**Teaching Introductory Statistics: Focus on Concepts and Data**

MAA Project NExT

Thursday, July 30 – Friday, July 31, 2020

Presenters: Allan Rossman and Beth Chance

Department of Statistics, Cal Poly – San Luis Obispo

[arossman@calpoly.edu](mailto:arossman@calpoly.edu) [bchance@calpoly.edu](mailto:bchance@calpoly.edu)

***Thursday, July 30*** (all times ET)

2:30 – 2:45 Introductions, Overview

2:45 – 3:15 Statistical investigation process, Simulation-based inference

*Activity 1: Friend or foe?* (introduction to p-value)

*Activity 2: Facial prototyping* (application)

3:15 – 3:45 Statistical thinking



*Activity 3: Graduate admissions* (proportional reasoning, Simpson’s paradox)

*Activity 4: Draft lottery* (measures of center, signal vs. noise)

*Activity 5: Home court disadvantage* (confounding variables)

*Activity 6: Car acceleration* (multivariable thinking)

3:45 – 4:15 Data collection and conclusions

*Activity 7: Lincoln’s words* (sampling bias, random sampling)

*Activity 8: Mandela’s age* (randomized experiment)

4:15 – 4:30 Q&A

***Friday, July 31*** (all times ET)

2:30 – 3:00 Probability

*Activity 9: Random babies* (probability, expected value)

*Activity 10: HIV testing* (reverse conditional probability, table of hypothetical counts)

3:00 – 4:15 Statistical inference

*Activity 11: Sleepless nights* (inference for population mean, *t*-test, 6-step process)

*Activity 12: Is yawning contagious?* (comparing two proportions)

*Activity 13: Lingering effects of sleep deprivation* (comparing two means)

*Activity 14: Cat households* (statistical significance vs. practical importance)

*Activity 15: Female senators* (limitations of inference)

4:15 – 4:30 Q&A

GAISE report: <https://www.amstat.org/education/GAISE/>

Applet collection: <http://www.rossmanchance.com/applets/>

Simulation-based inference blog: <https://www.causeweb.org/sbi/>

Ask good questions blog: <https://askgoodquestions.blog>/

**Activity 1: Friend or Foe?** (Introduction to significance, p-value)

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 (“helper”) and one where the climber was pushed back down the hill by another character (“hinderer”).  The infant was alternately shown these two scenarios several times. Then the child was presented with both pieces of wood (the helper and hinderer characters) and asked to pick one to play with.

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

b) The Methodology section states that for the 10-month-olds, the climber was a yellow triangle; helper and hinderer were a red square and a blue circle (counterbalanced). Also counterbalanced were which event (helping or hindering) the infants observed first and the positions of helper and hinderer when presented to the infants (on left or right). Why are these important considerations?

c) Researchers found that 14 of the 16 infants in the study selected the nice toy. 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) If infants have no preference, how many of the 16 would you have expected to select the nice toy? Would you *always* expect to see that many of the 16 infants select the nice toy? How many of the 16 infants would have to select the nice toy in order for you to fairly well convinced that the researchers’ conjecture is correct, that infants really do have a tendency to prefer the nice toy? Explain.

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 model**or**null hypothesis**.) We will answer this question 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](http://www.rossmanchance.com/applets/OneProp/OneProp.htm) ([http://www.rossmanchance.com/applets/](http://www.rossmanchance.com/applets/OneProp/OneProp.htm)) 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. Change the number of tosses to 16 and press Draw Samples. Uncheck the Animation box and repeat this a total of 5 times. Then change the number of repetitions to 995, which will produce a total of 1000 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 1000 repetitions produced 14 or more infants choosing the friend toy. (Enter 14 in the As extreme as box and press Count.) Also determine the *proportion* of these 1000 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 as extreme as the one observed, assuming that there is no genuine preference/ difference for the helper toy.
* A *small* p-value 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) Is this proportion small enough to consider the actual result obtained by the researchers surprising, assuming the null model that infants have no preference and so choose blindly between the two toys?

o) In light of your answers to the previous two questions, would you say that the experimental data obtained by the researchers provide strong evidence that infants in general have a genuine preference for the friend toy over the foe toy? Explain the reasoning process behind your answer.

p) The researchers next wanted to see whether the preference was made based on a social evaluation more than a perceptual preference (e.g., preference for upward motion). In a second experiment, the same process was repeated but with an inanimate object climbing the hill with no googly eyes attached and no self-propelled motion. For this experiment, they found 6 of twelve 10-month-olds chose the pushup toy. How would the analysis, p-value, and conclusion change for this study? Explain your reasoning and why your answers make intuitive sense.

**References**

Videos of this study are available at:

* *Nature* article Supplementary Information [page](http://www.nature.com/nature/journal/v450/n7169/suppinfo/nature06288.html)
* See also [Overview](https://www.youtube.com/watch?v=anCaGBsBOxM) video from Mind in the Making - The Essential Life Skills Every Child Needs, by Ellen Galinsky
* UBC Center for Infant Cognition [Lab](http://cic.psych.ubc.ca/example-stimuli/) website
* <https://www.youtube.com/watch?v=anCaGBsBOxM> (see especially the 0:30 – 2:00 minute mark)
* Example of online lab: www.rossmanchance.com/iscam3/labs/lab1/lab1\_1.html

**Activity 2: Facial Prototyping** (application with data collected on students)

A study in *Psychonomic Bulletin and Review* (Lea, Thomas, Lamkin, & Bell, 2007) presented evidence that “people use facial prototypes when they encounter different names.” Participants were given two faces and asked to identify which one was Tim and which one was Bob. The researchers wrote that their participants “overwhelmingly agreed” on which face belonged to Tim and which face belonged to Bob, but did not provide the exact results of their study. What can you conclude from the data we collected on you?

**Activity 3: Graduate admissions** (proportional reasoning, Simpson’s paradox)

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 4: Draft lottery** (measures of center, signal vs. noise)

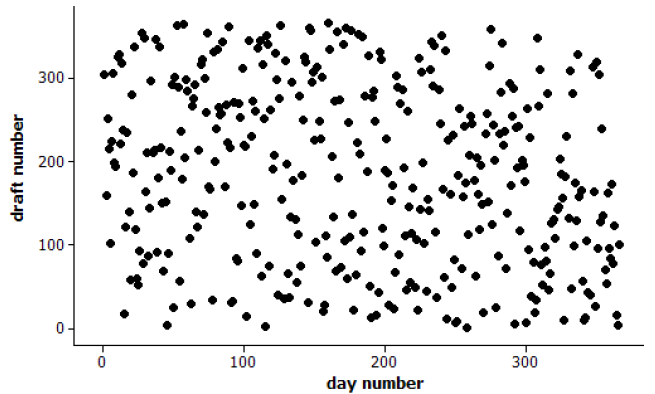
In the year 1970, the United States instituted a draft of young men into military service. In an effort to be as fair as possible, men were assigned draft numbers based on their birthdays. These numbers were determined by a random mechanism involving ping-pong balls. Men born on the date assigned a draft number of 1 were the first to be drafted, followed by those born on the date assigned draft number 2, and so on. See <https://www.youtube.com/watch?v=-p5X1FjyD_g> for a video of television coverage of this event.

The following table lists the draft numbers (1-366) assigned to birthdates in this 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 |

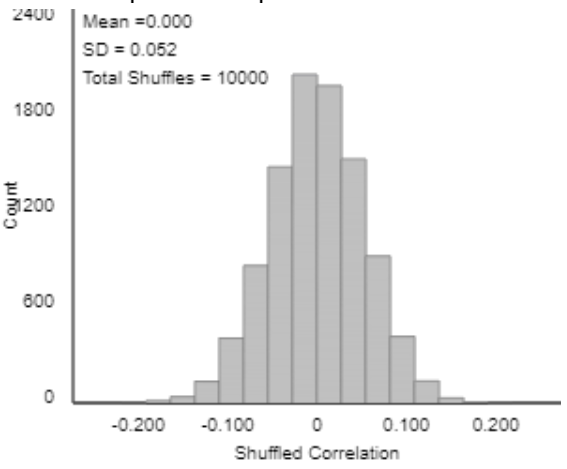
*Extension:*

The correlation coefficient between day number and draft number is -0.226.

g) Is it possible for such an extreme correlation value (i.e., so far from zero) to occur with a truly fair, random lottery?

h) Describe how we could investigate whether a correlation coefficient as extreme as -0.226 would be surprising to occur with a truly fair, random lottery.

The following graph displays the resulting correlation coefficients from 10,000 repetitions of a truly fair, random lottery process:



i) What conclusion would you draw from this simulation analysis? Also explain the reasoning process behind your conclusion.

**Activity 5: Home court disadvantage** (confounding variables)

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 × 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: Car Acceleration** (multivariable thinking)

One commonly cited metric in car commercials is the time it takes the car to accelerate from 0 to 60 mph. In this activity, we will consider what car features might help explain faster acceleration (shorter times) using data on 406 cars (Quinlan, 1993; used in the 1983 ASA Data Graphics Exposition). (This dataset has been around the internet for a while but details on the data collection are limited. We use the version with 392 complete records.)

The carmpg.txt file contains data on the weight of the car (whether or not the car weighs over 2800 kg) and the time it takes to accelerate to 60 mph (measured in seconds).

(a) Paste the carmpg data into the Multiple Variables applet (www.rossmanchance.com/applets). Drag the acceleration variable into the Response box and the weight variable into the Subset By box. Compare the distributions of acceleration times between light and heavy cars. Is this the relationship you expected to see?

(b) Because this is an observational study, we should at least be aware there could be confounding variables. With that in mind, the researchers also collected information on the horsepower of the car (the power of the engine) - whether it was more than 2800 horsepower or not. Exchange this variable with the weight variable in the Subset By box. Compare the distributions of acceleration times between low and high horsepower. Is this the relationship you expected to see?

(c) In order to be a confounding variable, horsepower would need to be associated with both acceleration and weight. The cross-tabulation of horsepower and weight helps to more easily determine whether there is an association between the two variables.

|  |  |  |  |
| --- | --- | --- | --- |
| **Horsepower** | **Heavy cars** | **Light cars** | **Total** |
| **High** | 151 | 16 | 167 |
| **Low** | 45 | 180 | 225 |
| **Total** | 196 | 196 | 392 |

Based on the table, is there an association between weight and horsepower? Make sure that you compute the relevant conditional proportions to make your case. (*Hints*: You can also move weight into the Response box and check Show descriptive.)

(d) Is horsepower a confounding variable on the relationship between weight and acceleration? Why or why not?

So we can think of variation in acceleration as coming from two sources:

acceleration = weight + horsepower + random error

Because of this, a better description of the association between acceleration and weight would first *adjust* the acceleration times based on the horsepower. We will do this by “subtracting off” the horsepower effects of acceleration, and then see if there is a relationship between what’s “left over” and weight. If high horsepower cars tend to lower acceleration by about 1.5 seconds and low horsepower cars tend to increase acceleration time by about 1.5 seconds, we can put the cars on a more level playing field (like a handicap in golf) by adding 1.5 seconds to all high horsepower cars and subtracting 1.5 seconds from all low horsepower cars.

(e) In the applet, use acceleration as the response and weight as the explanatory variable. What is the graph displaying?

(f) Now move horsepower to *above* weight in the explanatory variable box. What is the graph displaying?

(g) So if we follow through with our plan and add roughly 1.5 seconds to all high horsepower cars and subtract roughly 1.5 seconds from all low horsepower cars,

- How is the mean acceleration for the light cars going to change?

- How is the mean acceleration for the heavy cars going to change?

(h) Check the **Adjust values** box to execute our plan. Were your predictions correct? How do the horsepower-adjusted acceleration times for heavy cars compare to light cars?

***Extension*:**  Repeat this exercise using the quantitative version of the horsepower and weight variables (carmpg2.txt) to see an interesting example of Simpson’s Paradox!

**Activity 7: Lincoln’s words** (sampling bias, random sampling)

Consider the population of 268 words in the speech below.

C:\Users\bchance\Downloads\UNP 02.01.tif

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

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 press the Draw Samples button. Record the lengths of the words and the average length for the sample of 5 words.

n) Press 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. Then press 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. Report the mean of this distribution.

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 8: Mandela’s age** (randomized experiment)

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 9: Random babies** (probability, expected value)

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](http://www.rossmanchance.com/applets/randomBabies/RandomBabies.html) ([www.rossmanchance.com/applets/](http://www.rossmanchance.com/applets/)) 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 trials. 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 10: HIV testing** (reverse conditional probability, table of hypothetical counts)

The ELISA test for the human immunodeficiency virus (HIV) that causes AIDS has been used in the screening of blood donations. As with most medical diagnostic tests, the ELISA test is not infallible. If a person actually carries HIV, experts estimate that this test gives a positive result 97.7% of the time. (This number is called the *sensitivity* of the test.) If a person does not carry HIV, ELISA gives a negative result 92.6% of the time (the *specificity* of the test). Suppose that 0.5% of the public carries HIV (the *base rate* with the disease).

a) Suppose that someone tells you that they have tested positive. Given this information, how likely do you think it is that the person actually carries HIV? Circle which of the following you suspect is closest to this probability. 0.1 0.3 0.5 0.7 0.9

Imagine a hypothetical population of 1,000,000 people for whom these percentages hold exactly. You will fill in a two-way table as you derive *Bayes’ Theorem* to address the question above.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Positive test | Negative test | Total |
| Carries HIV | (c) | (c) | (b) |
| Does not carry HIV | (d) | (d) | (b) |
| Total | (e) | (e) | 1,000,000 |

b) Assuming that 0.5% of the population of 1,000,000 people carries HIV, how many such carriers are there in the population? How many non-carriers are there? (Record these in the table.)

c) Consider for now just the carriers. If 97.7% of them test positive, how many test positive? How many carriers does that leave who test negative? (Record these in the table.)

d) Now consider only the non-carriers. If 92.6% of them test negative, how many test negative? How many non-carriers does that leave who test positive? (Record these in the table.)

e) Determine the total number of positive test results and the total number of negative test results. (Record these in the table.)

f) Of those who test positive, what proportion actually carry HIV? How does this compare to your prediction in a)? Explain why this probability is smaller than most people expect.

g) Of those who test negative, what proportion are actually free of HIV?

**Activity 11: Sleepless Nights?** (inference for single mean, *t*-test, 6-step process)

*Step 1: Ask a research question*

How much do students at your school sleep on a typical night? Is the average less than the recommended eight hours? How can we estimate this average?

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

*Step 2: Design a study and collect data*

2) How much sleep (to the nearest quarter hour) did you get last night?

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

*Step 3: Explore the data*

4) In the **Sampling from a Finite Population** [applet](http://www.rossmanchance.com/applets/OneSample.html?population=gettysburg), press **Clear** and paste in the sleep data and press **Use Data**. Use the displayed graph to comment on the shape, center, and variability of this distribution. Are there any unusual observations that merit additional exploration? Is this distribution the sample or the population?

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

*Step 4: Draw inferences beyond the data*

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

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

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

1. Statistic

8) What do you suggest for the choice of a statistic? What was the observed value of the statistic in our sample?

2. Simulate

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

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

*Approach 2: Sampling from a finite population*

a) Select **Sleep 1** from the list of populations to display a distribution of a population of 18,000 students, assuming the null hypothesis is true. What are the mean and standard deviation of this population of sleep times?

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

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

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

*Approach 1: Sampling from a theoretical population*

Follow the Population model link.

*Approach 3: Bootstrapping*

Follow the Bootstrapping link.

3. Strength ofevidence

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

*Approach 4. Theory-based approach*

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

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

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

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

12) Because we don’t typically know , it is common to use *s* instead and calculate the *standard error* of the sample mean as *s*/. Calculate this value for our observed sample.

13) Using your answer to question 12 as the estimate of the sample to sample variation in the sample mean, construct a rough 95% confidence interval for the population mean. Interpret this interval in context.

14) Does the interval in question 13 include the value of 8 hours? Did you expect this based on your earlier analyses? What additional information is provided by this confidence interval?

*Step 5. Formulate conclusions*

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

*Step 6. Look back and ahead*

16) Did anything about the design and conclusions of this study concern you? Issues you may want to critique include:

* The match between the research question and the study design
* How the observational units were selected
* How the measurements were recorded
* Whether what we observed is of practical value

17) What should the researchers’ next steps be to fix the limitations or build on this knowledge?

***Extension:***

18) Suppose you reject the null hypothesis in this study. Does that mean you would have *proven* that the population mean differs from 8 hours? If not, why not?

19) Suppose we fail to reject the null hypothesis that mu = 7 hours. Would this mean you have *proven* that μ = 7 hours? If not, why not?

**Activity 12: Is yawning contagious?** (comparing two proportions)

The folks at *MythBusters*, a popular television program on the Discovery Channel, investigated whether yawning is contagious (video). We will analyze the data from the final study: 50 subjects were recruited at a local flea market and asked to sit in one of three small rooms for a short period of time. For some of the subjects, the attendee yawned while leading them to the room (planting a yawn “seed”), whereas for other subjects the attendee did not yawn. As time passed, the researchers watched (via a hidden camera) to see which subjects yawned.

(a) Summarize the research question for this study.

(b) Review the results of the design of this study and write a paragraph summarizing:

* Was this an observational study or an experiment? Explain how you are deciding.
* What is one observational unit in this study? How many are there?
* Briefly describe the sample that participated in the study. Was random sampling used in selecting the participants for the study?
* Identify the explanatory variable and response variable for these data. Why do you think they used a “treatment” group and a “control” group rather than only a treatment group?
* Was random assignment used in splitting the subjects into the treatment and control groups?

(c) Suppose the researchers found 11 of the 34 subjects who had been given a yawn seed then yawned within the time period, compared to 3 of the 16 subjects who had not been given a yawn seed. Organize this information in the two-way table in your lab report.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | *Explanatory variable* | |  |
|  |  | Seed | NoSeed | Total |
| *Response variable* | Yawn | *>>* | *>>* | *>>* |
| NoYawn | *>>* | *>>* | *>>* |
|  | Total | *>>* | *>>* | *>>* |

(d) *Numerical summary:* To numerically compare the two groups, calculate (by hand) the conditional proportion of participants who yawned in each treatment group, and observed difference as follows:

Of the 34 subjects in the Seeded group, what proportion of these subjects yawned?

*>>*

Of the 16 subjects in the Control (no seeding) group, what proportion of these subjects yawned?

*>>*

Calculate the observed difference in these conditional proportions:

Seeded Group Proportion – Control Group Proportion = Observed Difference

(e) *Graphical summary:* Use technology to create a segmented bar graph or mosaic plot comparing the yawning rates between the two groups (e.g., Analyzing Two-way Tables applet <http://www.rossmanchance.com/applets/ChisqShuffle.htm>).

Using the [Analyzing Two-way Tables](http://www.rossmanchance.com/applets/ChiSqShuffle.html?FET=1) applet.

|  |  |  |
| --- | --- | --- |
| * Please **Clear**. * Paste in the raw data and press **Use Data** or enter the titles and counts of a two-way table and press **Use Table**. (Or check the 2 ×2 box and enter the cell values.) * Use the pull-down menu to toggle between Bar graph and Mosaic plot. * Verify the observed difference in the proportion who yawn. |  |  |

*Introduce the graph to the reader/make a title>>*

*Copy and paste your graph into the space below.*

(f) Summarize what these numerical and graphical summaries reveal about the sample. In particular,

* How often did people yawn in this study?
* What did you learn about the differences between these two treatment groups? (cite your numerical and graphical evidence)
* Do the data appear to provide preliminary evidence for the claim that yawning is contagious? (are the observed results at least in the conjectured direction?)

*>>*

**Side Notes:** There is a difference between saying one treatment is effective (e.g., did most people in the group yawn) and saying one treatment is *more effective* than another (e.g., was the conditional proportion of success higher in one group than the other). There is also a difference between saying “a difference of 10%” and “a difference of 10 percentage points” – you mean the latter.

## Thought Questions (Think about briefly, discuss with partner, and then move on)

The above descriptive analysis tells us what we have learned about the 50 subjects in the study. But can we make any inferences beyond what happened in this study?

* What are the two possible explanations for why there were more yawners in the yawn seed group? (Why aren't we worried about confounding variables?)
* Consider the "random chance" explanation. Is it possible that the higher rate of yawning in the yawn seed group could have risen just from an unlucky random assignment? Do you think it was probable?
* Suggest a way to explore how we could use index cards to determine whether random assignment alone could have lead to this large of a difference.
  + Note: We will "fix" the number of yawners and non-yawners because we are assuming they will have yawned or not regardless of which group they were randomly assigned into.
  + How many cards will you have? How many will be blue and how many green? How many will you deal out to represent each treatment group?
* Why aren't we answering this question with coin flips anymore?

***Simulation under the null hypothesis***

(g) Record the results of your ONE “could have been” random assignment with the colored cards in the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Outcome** | **Treatment Group** | | **Total** |
| **Yawn Seed** | **No Seed** |
| **Yawn** | *>>* | *>>* | 14 |
| **No Yawn** | *>>* | *>>* | 36 |
| **Total** | 34 | 16 | 50 |

(h) Compute the difference in the proportion of yawners in each treatment group for this table. (Show your work.)

*>>*

(i) How does the "could have been" difference in proportions you simulated in (g) compare to the actual observed difference in conditional proportions found by the Mythbusters (more extreme or less extreme)?

*>>*

Using the [Analyzing Two-way Tables](http://www.rossmanchance.com/applets/ChiSqShuffle.html?FET=1) applet (with 2x2 box).

|  |  |
| --- | --- |
| * Check the **Show Shuffle Options** box. * Select the **Cards** display. * Press **Shuffle**. |  |

(j) Using the applet, examine one could-have-been outcome resulting from the random assignment of the subjects to the two treatment groups under the null hypothesis that the yawn seed did not make a difference. Report and verify (show your work) the value reported for the difference in proportions based on the generated table, in the space below.

*Paste a screen capture of your simulated two-way table here.*

*Calculation details of difference in simulated conditional proportions >>*

(k) Examine a second simulation of the random assignment process under the null hypothesis.

*Paste a screen capture of your second simulated two-way table here.*

Report the difference in the conditional proportions of the two treatment groups. Is the difference in conditional proportions this time the same as that obtained in (j)? Did you expect them to be the same? Explain.

*>>*

(l) Now look at a large number of such shuffles to see the long-term pattern in the results.

*Introduce the graph to the reader/make a title>>*

*Copy and paste in the space below a screen capture of the null distribution dotplot generated with the shaded values and the approximate p-value displayed.*

(m) What values are shaded red for the p-value and why? [Hint: If you had shaded them yourself, how would you know which ones to shade?]

*>>*

**Analysis**

(n) For the simulation you created, map the components to the research study:

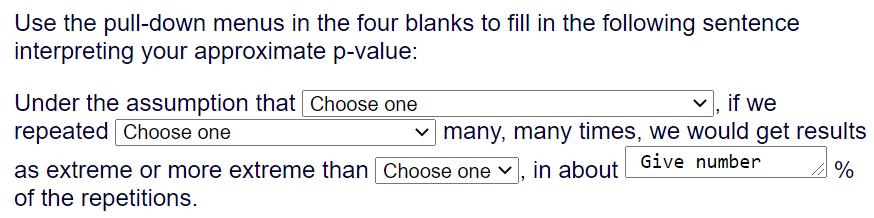
|  |  |
| --- | --- |
| Null hypothesis: | *>>* |
| One repetition: | *>>* |
| Statistic: | *>>* |

STOP: The previous question is pretty important, ask the instructor if you are not sure.

(o) Confirm that the mean of your dotplot results is approximately zero. Explain why it makes sense for the observations in this dotplot to have a mean of about zero.

*>>*

(p) *Interpret* the resulting p-value in context (see hints in online lab):



STOP: The previous question is pretty important, ask the instructor if you are not sure.

(q) *Evaluate* the p-value: Does this p-value provide strong evidence against the null hypothesis? Clearly indicate how you are deciding. [Remember the guidelines from Lab 1.]

*>>*

We actually have a small confession to make. We fudged the data from this study a bit! Recall from the video that the actual data showed that 10 of the 34 subjects in the yawn seed group (29.4%) and 4 of the 16 subjects in the control group (25.0%) yawned.

(r) Calculate the observed statistic for this study. How has it changed from the statistic you used earlier? Predict how this will change the p-value of the study and therefore the strength of evidence against the null hypothesis.

*>>*

(s) Determine the p-value for the actual study. Was your prediction correct?

*>>*

**Conclusions**

(t) Write a paragraph summarizing your analysis **of the actual study result** and addressing the following issues:

*Significance*: In light of your answers to the previous question, would you say the results obtained by the researcher’s provide strong evidence that yawn seeding is more effective than the no-yawn seeding in creating yawns? Do you agree with the Mythbusters’ conclusion that there is “little doubt, yawning seems to be contagious”? Explain, as if discussing this with the hosts Jamie and Adam!

*Causation*: Also consider the design of the study. Is it appropriate to draw a cause-and-effect conclusion here? Justify your answer.

*Generalizability*: Again consider the design of the study. To what population are we willing to generalize the results of this study? Justify your answer.

**Exact p-value**

The simulation you conducted approximated the p-value for two-way tables arising from random assignment by assuming the row and column totals are fixed. In this case, the probability of obtaining a specific number of successes in one group can be calculated exactly using the *hypergeometric* probability distribution. Using the hypergeometric distribution to calculate the p-value is referred to as *Fisher’s Exact Test*.

P(X = 10) = C(14, 10) × C(36, 24) / C(50, 34) = 0.2545.

p-value = P(X = 10) + P(X = 11) + P(X = 12) + P(X = 13) + P(X =14).

Why do we stop at 14?

Sum all five probabilities together (including P(X = 10)) to determine the exact p-value for the yawning study. How does this p-value compare to the empirical p-value from the applet simulation? Write a one or two sentence interpretation of this p-value.

Exact p-value: Comparison:

Interpretation:

In the applet, check the box to **Show Fisher’s Exact Test**.

**Activity 13: Lingering Effects of Sleep Deprivation** (comparing two means)

Researchers have established that sleep deprivation has a harmful effect on visual learning. But do these effects linger for several days, or can a person “make up” for sleep deprivation by getting a full night’s sleep in subsequent nights? A recent study (Stickgold, James, and Hobson, 2000) investigated this question by randomly assigning 21 subjects (volunteers between the ages of 18 and 25) to one of two groups: one group was deprived of sleep on the night following training and pre-testing with a visual discrimination task, and the other group was permitted unrestricted sleep on that first night. Both groups were then allowed as much sleep as they wanted on the following two nights. All subjects were then re-tested on the third day. Subjects’ performance on the test was recorded as the minimum time (in milli-seconds) between stimuli appearing on a computer screen for which they could accurately report what they had seen on the screen.

(a) Identify the explanatory and response variables in this study. Also classify them as either categorical or quantitative.

(b) Was this an experiment or an observational study? Explain how you are deciding.

The sorted data and dotplots presented here are the improvements in those reaction times (in milliseconds) between the pre-test and post-test (a negative value indicates a decrease in performance):

Sleep deprivation (n = 11): -14.7, -10.7, -10.7, 2.2, 2.4, 4.5, 7.2, 9.6, 10.0, 21.3, 21.8

Unrestricted sleep (n = 10): -7.0, 11.6, 12.1, 12.6, 14.5, 18.6, 25.2, 30.5, 34.5, 45.6



(c) Does it appear that subjects who got unrestricted sleep on the first night tended to have higher improvement scores than subjects who were sleep deprived on the first night? Explain briefly.

(d) Calculate the median of the improvement scores for each group. Is the median improvement higher for those who got unrestricted sleep? By a lot?

(e) What are two possible explanations for the tendency for higher improvement groups for the unrestricted group?

The dotplots and medians provide at least some support for the researchers’ conjecture that sleep deprivation still has harmful effects three days later. Nine of the ten lowest improvement scores belong to subjects who were sleep deprived, and the median improvement score was more than 12 milli-seconds better in the unrestricted sleep group (16.55 ms vs. 4.50 ms). The average (mean) improvement scores reveal an even larger advantage for the unrestricted sleep group (19.82 ms vs. 3.90 ms). But before we conclude that sleep deprivation is harmful three days later, we should consider once again this question:

(f) Is it possible that there’s really no harmful effect of sleep deprivation, and random chance alone produced the observed differences between these two groups?

You will notice that this question is very similar to questions asked of the helper/hinderer toy study. Once again, the answer is yes, this is indeed possible. Also once again, the key question is how likely it would be for random chance to produce experimental data that favor the unrestricted sleep group by as much as the observed data do.

(g) In words, state the null and the alternative hypotheses to investigate whether sleep deprivation has a negative effect on improvement in performance on visual discrimination tasks. (*Hint*: For the alternative hypothesis: Do you expect the people to do better or worse when sleep deprived? Based on your answer, what sign/direction should you choose for the alternative hypothesis?)

(h) Describe how you might go about deciding whether the observed difference between the two sample medians in this randomized experiment is statistically significant.

We will answer that question using the same 3S simulation strategy that we used earlier:

**1. Statistic:** We need a measure of the difference in the centers of the two groups, such as the difference in medians (or means).

**2. Simulate:** We will assume that the null hypothesis is true and replicate the random assignment process a large number of times, in order to get a sense for what’s expected and what’s surprising for values of the statistic under the null model.

**3. Strength of evidence?** If the result observed by the researcher is in the tail of the null model’s “what if” distribution, we will reject that null model.

Because the null hypothesis asserts that improvement score is not associated with sleep condition, we will assume that the 21 subjects would have had exactly the same improvement scores as they did, *regardless* of which sleep condition group (unrestricted or deprived) the subject had been assigned.

But there’s an important difference here as opposed to those earlier studies: the data that the researchers recorded on each subject are not yes/no responses. In this experiment the data recorded on each subject are numerical measurements: improvements in reaction times between pre-test and post-test. So, the complication this time is that after we do the random assignment, we must do more than just count yes/no responses in the groups.

What we will do instead is, after each new random assignment, calculate the median improvement in each group and determine the difference between them. After we do this a large number of times, we will have a good sense for whether the difference in group medians actually observed by the researchers is surprising under the null model of no real difference between the two groups (no treatment effect). Note that we could just as easily use the means instead of the medians, which is a very nice feature of this analysis strategy.

One way to implement the simulated random assignment is to use 21 index cards. On each card, we would write one subject’s improvement score. Then shuffle the cards and randomly deal out 11 for the sleep deprivation group, with the remaining 10 for the unrestricted sleep group.

This time will turn to technology right away:

(i) Go to the **two means** [applet](http://www.rossmanchance.com/applets/AnovaShuffle.htm?hideExtras=2), change the statistic to **Difference in Medians**,check the **Show Shuffle Options** box, select the **Plot** display, and press **Shuffle Responses**. What is the applet doing?

(j) Press Shuffle Responses again to repeat the re-randomization process. Did you obtain the same result as before?

(k) Now ask for 998 more re-randomizations (for 1000 total). Look at the distribution of the 1000 simulated differences in group medians. Is the center where you would expect? Does the shape have a recognizable pattern? Explain.

(l) Count how many of your 1000 simulated differences in group medians are as extreme (or more extreme), in the direction favoring the unrestricted sleep group, as the researchers’ actual result (which was?) and determine the approximate p-value by calculating the proportion of your simulated differences that are at least this extreme.

(m) Do these simulation analyses reveal that the researchers’ data provide strong evidence that sleep deprivation has harmful effects three days later? Explain the reasoning process by which your conclusion follows from the simulation analyses.

(n) Based on your analysis and the study design, it legitimate to draw a cause-and-effect conclusion between sleep deprivation and lower improvement scores? To what population are you willing to generalize these results? Explain.

**Possible Assessment/Follow-up Questions:**

(o) Redo this simulation analysis, using the difference in group means rather than medians. When you re-run the simulation, you will have to tell the computer or calculator to calculate means rather than medians. Report the approximate p-value, and summarize your conclusion. Also describe how your conclusion differs (if at all) from this analysis as compared to the analysis based on group medians.

(p) What does one dot on the dotplot represent? (*Hint*: Think about what you would have to do to put another dot on the graph.)

(q) How would you *interpret* the p-value (*Hint*: Describe the process you used to obtain it.)

(r) Use the 2SD Method to approximate a 95% confidence interval for the difference in long-run mean improvement score for subjects who get unrestricted sleep minus the long-run mean improvement score for subjects who are sleep deprived. (*Hints*: Remember the observed value of the difference in group means, and obtain the SD of the difference in group means from the applet’s simulation results. The interval should be observed difference in means ± 2SD, where SD represents the standard deviation of the null distribution of the difference in group means.) Interpret what this confidence interval reveals, paying particular attention to whether the interval is entirely positive, entirely negative, or contains zero. (*Hint*: Be sure to mention “direction” in your interpretation by saying how much larger improvement scores are on average for the treatment you find to have the larger long-run mean: I’m 95% confident that the long-run mean improvement score is to higher with the \_\_ \_\_\_\_\_\_\_\_ treatment as opposed to the \_\_\_\_\_\_\_\_\_\_\_\_ treatment.)

(s) Looking back and ahead: Did anything about the design and conclusions of this study concern you?What should the researchers’ next steps be to fix the limitations or build on this knowledge?

(t) Carry out a two-sample *t*-test for these data. Are the p-values and conclusions from these two approaches (simulation of randomization test vs. conventional test) similar? Discuss advantages and disadvantages of these two approaches to significance testing.

**Activity 14: Cat households** (statistical significance vs. practical importance)

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 15: Senators** (limitations of statistical inference)

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.