. Then write a sentence using/interpreting this value.

- The **relative risk** of an outcome between two groups is the *ratio* of the proportion having that outcome between the two groups.

k) Suggest how the percentage difference between two proportions can be calculated from the relative risk.

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: Random babies

Suppose that on one night at a certain hospital, four mothers give birth to baby boys. As a very sick joke, the hospital staff decides to return babies to their mothers completely at random.

a) Simulate this process by shuffling four index cards marked with the babies' first names and dealing them randomly onto a sheet marked with the mothers' last names. Do this three times, in each case recording the number of mothers who get the right baby (a "match").

Repetition #1:

Repetition #2:

Repetition #3:

b) Combine your results with the rest of the class, filling in the "count" row of the table:

# of matches:	0	1	2	3	4
count					
proportion					

c) Determine the proportion of times that there are 0 matches, 1 match, and so on. Record these in the "proportion" row of the table above.

- A process is **random** if individual outcomes are uncertain but there is a regular distribution of outcomes in a large number of repetitions.
- The **probability** of any outcome in a random process is the *proportion* of times that the outcome would occur in a very large number of repetitions (also known as *relative frequency*).
- A probability can be approximated by **simulating** (artificially re-creating) the random process a large number of times and determining the relative frequency of occurrences.

d) Use the **Random Babies** [applet](#) to simulate this random process a total of 10,000 times. First do 1 at a time for 5 repetitions with the "animate" box checked. [Warning: The applet contains explicit visual images revealing where babies come from!] Then do 9995 more. Record the resulting approximate probabilities in the table:

# of matches:	0	1	2	3	4
Approx prob					

e) Click on the bar in the graph corresponding to 0 matches, and the applet will reveal a graph of the relative frequency as the number of repetitions increased. Does this graph indicate that the relative frequency is varying less as time goes on, perhaps approaching a specific limiting value?

A theoretical analysis of this process would consider all of the possible ways to distribute the four babies to the four mothers. All of the possibilities are listed here:

1234	1243	1324	1342	1423	1432
2134	2143	2314	2341	2413	2431
3124	3142	3214	3241	3412	3421
4123	4132	4213	4231	4312	4321

f) How many possibilities are there for returning the four babies to their mothers?

g) For each of these possibilities, indicate how many mothers get the correct baby.

h) Count how many ways there are to get 0 matches, 1 match, and so on. Record these in the middle row of the table below:

# of matches:	0	1	2	3	4
# possibilities					
probability					

i) Determine the probability of each event by dividing these counts by your answer to (e). Record these in the last row of the table above.

j) Are these theoretical probabilities close to the ones you approximated by simulation?

k) What is the probability that at least one mother gets the correct baby? Suggest two different ways to calculate this.

- The listing of all possible outcomes is called the **sample space** of the random process.
- A **probability model** describes all possible outcomes and assigns probabilities to them.
- In some cases it is reasonable to assume that the sample space outcomes are **equally likely**.

l) For the applet simulation results, calculate the average (mean) number of matches per repetition of the process. (Yes, I'm asking you to add 10,000 numbers and divide the sum by 10,000.)

- The long-run average value achieved by a numerical random process is its **expected value**.
- To calculate this expected value from the (exact) probability distribution, multiply each possibility by its probability, and then add these up across all of the possible outcomes.

m) Calculate the theoretical expected number of matches from the (exact) probability distribution, and compare that to the average number of matches from the simulated data.

Activity 8: Senators

Suppose that a member of the U.S. Senate is to be selected at random.

- a) How many U.S. Senators are there?
- b) Make an educated guess for the number of Senators who are Democrats and the number who are Republicans.
- c) Make an educated guess for the number of Senators who are women and the number who are men.
- d) Name the two U.S. Senators from your home state.

The 2020 U.S. Senate consists of 53 Republicans and 47 Democrats. (Two senators are Independents but caucus with Democrats.) There are currently 26 women and 74 men in the Senate, which is the most women that have ever been in the Senate.

Now suppose that one Senator is to be selected at random.

- e) What is the probability that the Senator selected at random turns out to be a Republican?
- f) What is the probability that the Senator selected at random turns out to be a woman?
- g) Fill in as many counts as you can in the following table:

	Democrat	Republican	Total
Man			
Woman			
Total			

- h) What additional information do you need in order to fill in the rest of this table? What's the minimum amount of additional information that you need?

A listing of the 26 women Senators, along with their party (D or R) and state, appears on the next page.

Tammy Baldwin (D-WI)	Marsha Blackburn (R-TN)	Maria Cantwell (D-WA)
Shelley Moore Capito (R-WV)	Susan Collins (R-ME)	Catherine Cortez Masto (D-NV)
Tammy Duckworth (D-IL)	Joni Ernst (R-IA)	Dianne Feinstein (D-CA)
Deb Fischer (R-NE)	Kirsten Gillibrand (D-NY)	Kamala Harris (D-CA)
Margaret Wood Hassan (D-NH)	Mazie Hirono (D-HI)	Cindy Hyde-Smith (R-MS)
Amy Klobuchar (D-MN)	Kelly Loeffler (R-GA)	Martha McSally (R-AZ)
Lisa Murkowski (R-AK)	Patty Murray (D-WA)	Jacky Rosen (D-NV)
Jeanne Shaheen (D-NH)	Krysten Sinema (D-AZ)	Tina Smith (D-MN)
Debbie Stabenow (D-MI)	Elizabeth Warren (D-MA)	

i) Use this information to fill in the remaining counts in the table:

	Democrat	Republican	Total
Man			74
Woman			26
Total	47	53	100

j) Now suppose that you learn that the senator selected at random is a Republican. Given this (partial) information, what is the updated (conditional) probability that the Senator selected at random is a woman?

k) Fill in the blanks to complete this sentence to interpret the conditional probability calculation:

_____ % of _____ are _____ .

l) Determine the conditional probability that the senator selected is a woman, given that the senator selected is a Democrat.

m) How do the conditional probabilities (that the senator selected is a woman) compare, depending on whether the partial information is that the senator is a Republican or Democrat? How do these conditional probabilities compare to the (unconditional) probability that the Senator selected at random is a woman?

n) Based on your calculations, would you say that a senator's gender is *independent* of their political party? Give a numerical justification for your answer. Also explain what this means.

o) Determine the conditional probability that the selected Senator is a Republican, given that the Senator is a woman. Use this conditional probability in a sentence.

p) How does this conditional probability compare to the first conditional probability that you calculated above? How are the two questions different?

Now consider selecting a person at random from the approximately 328.5 million residents of the United States.

q) What is the conditional probability that the person selected is a male, given that the person is a U.S. Senator?

r) What is the approximate conditional probability that the person is a U.S. Senator, given that the person is male? Is this close to your previous answer? Explain how the questions are different.

Activity 9: 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.

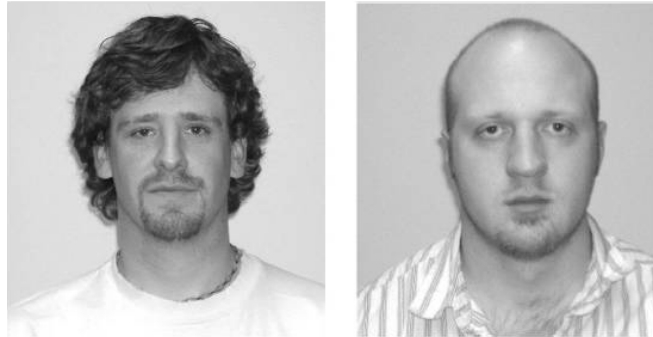
- This proportion is called an *approximate p-value*. A p-value is the probability of obtaining a result at least as extreme as the one observed, assuming that there is no genuine preference/difference.
- A *small* p-value provides evidence against (i.e., casts doubt on) the null model/hypothesis used 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 10: 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 fall 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 11: Halloween treats

Stemming from concern over the nation's obesity epidemic, researchers investigated whether children might be as tempted by toys as by candy for Halloween treats. Test households in five Connecticut neighborhoods offered children two plates: one with lollipops or fruit candy and one containing small, inexpensive Halloween toys, like plastic bugs that glow in the dark. The researchers observed the selections of 283 trick-or-treaters between the ages of 3 and 14 (Schwartz, Chen, and Brownell, 2003).

- a) Identify the observational units and variable in this study. Also classify the variable as categorical or quantitative.
- b) Describe (in words) a parameter of interest in this study. Also indicate the appropriate symbol to denote this parameter.
- c) State the null and alternative hypotheses to be tested. (*Hint: The researchers want to know whether there is a preference for either kind of treat: candy or toys.*)

The researchers found that 148 trick-or-treaters selected the candy, and 135 chose a toy.

- d) Calculate the value of the relevant statistic for this study.
- e) Use the One Proportion Inference applet to conduct a simulation analysis with 10,000 repetitions, assuming the null hypothesis. Comment on the shape, center, and variability of the simulated null distribution of sample proportions.

f) Use the simulation results to approximate the p-value for this study. (*Hint*: Be sure to consider your alternative hypothesis when deciding which outcomes you want to consider more extreme than the observed statistic.) Describe which values the applet counts to determine this p-value, and explain why these are the correct values to count.

Keep in mind that the goal of the p-value is to measure the strength of evidence against the null hypothesis. That is, it provides a way of quantifying how unusual our observed statistic is when the null hypothesis is true. Another way to measure how far out we are in the tail is to look at the distance between the observed value of the statistic and the hypothesized value of the parameter. Better yet, we can *standardize* this distance by using standard deviation units. We will consider observations that fall more than 2 or 3 standard deviations from the mean of the distribution to be in the tail.

g) Calculate the standardized value of the statistic for this study. (*Hint*: Subtract the mean of the simulated sample proportions from the observed value of the sample proportion for this study, and then divide by the standard deviation of the simulated sample proportions.)

h) Interpret what this standardized value means.

i) Report the (approximate) p-value that the applet provides for the normal approximation.

j) Summarize your conclusion from the simulation analysis and standardized statistic from this study.

Extensions:

- k) Describe how your analysis would change if you used candy rather than toy as a “success.”
- l) Are the sample size conditions for using the normal approximation satisfied for this test? Just barely, or by a good bit? Justify your answer.
- m) Are the (approximate) p-values from the normal distribution and from the simulation analysis very close? About how much do they differ in percentage terms?
- n) In which of the two approximations (simulation or normal-based) do you have more confidence/trust? Explain why.
- o) Describe how you could decide which of the two approximate p-values comes closer to the exact p-value in this case.
- p) Click on “Exact Binomial” in the applet. Report the (exact) p-value. Which approximation came closer – simulation or normal?
- q) Make a conjecture for why the normal approximation performed fairly poorly with these data.
- r) Suggest a possible “fix” for the problematic issue, and then use the applet to implement the fix. Comment on how well this fix has worked.

Activity 12: 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:

<i>Simulation repetition #1</i>	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:

<i>Simulation repetition #2</i>	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 13: Tattoos

A Harris Poll that surveyed 2225 adult Americans on October 14-19, 2015 found that 29% reported having at least one tattoo.

- a) Is 29% (.29) a parameter or a statistic? What symbol do we use for it?
- b) Determine (by hand) a 95% confidence interval for the relevant parameter.
- c) Interpret this interval: We are 95% confident that _____ is between ____ and ____ .

Next we will explore the question of what it means to be “95% confident.”

- Interpreting confidence intervals correctly requires us to think about what would happen if we took random samples from the population over and over again, constructing a CI for the unknown population parameter from each sample.
- The *confidence level* (e.g., 95%) indicates the percentage of samples that would produce a CI that successfully contains the actual value of the population parameter.

We will turn to an applet called “Simulating Confidence Intervals” to illustrate this. First we will make sure that the method is set to “Proportions” and “Wald.” Let’s assume that 35% of all adult Americans have a tattoo. Let’s start with a sample size of 75 and a 95% confidence level.

- d) As we take new samples, what do you notice about the intervals? Are they all the same?
- e) Does the value of the population proportion change as we take new samples?
- f) As we take hundreds and then thousands of samples and construct their intervals, about what percentage seem to be successful at capturing the population proportion?
- g) Sort the intervals, and comment on what the intervals that fail to capture the population proportion have in common.
- h) Now change the confidence level to 80%. Before pressing the Recalculate button, what changes do you expect to see? Then press the button. What two things change about the intervals?

- i) Now change the confidence level to 99%. What two things change about the intervals?
- j) Now change the sample size to 300 (still with a confidence level of 99%). Does this produce a higher percentage of successful intervals? What does change about the intervals?
- k) Is it desirable to have larger (e.g., 99%) or smaller (e.g., 80%) confidence *levels*? Explain.
- l) Is it desirable to have wider or narrower confidence intervals? Explain.
- m) Explain how your previous two answers produce a tension/conflict.
- n) What's a drawback of using a very high confidence level such as 99.9%?
- o) What would it take to achieve the best of both worlds: a very high confidence and a very narrow interval? Why is this so difficult to achieve?
- p) Unlike in this exercise, in most situations you do *not* know the value of the population proportion, and you can only afford to take *one* sample. In such cases, do you know whether or not your confidence interval succeeds in capturing the unknown value of the population proportion? In what sense can you be confident that your interval succeeds?

Activity 14: 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.

Activity 15: 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 16: Senators (revisited)

Suppose that an alien lands on Earth and wants 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 2020 U.S. Senate as its sample of human beings, so it finds 26 women (the most ever to serve in the U.S. Senate!) and 74 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 2020 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 2020 U.S. Senators is 0.26 (exactly!), so it is senseless to construct a confidence interval from these data.