Section 4.1 Part 1: Bad Sampling & Bias, Simple Random Sampling

Today, you will learn how to:
- Identify the population and sample in a statistical study
- Identify convenience sampling and voluntary response sampling, & explain how they lead to bias
- Describe how to obtain a random sample using the Hat Method, Technology, or a table of random digits (TABLE D)

Activity: CRAZY IN LOVE

1. Quickly and individually choose a random sample of 5 words from the Crazy in Love lyrics sheet. Write them below. How many letters are in each word?

2. What is the average word length of your sample? \( \bar{x} = 5 \)

3. Put your average on the dot plot on the white board at the front of the room. Copy the class dot plot below.

4. Find a new sample of 5 words using a random number generator. Put your average on the dot plot on the white board at the front of the room. Copy the class dot plot below.

5. How is the dot plot from #4 different than the dot plot for #3? Which do you think is a better estimator of the true mean word length?

6. It is known that Beyoncé wrote the lyrics for all of the Destiny’s child songs. The average word length for these songs is 3.64 letters. Based on your samples, do you have good evidence that Beyonce did not write the lyrics for “Crazy in Love”. Explain.

No, we don’t. The true mean word length is 3.53 letters for Crazy in Love, which is close to Destiny Child’s mean word length of 3.64 letters. We don’t have evidence that she didn’t write Crazy in Love.
Big Ideas from Section 4.1: Key Terminology & Bad Sampling

<table>
<thead>
<tr>
<th>Population: entire group of individuals</th>
<th>Bias: consistently overestimating or underestimating population parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>population size notation is N</td>
<td>Convenience Sample: sampling individuals who are easy to reach</td>
</tr>
<tr>
<td>Census: Choosing data from every individual in population</td>
<td>Voluntary Response: people choose themselves to be part of a sample by responding to a general invitation</td>
</tr>
<tr>
<td>Sample: subset of individuals randomly chosen from the population that represents the population (easier, faster, cheaper)</td>
<td></td>
</tr>
</tbody>
</table>

Check Your Understanding Part 1 (warm up) 10/17/18

1. In June 2008 Parade magazine posed the following question: “Should drivers be banned from using all cell phones?” Readers were encouraged to vote online at www.parade.com. The July 13, 2008, issue of Parade reported the results: 2407 (85%) said “Yes” and 410 (15%) said “No.”

   a. What type of sample did the Parade survey obtain? Voluntary Response Sample

   b. Explain why this sampling method is biased. This sampling method is biased because only people who are passionate about the ban will call in. They don’t represent the population.

   c. Is 85% likely to be greater than or less than the percentage of all adults who believe that cell-phone use while driving should be banned? Why? 85% is likely greater than the true proportion because people who call in feel strongly that they should be banned. People who don’t care wouldn’t call.

2. To help eliminate bias, a reporter from Parade decides she will go out and ask people in person if they think drivers should be banned from using cell phones. She lives close to the local high school so she goes to the parking lot at 3:00 pm and asks the first 100 people she sees.

   a. What type of sample did the reporter obtain? Convenience sample

   b. Explain why this sampling method is biased. This sampling method is biased because most of the people she talks to are probably students, which don’t represent the population.

   c. How could Parade magazine avoid the bias described above? They should do a simple random sample from the population.
How to Sample Well: SIMPLE RANDOM SAMPLING (Sampling Method #1)

Let chance choose the sample!

Let’s write out 3 ways that we can choose a simple random sample for the Beyoncé’s Crazy in Love:

A. Choosing an SRS with the Hat Method:

Number the words in the song “Crazy in Love” from 1 to 297 and write them on 297 identical slips of paper. Place the identical slips of paper in a hat and mix thoroughly. Without looking at the hat, select 5 different slips of paper. The words that correspond to the 5 numbers selected will be chosen for the sample.

B. Choosing an SRS with Table D (the Table of Random Digits)

* Three digits are needed to label the 297 words *

Label the words 001 to 297 in the order they are sang. Using line ___ in Table D, select 3-digit numbers between 001 to 297 reading from left to right until five different numbers are selected. Skip any repeated numbers or numbers outside the range 001-297. The five words that correspond to the 5 numbers obtained will be chosen for the sample.

C. Choosing an SRS with Technology:

Label the words 1 to 297 in the order they are sang. Use a random number generator to select five unique numbers from 1 to 297. Skip any repeated numbers or numbers outside of 1 to 297. The five words that correspond to the 5 numbers obtained will be chosen for the sample.
Check Your Understanding

1. Suppose you want to survey 100 students at WGHS about ___________________________. You are given a list of names for all students who attend WGHS. Describe a simple random sampling procedure using a hat method to obtain a random sample of 100 students for you to survey.

Label all student names alphabetically from 1 to 1711, and write the numbers on 1711 identical slips of paper. Place the identical slips of paper in a hat and mix thoroughly. Select 100 different slips of paper without replacement. The students that correspond to the 100 numbers will be selected for the sample.

2. An administrator at a large university wants to conduct a survey to estimate the proportion of students who are satisfied with the appearance of the university buildings and grounds. Describe a simple random sampling procedure using a random number generator that the administrator could use to obtain a sample of 500 students from the 70,000 students at the university.

Label all student names alphabetically from 1 to 70,000. Use a random number generator to select 500 unique numbers between 1 and 70,000. Skip any repeated numbers. The students that correspond to the 500 numbers will be selected for the university's survey.

3. The management company of a local mall with 21 stores plans to survey 3 of them to determine the hours they would like to stay open during the holiday season. Use Table D at line 101 to select an SRS of 3 stores.

Label the stores from 01 to 21. Using line 101 in Table D reading from left to right, select two-digit numbers between 01 and 21 until three different numbers are selected. Skip any repeated numbers or numbers outside of 01 – 21. The three stores that correspond to the numbers obtained will be surveyed.

SRS: 19, 05, 13 correspond to Star Jewelers, Carter's, & Mrs. Fields.

4. The principal of Thomas Jefferson Elementary School wants to assess reading achievement of third graders. He asks the vice principal to put together a list of all the third graders. Then, the principal randomly selects a student from the first three students on the list. Starting with that student, the principal selects every third student for the assessment. For example, if student number 2 were the first student selected, the sample would consist of students number 2, 5, 8, 11, 14, etc. Is this an example of simple random sampling? Explain.

No, this is not an example of simple random sampling because each possible sample is not equally likely to occur. For example, if student #2 were in the sample, then the sample would never include students 1 or 3.
Crazy in Love (Beyoncé)

I look and stare so deep in your eyes
I touch on you more and more every time
When you leave I'm begging you not to go
Call your name two or three times in a row
Such a funny thing for me to try to explain
How I'm feeling and my pride is the one to blame
'Cuz I know I don't understand
Just how your love can do what no one else can

Got me looking so crazy right now, your love's
Got me looking so crazy right now (in love)
Got me looking so crazy right now, your touch
Got me looking so crazy right now (your touch)
Got me hoping you'll page me right now, your kiss
Got me hoping you'll save me right now
Looking so crazy in love's
Got me looking, got me looking so crazy in love

When I talk to my friends so quietly
Who he think he is? Look at what you did to me
Tennis shoes, don't even need to buy a new dress
If you ain't there ain't nobody else to impress
The way that you know what I thought I knew
It's the beat my heart skips when I'm with you
But I still don't understand
Just how the love your doing no one else can

I'm Looking so crazy in love's
Got me looking, got me looking so crazy in love

Got me looking, so crazy, my baby
I'm not myself, lately I'm foolish, I don't do this
I've been playing myself, baby I don't care
'Cuz your love's got the best of me
And baby you're making a fool of me
You got me sprung and I don't care who sees
'Cuz baby you got me, you got me, so crazy baby
HEY!
| 1 I   | 51 and | 101 your | 151 so | 201 I'm | 251 this |
| 2 look | 52 my  | 102 touch| 152 quietly | 202 with | 252 I've |
| 3 and | 53 pride| 103 Got | 153 Who | 203 you | 253 been |
| 4 stare| 54 is  | 104 me | 154 he | 204 But | 254 playing |
| 5 so  | 55 the | 105 looking | 155 think | 205 I | 255 myself |
| 6 deep | 56 one | 106 so | 156 he | 206 still | 256 baby |
| 7 in  | 57 to | 107 crazy | 157 is | 207 don't | 257 I |
| 8 your | 58 blame | 108 right | 158 Look | 208 understand | 258 don't |
| 9 eyes | 59 'Cuz | 109 now | 159 at | 209 Just | 259 care |
| 10 I  | 60 I  | 110 (your | 160 what | 210 how | 260 'Cuz |
| 11 touch | 61 know | 111 touch) | 161 you | 211 the | 261 your |
| 12 on | 62 I | 112 Got | 162 did | 212 love | 262 love's |
| 13 you | 63 don't | 113 me | 163 to | 213 your | 263 got |
| 14 more | 64 understand | 114 hoping | 164 me | 214 doing | 264 the |
| 15 and | 65 Just | 115 you'll | 165 Tennis | 215 no | 265 best |
| 16 more | 66 how | 116 page | 166 shoes | 216 one | 266 of |
| 17 every | 67 your | 117 me | 167 don't | 217 else | 267 me |
| 18 time | 68 love | 118 right | 168 even | 218 can | 268 And |
| 19 When | 69 can | 119 now, | 169 need | 219 I'm | 269 baby |
| 20 you | 70 do | 120 your | 170 to | 220 Looking | 270 you're |
| 21 leave | 71 what | 121 kiss | 171 buy | 221 so | 271 making |
| 22 I'm | 72 no | 122 Got | 172 a | 222 crazy | 272 a |
| 23 begging | 73 one | 123 me | 173 new | 223 in | 273 fool |
| 24 you | 74 else | 124 hoping | 174 dress | 224 love's | 274 of |
| 25 not | 75 can | 125 you'll | 175 if | 225 Got | 275 me |
| 26 to | 76 Got | 126 save | 176 you | 226 me | 276 You |
| 27 go | 77 me | 127 me | 177 ain't | 227 looking, | 277 got |
| 28 Call | 78 looking | 128 right | 178 there | 228 got | 278 me |
| 29 your | 79 so | 129 now | 179 ain't | 229 me | 279 sprung |
| 30 name | 80 crazy | 130 Looking | 180 nobody | 230 looking | 280 and |
| 31 two | 81 right | 131 so | 181 else | 231 so | 281 I |
| 32 or | 82 now, | 132 crazy | 182 to | 232 crazy | 282 don't |
| 33 three | 83 your | 133 in | 183 impress | 233 in | 283 care |
| 34 times | 84 love's | 134 love's | 184 The | 234 love | 284 who |
| 35 in | 85 Got | 135 Got | 185 way | 235 Got | 285 sees |
| 36 a | 86 me | 136 me | 186 that | 236 me | 286 'Cuz |
| 37 row | 87 looking | 137 looking, | 187 you | 237 looking, | 287 baby |
| 38 Such | 88 so | 138 got | 188 know | 238 so | 288 you |
| 39 a | 89 crazy | 139 me | 189 what | 239 crazy | 289 got |
| 40 funny | 90 right | 140 looking | 190 I | 240 my | 290 me, |
| 41 thing | 91 now | 141 so | 191 thought | 241 baby | 291 you |
| 42 for | 92 (in | 142 crazy | 192 I | 242 I'm | 292 got |
| 43 me | 93 love) | 143 in | 193 knew | 243 not | 293 me, |
| 44 to | 94 Got | 144 love | 194 It's | 244 myself, | 294 so |
| 45 try | 95 me | 145 When | 195 the | 245 lately | 295 crazy |
| 46 to | 96 looking | 146 I | 196 beat | 246 I'm | 296 baby |
| 47 explain | 97 so | 147 talk | 197 my | 247 foolish, | 297 HEY |
| 48 How | 98 crazy | 148 to | 198 heart | 248 I | 298 you, |
| 49 I'm | 99 right | 149 my | 199 skips | 249 don't | 299 me, |
| 50 feeling | 100 now, | 150 friends | 200 when | 250 do | 300 you, |

Credit: Gabe Yonker
Section 4.1 Part 2: Stratified Random Sampling & Cluster Sampling

Today, you will learn how to:
- Identify the population and sample in a statistical study
- Identify convenience sampling and voluntary response sampling, & explain how they lead to bias
- Describe how to obtain a random sample using the Hat Method, Technology, or a table of random digits (TABLE D)

Activity:
Justin Timberlake’s concert promoter wants to find out how much the 8,000 fans enjoy his concert. He wants to ask attendees, “From 1 to 10, where 10 is the most, how much did you enjoy Justin’s concert?” The concert area surveyed is square and divided into 16 equally sized sections (4 section rows x 4 section columns). Each section holds 500 seats. The stage runs along the Northern edge of the venue (where Justin is pictured). The concert promoter wants to survey a sample of 500 fans.

1. Describe a method to pick a Simple Random Sample of 500 fans.
   Label each fan’s seat from 1 to 8000. Use a random number generator to select 500 different seat numbers. Skip any repeated numbers. The fans that correspond to the 500 unique numbers selected will be surveyed.

2. Describe how to select a Stratified Random Sample of 500 fans. Explain your choice of strata.
   - It is best to stratify by row because people that sit closer to J.T. might enjoy the concert more than people that sit farther away.
   - Stratify the 8000 fans by row sections 1 to 4. Take an SRS of 125 people from each row section. Combine the people selected from each row to form the sample of 500 people to be surveyed.

3. Describe how to select a Cluster Sample of 500 fans. Explain your choice of clusters.
   - It is best to cluster by column section because each column contains fans who are different distances from the stage, which is representative of the entire population of 8000 fans.
   - Cluster the 8000 J.T. fans by column sections 1 to 16. Use a random number generator to obtain one column section of 500 fans. All 500 fans in this column cluster will be surveyed.

4. Which method is most appropriate to select a representative sample of fans?
   The stratified random sample with rows as strata would be best because a fan's distance from J.T.'s stage may affect how much they enjoy his concert. This reduces variability.
<table>
<thead>
<tr>
<th>Simple Random Sample</th>
<th>Stratified Random Sample</th>
<th>Cluster Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Choosing a sample n from the population N so that every individual or group of individuals has an equal chance of being selected.</strong></td>
<td><strong>Stratifies (splits) the population into groups (strata) that are different between but similar within. Then take an SRS from each stratum, and combine SRS’s to form the sample.</strong></td>
<td><strong>Classify population into clusters of individuals that are located near each other and each have the same variability as the population, then randomly select clusters and take everyone from chosen cluster to form the sample.</strong></td>
</tr>
</tbody>
</table>

**Check Your Understanding**

1. The manager of Oracle arena wants to know the financial status of the people who are attending the October 22nd Warriors home game against the Phoenix Suns. He would like to give a survey to a representative sample of 20,000 fans in attendance. Ticket prices for the game vary a great deal: seats near the court cost up to $790, while seats in the top rows of the arena can be as low as $50. The arena is divided into a North, South, East, and West side with 76 total numbered sections. Each section has 52 rows of seats labeled with letters from A (nearest to court) to AA, to B, to BB (nearest the court) .... to Z, to ZZ (top row of arena furthest from court).

   a) Explain why it would be difficult to give the survey to an SRS of 200 fans.

   "This would be very time-consuming. You would have to number every taken seat in the arena, randomly select 200 of them, go to those seats in the arena, find the people who are sitting there, and survey these people. This would take a very long time."

   b) Which would be a better way to take a stratified random sample: using the N, S, E, W sides as strata, or the lettered rows as strata?

   It is best to create strata where the people within the stratum are very similar to each other, but different than people in other strata. In this case, it would be better to take the lettered rows as strata because people within each row have the same ticket prices & distances from the court, but each stratum (row) is different than other strata.

   c) Which would be a better way to take a cluster sample: using the N, S, E, W sides as clusters, or the lettered rows as clusters?

   It is best to select clusters where people in it reflect the variability found in the population. In this case, it would be better to take the N,S,E,W sections as clusters because they include all the different seat prices.
A school librarian wants to know the average number of pages in all the books in the library. The library has 20,000 books, arranged by type (fiction, biography, history, and so on) in shelves that hold about 50 books each.

(a) Explain how to select a simple random sample of 500 books.

Label the books from 1 to 20,000 alphabetically. Use a random number generator to select 500 different numbers between 1 and 20,000. Skip any repeated numbers. The books that correspond to the 500 unique numbers will be chosen for the sample.

(b) Explain how to select a stratified random sample of 500 books. Explain your choice of strata and one reason why this method might be chosen.

- We should stratify by book type (fiction, etc.) because different types of books might be longer or shorter than other types. This would provide a more precise estimate of the average page length than a simple random sample.

- Stratify the library's 20,000 books by type. Take an appropriately sized (proportional) SRS of each type of book genre. Combine the books selected from each type (strata) to form the sample.

(c) Explain how to select a cluster sample of 500 books. Explain your choice of cluster and one reason why this method might be chosen.

- We should form clusters by shelves since each shelf contains mixed types of books of each genre that is representative of the entire population of 20,000 books. This would make it easier for the librarian since it would take less time because the books are already close together.

- Cluster the library's 20,000 books by shelves. Label the shelves from 1 to 400. Use a random number generator to select 10 different shelves using a similar method described in part (a). Combine the 500 books from the 10 selected shelves to form the sample.
A factory runs 24 hours a day, producing wood pencils on three 8-hour shifts—day, evening, and overnight. In the last stage of manufacturing, the pencils are packaged in boxes of 10 pencils each. Each day a sample of 300 pencils is selected and inspected for quality.

a) Describe how to select a stratified random sample of 300 pencils. Explain your choice of strata.

- We should stratify by shift (day, evening, overnight) because the quality of the pencils may be affected by how long the machines have been running & the workers in each shift.

- Stratify all pencils produced in a day by shift they are produced in. Label all pencils in each shift 1 to N, then take an SRS of 100 pencils from each shift, combine the pencils selected from each shift to form the sample of 300 pencils. Inspect these pencils for quality.

b) Describe how to select a cluster sample of 300 pencils. Explain your choice of clusters.

- We should cluster by boxes of 10 pencils because each box is assumed to contain pencils from all 3 shifts.

- Cluster all pencils produced in a day by the boxes that they are packaged in. Label every box of 10 pencils from 1 to N, then take an SRS of 30 boxes. Combine the 300 pencils from the 30 randomly selected boxes to form the sample. Inspect these pencils for quality.

c) Explain a benefit of using a stratified random sample and a benefit of using a cluster random sample in this context.

- A benefit of a stratified random sample is that we get 100 pencils from every shift, which will give us a more precise estimate.

- A benefit of a cluster sample is that it is efficient and simplifies the process. We don't have to label every pencil, just every box.
Lesson 4.1: Day 3: What is wrong with these surveys?

Identify what is wrong in each of these surveys. Be sure to explain.

1. The mayor of Springfield is interested in finding out the average age of people in the city. He obtains a list of all of the landline telephones in the city, and then contacts a simple random sample of 300 people. He uses the data from the sample to estimate the average age of all the people in the city.
   a. What is wrong with this survey?
   He is only contacting people with landline phones. People without landlines aren't surveyed.
   b. Do you think the Mayor will over or underestimate the true mean age of people in Springfield? Why?
   Overestimate. Usually people with landline phones are old.

2. The administration at a school wants to know the proportion of students that did all of their homework last night. They select a simple random sample of 100 students and send an email to each of them asking if they did all of their homework last night. Of the 40 responses, 36 of the students said that they did all of their homework last night (90%).
   a. What is wrong with this survey?
   Only 40 of the 100 responded
   b. Do you think the administration will over or underestimate the true proportion of students who did all of their homework last night? Why?
   Overestimate. Students might lie because the admin is asking. Or they might not respond if they didn't.

3. Boy Scout Peter M. wants to know the proportion of people in his neighborhood who support the Boy Scouts. He takes a random sample of 30 homes and visits them dressed in his uniform.
   a. What is wrong with this survey? He is influencing responses.
   b. Do you think Peter will over or underestimate the true proportion of his neighbors who support the Boy Scouts? Why?
   Overestimate, people may say they support boy scouts even if they don't. Because of him.
Lesson 4.1: Day 3: Sample Surveys: What else can go wrong?

Big Ideas:

Undercoverage: When some members of the population cannot or are less likely to be chosen.

Ex: land lines

Nonresponse: When an individual is selected to be in a sample but chooses not to participate or isn't reached. *Different from voluntary response*

Response Bias: Pattern of inaccurate responses. Could be due to wording, interviewer, lying, etc.

Check Your Understanding:

1. Each of the following is a possible source of bias in a sample survey. Name the type of bias that could result.

(a) The sample is chosen at random from a telephone directory.

  Undercoverage, only people with numbers in the phone directory can be chosen.

(b) Some people cannot be contacted in five calls.

  Nonresponse, some of the sample can't be reached, so they don't respond.

(c) Interviewers choose people walking by on the sidewalk to interview.

  Convenience sample, all the people on the sidewalk could have some commonality.

2. A survey paid for by makers of disposable diapers found that 84% of the sample opposed banning disposable diapers.

Here is the actual question: "It is estimated that disposable diapers account for less than 2% of the trash in today's landfills. In contrast, beverage containers, third-class mail, and yard wastes are estimated to account for about 21% of the trash in landfills. Given this, in your opinion, would it be fair to ban disposable diapers?"

You think the estimate of 84% is less than, greater than, or about equal to the percent of all people in the population who would oppose banning disposable diapers? Explain your reasoning.

84% is likely greater. The wording of the question makes it sound like diapers are not a problem in landfills. The question is leading.