Chapter 2 -- Producing Data

OBJECTIVES FOR THIS CHAPTER

- Differentiate between a population and a sample.
- Differentiate between a parameter and a statistic.
- Introduce the concept that values of a statistic vary.
- Understand various ways bias can enter into the results.
- Demonstrate various types of sampling methods.

IDEAS FOR TEACHING THIS CHAPTER

This chapter starts out with a fun activity to do in the classroom -- the counting the number F's activity on the first 2 pages of Chapter 2. We suggest that you try to structure your lectures so that you can do this activity towards the end of the lecture that you finish up Chapter 1. If you tell your students to read Chapter 2 for next class and start out the class with the counting F's activity, many of the students will have read the answer. The activity can take as little as 5 minutes, or it can be extended by adding a quick graphical display of the results. Here is a brief outline for how you might do this activity: You have just finished Chapter 1. Ask your students to close their textbooks. Explain to your students: "I have a task for you. I need your help in counting something. In a moment I will place a sentence on the overhead and I want you to count the number of times that a certain letter appears in the sentence. I will give you 10 seconds to complete the task. And I will tell you which letter to count in just a moment, but is everyone clear on the directions? The letter to count is F (as in Frank)." Place the transparency with the sentence on the overhead and quietly count to 10, then remove the transparency. Ask your students: "Please tell me how many F's were in the sentence." Gather responses from some students, either volunteers or call on some students. You will certainly get varying answers. As you do, you ask, with a puzzled look on your face, "What happened? How many of you counted 3 F's? 4 F's? 5? 6? Anyone else have a different answer? (You could make a quick graph of the results up front.) Were the directions not clear enough? How come, when we took a complete census of a sentence, we did not all get the same value?" The moral to this activity is that a census, a sample consisting of the entire population, is not foolproof. You can then continue with a discussion about additional reasons why a census may not be possible. After a discussion on parameter versus statistic comes the section on bias, which can be enhanced by bringing in some current articles that present some data or study results and discuss possible sources of bias.

In this chapter, the students learn about the various sampling methods by actually doing them -- sampling their fellow classmates. A few of the Let's Do It! exercises rely on group work. If in Chapter 1 you have made an effort to have students work together at least in groups of 2 or 3, then forming larger groups for these sampling exercises should come fairly easy. Some of the LDI exercises can take 15 to 20 minutes. Thus it is important to watch the time. For our larger classrooms it is not easy to start a group project and have the exact same groups be able to continue on it at the start of the next class. And if you do not allow for a few minutes to wrap up and recap, the main ideas or features of the LDI exercise can be forgotten. If you spend the time on simple random sampling so students have a good understanding of it, the other sampling methods follow more easily and quickly.

This is the first chapter which can make use of a calculator or computer. If you are using a TI-84 graphing calculator, details on how to generate random integers with the TI are presented on page 102 and in the TI Quick Steps appendix which follows the exercises at the end of this chapter (page 144). The steps are fairly straight forward and no data entry is required, so students are not overwhelmed with TI details and options. They become familiar and comfortable with the calculator gradually throughout this text. If you do just one sampling example using the random number table with multiple labels and some labels unused, and then do the same example using the calculator, students see the benefit of using a calculator or computer. The random number generating steps for other TI graphing calculators are similar, and if the same seed value is used, the output should be the same.

Let's Do It (LDI) Solutions

Let's Do It! 2.1 Parameter or Statistic?

According to the Campus Housing Fact Sheet at a Big-Ten University, 60% of the students living in campus housing are in-state residents. In a sample of 200 students living in campus housing, 56.5% were found to be in-state residents. Circle your answer.

(a) In this particular situation, the value of 60% is a (**parameter**, statistic).
(b) In this particular situation, the value of 56.6% is a (parameter, **statistic**).

LDI 2.1

How long? 2 minutes

How might it be done? Ask students to read through the scenario (or you read it aloud with the class), complete the choices, and compare with a neighbor.

How important? We recommend you do this exercise and/or bring in other examples from recent news to share with the class.

Let's Do It! 2.2 Is It Biased?

A television show conducted the following opinion poll:

Should gun control be tougher? Let us know what you, the public, think in a special call-in poll tonight.
If yes, call 1-900-446-6444. If no, call 1-900-446-6445. Charge is 50 cents for the first minute.

Would you consider the results of this opinion poll to be trustworthy? Explain.

No. A call-in poll is typically biased because it is based on a volunteer sample. Only those individuals who are watching the program even have the opportunity to call in. Among those who are watching, individuals who have a strong opinion about the subject are more likely to pay the 50 cents to call.

LDI 2.2

How long? 2-3 minutes

How might it be done? Ask students to read through the scenario and discuss possible answers with a neighbor.

How important? This is a nice, short exercise that reinforces the idea of a voluntary sample. Such call-in polls are common enough, that you may even be on the look out for an actual example to share with the class.

Let's Do It! 2.3 Family Size

A study was conducted to estimate the average size of ***households*** in the U.S. A total of 1000 ***people*** were randomly selected from the population and they were asked to report the number of people in their household. The average of these 1000 responses was found to be 4.6.

(a) What is the population of interest? **All households in the U.S.**
(b) What is the variable of interest? **Size of (number of people in) the household.**
(c) What is the parameter of interest? **Average household size.**
(d) An average computed in the above manner would tend to be larger than the true average size of households in U.S. Explain why this would be the case. **Larger households have more people in the list, so members of a large household are more likely to be selected.**
(e) To better estimate the average size of households in U.S., ***the units*** that should be labeled, and thus sampled from, are ***not*** the individual people, but rather the ___**households**___.

LDI 2.3

How long? 4-5 minutes

How might it be done? Ask students to read through the scenario and discuss possible answers with a neighbor.

How important? This exercise reinforces the ideas of population, parameter, and sampling the wrong unit. This is also an example of length-biased sampling.

Let's Do It! 2.4 A Simple Random Sample of Companies

An investment magazine publishes data on sales, profits, assets, dividends, shares, and earnings per share for the nation's 500 most valuable companies. You are to select a simple random sample of 10 companies from the list of 500 companies. Explain how you would label the companies and then use your calculator (with a seed value of 53) or the random number table (Row 26, Column 1, reading from left to right) to identify the labels of the 10 companies that would be selected from the list of 500 companies.

With the TI: **Label the 500 companies from 1 to 500.**
Using a seed of 53, the labels of the 10 selected companies are: 258, 182, 473, 435,
198, 251, 122, 481, 372, 14.

With the Table: **Give each of the 500 companies 2 labels as follows:**

Company 1:	**001**
Company 2:	**002**
	etc ...
Company 499:	**499**
Company 500:	**500**

Starting at row 26, column 1, the labels of the 10 selected companies are: 815 (skip), **257, 229,** 504 (skip), 839 (skip), 964 (skip), **232, 487**, 882 (skip), 651 (skip), 665 (skip), 661 (skip), **477,** 876 (skip), 797 (skip), **147,** 801 (skip), **330, 087, 074,** 796 (skip), 669 (skip), 572 (skip), 529 (skip), 676 (skip), **205.**

LDI 2.4

How long? 5 minutes

How might it be done? Recap the scenario with the whole class. Most of the time we work through this exercise together as the whole class and then have them work own their own in groups for the next LDI 2.5 exercise. Alternatively if, you spent time going over the material from pages 86 to 89, you might have students work with their neighboring classmate on this exercise to practice taking a simple random sample.

How important? Not completely necessary to do in class, but it is a good basic exercise on taking a simple random sample, to reinforce the basic steps. This exercise could be assigned as a homework problem.

Let's Do It! 2.5 Simple Random Sampling

Form a group of 10 students. The population of interest is your group.
Your task is to **select a simple random sample of size *n* = 3** from your group.

Steps:

1. In the space provided below, write the names of the people in your group.

Labels if using the random number table				**Labels if using a calculator**			
0	Susan	5	Matt	1	Susan	6	Matt
1	Martha	6	Linda	2	Martha	7	Linda
2	Peter	7	Kathy	3	Peter	8	Kathy
3	Brenda	8	Karl	4	Brenda	9	Karl
4	John	9	Albert	5	John	10	Albert

2. Assign a different label to each of the names in your list.
   Be sure that everyone in your group assigns the same label to the same names!
   Note: many ways to label -- and some groups may have fewer or more than 10 in each.
   You may wish to ask students to think about how to label if these are the cases.

3. Select your sample by selecting labels at random.
If you will be using a calculator, use a seed of 21 and your population size N = 10.
If you will be using the random number table, start at Row 13, Column 1.

What is the first label selected? = **TI: 7, Table: 0**
Who is the first person selected from your group? **TI: Linda, Table: Susan**

What is the second label selected? = **TI: 10, Table: 9**
Who is the second person selected from your group? **TI: Albert, Table: Albert**

What is the third label selected? = **TI: 8, Table: 4**
Who is the third person selected from your group? **TI: Kathy, Table: John**

Suppose we wish to learn about the **proportion** of women in your **population**, denoted by p.

Count the number of women in your population, COUNT = **5**
Count the number of people in your population, N = **10**
Compute the proportion of women in your population $p = \text{COUNT}/N =$ **5/10**

Next, let's look at the results for your simple random sample of size n = 3. In many cases, the corresponding symbol computed for the sample is the same as that for the population, but a hat " ^" is written over the top, like this $\hat{p}$ (read p-hat)

Count the number of people in your sample: n = **3**
Count the number of women in your sample, count = **TI: 2, Table: 1**
Compute the proportion of women in your sample, $\hat{p} = \text{count}/n =$ **TI: 2/3, Table: 1/3**

In this example, p is a ___________________ select one: **parameter** or statistic
and $\hat{p}$ is a ___________________ select one: parameter or **statistic**.

Does $\hat{p} = p$? __**No**__ Will this always be the case? __ **No** __

LDI 2.5

How long? 15-20 minutes

How might it be done? There are many ways to approach this exercise so it can be adaptable to fit your needs. Since we have large lecture sections and this is the first sampling exercise, we often take about 5 minutes to briefly go through a various steps with our own mock population of 10 people. You could prepare the transparency before class, having already filled in names (step 1). With the whole class, discuss how you might label the units (step 2 -- if you are only using the table, perhaps ask what if we had 12 in our population), select the first person at random (step 3, using a DIFFERENT seed so they will have to do this themselves in their groups later), and compute your population proportion. Then explain that it is their turn -- to form groups of about 10 (this number may have to be different for some groups) and complete the full exercise. If they have questions, they should first ask their neighboring classmate, and then raise their hand. After most to all groups have finished, gather the class back together to review the last part of step 3. You might finish out your mock example to wrap this up, then go right into a discussion of the think about it questions following LDI 2.5.

How important? Fairly important. Since simple random sampling is used within the remaining sampling methods (random selection within a stratum, of a systematic starting point, of a cluster), it is important that students understand how to do it. By having them do it themselves, with their own population, it reinforces the basic steps.

Let's Do It! 2.6 Accounting Practices

Accountants often use stratified random sampling during audits to verify a company's records of such things as accounts receivable. The stratification is based on the dollar amount of the item, and often includes 100% sampling of the largest items. One company reports 5000 accounts receivable. Of these, 200 are in amounts over $100,000, another 1000 are in amounts between $10,000 and $100,000, and the remaining 3800 are in amounts under $10,000. Using these groups as strata, you decide to verify all of the largest accounts (over $100,000) and to take a simple random sample of 5% of the midsize accounts ($10,000 to $100,000) and 1% of the small accounts (under $10,000).

(a) Based on this sampling design, how many accounts will be sampled?

# of large accounts to be sampled = 100% of 200 **200**
# of midsize accounts to be sampled = 5% of 1000 **50**
# of small accounts to be sampled = 1% of 3800 **38**
----
total # of accounts to be sampled **288**

(b) Describe how you will label the accounts in the ***small*** stratum to select some small accounts to be included in the sample. Use your calculator (seed value = 25) or the random number table (Row 18, Column 1) to select only the first 5 small accounts to be verified. Think carefully about how many small accounts there are to label.

TI: Label the 3800 small accounts from 1 to 3800. Using a seed value of 25 and *N* = 3800, the first 5 selected small accounts have the following labels: 2429, 1511, 2410, 2777, 636.

Table: Label the 3800 small accounts from 0001 to 3800. Starting at row 18, column 1, the first 5 selected small accounts have the following labels:
0101, **1540**, 9233 (skip), **3629**, 4904 (skip), **3127**, **3041**

LDI 2.6

How long?	5-7 minutes
How might it be done?	Recap the scenario with the whole class. Have students work with their neighboring classmate to practice setting up and starting to take a stratified random sample. Sometimes we work through this exercise together with the whole class.
How important?	Not completely necessary to do in class, but it is a good basic exercise on stratified random sample, to reinforce the basic steps. Many students rush into it and make the mistake of taking a sample of 5 small accounts from *some* list of 200 small accounts, instead of sampling from a list of *N* = 3800 small accounts. This exercise could be assigned as a homework problem.

Let's Do It! 2.7 Stratified Random Sampling

Form a group of about eight students. You need to have at least two females and two males in your group. As before, the population of interest is your entire group.

The question of interest: How many times per year do you get a haircut?
We want to learn about the average number of haircuts per year for your population.
(Note: you may come up with a different question of interest.)

In the space provided below, write the names of the people in your group.

Name	# haircuts per year	Name	# haircuts per year
Monica	**3**	**Sarah**	**4**
Jon	**3**	**Emily**	**4**
David	**2**	**Mike**	**5**
Mary	**1**	**Steve**	**2**

Ask the question of interest to each member of your population and record their responses next to their name. Compute the average response for your population. Add up all of the responses and then divide by the number of students in your population, *N*.

$$\textbf{Average} = \frac{\text{SUM}}{N} = \mathbf{24/8 = 3}$$

This number is a (**parameter** , statistic)

You are able to take a sample of size *n* = 4. Take a **simple random sample** of size *n* = 4.

Steps:

1. Assign a different label to each of the names in your list

Label	**Name**	**# haircuts per year**	**Label**	**Name**	**# haircuts per year**
1	**Monica**	**3**	**5**	**Sarah**	**4**
2	**Jon**	**3**	**6**	**Emily**	**4**
3	**David**	**2**	**7**	**Mike**	**5**
4	**Mary**	**1**	**8**	**Steve**	**2**

2. Select a place to start in your random number table (row 10, column 22) and read off labels until 2 different labels have been selected or use your calculator (seed value = 270) to select your sample of size 4 from your population of *N*. Who did you select from your group and what are their responses?

Table (row 10, column 22): 3 / 0 / 6 / 0 / 5 / 9 / 5 / 3 / 3 / 3/ 8 / 8 / 6 / 7 / etc.

=> the selected sample is	**label = 3**	**David with response**	**2**
	label = 6	**Emily with response**	**4**
	label = 5	**Sarah with response**	**4**
	label = 8	**Steve with response**	**2**

TI: (seed of 270): => the selected sample is	**label = 3**	**David with response**	**2**
	label = 1	**Monica with response**	**3**
	label = 7	**Mike with response**	**5**
	label = 6	**Emily with response**	**4**

3. Compute the average response for your simple random sample of size *n* = 4 -- add up the above 4 responses and divide by 4.

Table: $\textbf{Average} = {}^{\text{sum}}\!/_{n}$ **= 12/4 = 3** **TI:** $\textbf{Average} = {}^{\text{sum}}\!/_{n}$ **= 14/4 = 3.5**

This number is a (parameter , **statistic**)

Now you are able to take a sample of size 4, but you want to have 2 females and 2 males in your sample. How? **STRATIFY**!

Steps:

1. In the space provided below, write a list of all the males and all of the females in your group, that is, form the strata. Also include their response next to their name in parentheses, for example, Mary (2).

Label	**FEMALES (Stratum 1)**	**Label**	**MALES (Stratum 2)**
1	**Monica (3)**	**1**	**Jon (3)**
2	**Mary (1)**	**2**	**David (2)**
3	**Sarah (4)**	**3**	**Mike (5)**
4	**Emily (4)**	**4**	**Steve (2)**

2. Assign a label to each unit in each stratum. Note that you can start with the same label for each stratum. For example, if there were four females and four males in your group, the females could be labeled 1 through 4 and the males could be labeled 1 through 4.

3. Select a simple random sample of size *n* = 2 females (start at row 14, column 1 or use your calculator with a seed value = 24) and a simple random sample of size *n* = 2 males (start at row 23, column 20 or use your calculator with a seed value = 35). Record the selected responses below.

Table: Monica and Sarah are selected for ... responses: 3, 4
Jon and Mike are selected for ... responses: 3, 5
TI: Emily and Sarah are selected for ... responses: 4, 4
Jon and Steve are selected for ... responses: 3, 2

4. Compute the estimated average response in each stratum separately:

Table: Stratum 1, Females: Estimated Average = sum/n **= 7/2 = 3.5**

Stratum 2, Males: Estimated Average = sum/n **= 8/2 = 4**

TI: Stratum 1, Females: Estimated Average = sum/n **= 8/2 = 4**

Stratum 2, Males: Estimated Average = sum/n **= 5/2 = 2.5**

5. Compute the **overall sample average response by pooling** the averages from each stratum together. Since the size of the strata may differ, we take a weighted average of the individual stratum averages. Each stratum average is weighted by the proportion of units in the population that make up that stratum.

Overall Sample Average

$$\left(\frac{\#\text{units in stratum1}}{N}\right)\left(\begin{array}{c}\text{stratum1}\\ \text{estimated average}\end{array}\right)+\left(\frac{\#\text{units in stratum2}}{N}\right)\left(\begin{array}{c}\text{stratum2}\\ \text{estimated average}\end{array}\right)=$$

Table: $\left(\frac{4}{8}\right)(3.5)+\left(\frac{4}{8}\right)(4)=3.75$ **TI:** $\left(\frac{4}{8}\right)(4)+\left(\frac{4}{8}\right)(2.5)=3.25$

This number is a (parameter , **statistic**)

How does it compare to the average for the entire population? **Not exactly equal to the population mean of 3.**

LDI 2.7

How long? 15-20 minutes

How might it be done? As with LDI 2.5, there are many ways to approach this exercise. You could take a few minutes to briefly go through a various steps with our own mock population of 8 (or actually any number you wish) people. If you decide to increase the population size, you might also increase the sample size from within each stratum. You could discuss some or all of the steps with your mock population. We have found that if you have spent enough time teaching the details about taking a simple random sample, students do not have much difficulty performing the remaining sampling methods. After a brief introduction to the idea of stratified random sampling and doing LDI 2.6, we have gone right into asking students to form groups and complete the full exercise. You might need to spend some of the time wrapping up this class exercise explaining the weighted average concept (also presented in Example 2.16 which precedes LDI 2.7).

How important? Fairly important. By having them do it themselves, with their own population, it reinforces the basic steps. They see that the labeling is assigned separately within each stratum. They experience computing a weighted average.

Let's Do It! 2.8 Faculty Salaries

A study is being conducted of the faculty salaries for a public university. Of the 1200 tenure-track faculty, 480 are full professors, 336 are associate professors and the remaining are assistant professors. A simple random sample of 100 full professors, 50 associate professors and 50 assistant professors will be taken and information about salary will be obtained.

(a) What type of sampling method is used to obtain the 200 selected faculty members? **Stratified random sampling.**

(b) Use your calculator (with a seed value of 18) or the random number table (Row 14, Column 1) to give the labels for the first 5 full professors to be selected.

TI: 202, 195, 28, 215, and 174.
Table: 103, 112, 267, 339, and 401.

(c) The table summarizes the average salary for the sampled faculty members by rank.

Rank	*Sample Size*	*Average Salary*
Full	100	$95,000
Associate	50	$70,000
Assistant	50	$55,000

Give the overall estimate of the average salary for all faculty members based on this sample information. Show all work and include your units.

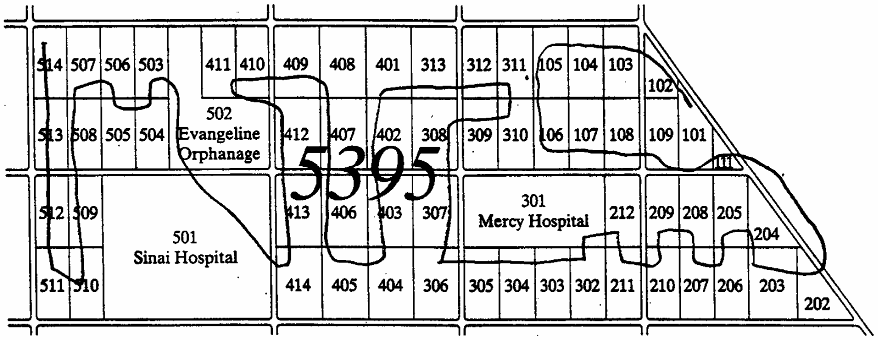
(480/1200)($95,000) + (336/1200)($70,000) + (384/1200)($55,000) = $75,200.

LDI 2.8

How long?	5 - 7 minutes
How might it be done?	Recap the scenario with the whole class. Go through part (a) as a group. It should be somewhat obvious since the exercise is in the Stratified Random Sampling section. You might have students work on parts (b) and (c) in groups of size 2 or 3. When polling the class for their answers to part (c), see if they included the units.
How important?	Not completely necessary to do in class, but it is a good basic exercise on stratified random sample, to reinforce the basic steps. It gives students another opportunity to apply the idea of a weighted average to obtain an overall estimate.

Let's Do It! 2.9 Systematic Sampling of Census Tract Blocks

The accompanying figure is a map of census tract, #5395 in Detroit, Michigan. Census tracts are small homogeneous areas with an average population of about 4000. Each block in the tract is marked with a Census Bureau identifying number.



In the map, the ID numbers start at 101, which is along the right side, and end at 514 in the upper left corner. Notice that some of the numbers in between are skipped. For example, the numbers jump from 414, located in the middle of the bottom edge, to 501 for Sinai Hospital. Interviewers can be assigned to a *set* of blocks, such as the blocks with ID numbers in the 100s. The ID numbers are also assigned in a serpentine pattern so that the blocks that form a set are close to one another.

Your task is as follows: Take a ***1-in-10*** systematic sample of the blocks in this tract.

(a) Following the ID numbers for ordering the blocks in this population, use your pen to ***continue*** to trace out this order on the above map (that is, connect the blocks with a line in ascending order of ID number).

(b) For a 1-in-10 systematic sample the first ten blocks in the map form your first group. Label these 10 blocks and use your calculator (seed value = 39 and *N* = 10) or the random number table (row 16, column 6) to randomly select your starting block, which is the first block in your sample.

What is the ID number of the first block selected?

TI: Labels are #101=1, #102=2, #103=3, #104=4, #105=5,
#106=6, #107=7, #108=8, #109=9, #111=10
Using a seed value of 39 and *N* = 10, the first selected label is 1, which is block #101.

Table: Labels are #101=1, #102=2, #103=3, #104=4, #105=5,
#106 =6, #107=7, #108=8, #109=9, #111=0
Using a row 16, column 6, the first selected label is 1, which is block #101.

(c) Now starting at the first block selected, count off every tenth block, in order, to be included in your sample. List the block ID numbers which form your sample: **#101, #202, #212, #310, #407, #503, #513**

(d) How many blocks are in your sample? ***n* = 7 blocks**
Is the sample size fixed? Explain. **No, the number of blocks in the population is not a multiple of 10. When we started with block #101 we sampled *n* = 7 blocks. However, if you had started with block #104, you would have sampled *n* = 6 blocks.**

LDI 2.9

How long?	8-10 minutes
How might it be done?	Recap the scenario with the whole class. Have students work in groups of size 2 or 3. You might do the tracing to form the list up front, so that everyone starts off with the correct list.
How important?	This is a nice, real-data exercise to do in class. Since the ID numbers are in order but skip some numbers, the students do have to 'count' out every tenth block to be included in the sample. This exercise could be assigned as a homework problem.

Let's Do It! 2.10 Systematic Sampling of Presentation Attendees

An annual meeting for computer programmers was held in a convention center. A total of 1268 computer programmers were in attendance. At 10:00 A.M. on the last day of presentations, the 1268 programmers were attending exactly one of the 20 presentations that they had previously registered for. The 1268 programmers were given consecutive ID numbers based on the registration for the last presentation, as shown:

Presentation	1	2	3	4	5
ID Numbers	1-61	62-85	86-96	97-138	139-150

Presentation	6	7	8	9	10
ID Numbers	151-182	183-240	241-408	409-510	511-544

Presentation	11	12	13	14	15
ID Numbers	545-789	790-816	817-825	826-870	871-892

Presentation	16	17	18	19	20
ID Numbers	893-960	962-1017	1018-1120	1121-1249	1250-1268

The meeting organizers would like to survey the programmers to learn of their impressions regarding the annual meeting. It was decided to take a 1-in-50 systematic sample of the programmers using the ID number.

(a) With this type of sampling plan, is each presentation guaranteed to be represented? That is, will the sample include at least one programmer from each presentation?

Circle one: Yes [✓ No]
Explain: **Some presentations had less than 50 registered.**
If yes, use your calculator with seed value = 18 (or Row 33, Column 1) and give the ID numbers for the programmers that will be selected. If no, give the maximum value for *k* such that the 1-in-*k* systematic sample will always include at least one programmer from each presentation.
The smallest number of programmers registered is 9 for presentation #13, so the maximum value of *k* is 9.

(b) With the correct *k* from part (a), use your calculator with seed value = 18 (or Row 33, Column 1) and give the first 10 ID numbers for the programmers that will be selected.
Using a *k* = 9 and the TI, we have 4, 13, 22, 31, 40, 49, 58, 67, 76, 85.
Using a *k* = 9 and the Table, we have 6, 15, 24,33, 42, 51, 60, 69, 78, 87.

(c) How many programmers will be included in your systematic sample? **We have that 1268/9 = 140 with a remainder of 8. With either the TI or the Table we will sample one more from the last 8 attendees, so the total sample size will be 141. If the first selected digit were a 9, we would have just 140.**

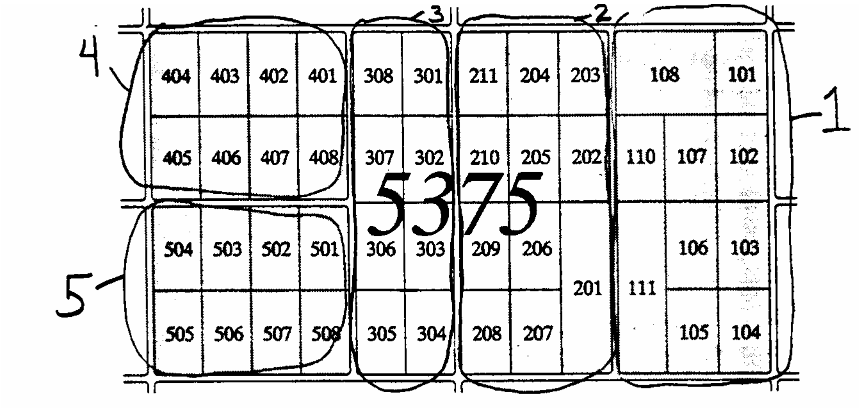
(d) Suppose the organizers would like to survey exactly two programmers from each presentation. Suggest a sampling plan to accomplish this.
They should do a stratified random sample with the 20 presentations representing 20 strata.

LDI 2.10

How long?	10-15 minutes
How might it be done?	Talk through the scenario with the whole class. Have students work in groups of size 2 or 3 to discuss part (a). You might gather the class back together to discuss and arrive at *k* =9, so that everyone can do the remaining parts with the same value for *k*.
How important?	Very Important. Students have to really “think” to complete this exercise. It should generate some good discussion. You could extend part (d) as a homework exercise by having students carry out their suggested plan, showing all details. We have also assigned this as a homework question and given them some time in their weekly lab to discuss it.

Let’s Do It! 2.11 Cluster Sampling of Census Tract Blocks

The following figure is a map of census tract #5375 in Detroit, Michigan.



The blocks in the tract have been grouped to form clusters of blocks, corresponding to the Census Bureau identifying numbers.

In the map there are five clusters of blocks based on the hundredths value for the ID number. Your task is as follows: Take a **cluster sample** by selecting two of the clusters of blocks at random.

(a) Label the 5 clusters. **Cluster Label 1 corresponds to the 100's,**
Cluster Label 2 corresponds to the 200's.
Cluster Label 3 corresponds to the 300's.
Cluster Label 4 corresponds to the 400's.
Cluster Label 5 corresponds to the 500's.

(b) Take a simple random sample of 2 clusters. Use your Calculator (with seed value = 10 and N = 5) or the random number table (Row 24, Column 31) to select two cluster labels at random.
What is the label of the first cluster selected?
What is the label of the second cluster selected?
TI: selected clusters are 3 and 2. **Table: selected clusters are 4 and 5.**

(c) The sample would consist of all of the blocks in those 2 selected clusters. List the ID numbers of the blocks in your sample:
TI: **201, 202, 203, 204, 205, 206, 207, 208, 209, 210, 211,**
301, 302, 303, 304, 305, 306, 307, 308.

Table: **401, 402, 403, 404, 405, 406, 407, 408,**
501, 502, 503, 504, 505, 506, 507, 508.

(d) What is the chance that a block will be selected for the sample? **2/5**

LDI 2.11

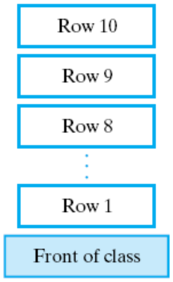
How long? 5-6 minutes

How might it be done? Recap the scenario with the whole class. Have students work in groups of size 2 or 3. Students generally find that this exercise is fairly straight forward. You may wish to remind them that only the clusters of blocks need to receive a label, not the individual blocks.

How important? Again a nice, real-data exercise to do in class, which does not take too much time.

Let's Do It! 2.12 Cluster Sampling of Students

The population of interest is the students in your classroom today. Each row of students will form a cluster.



(a) In the space provided, sketch a map portraying the relative positions of the rows (clusters) in your classroom.

(b) Assign a label to each cluster (each row). What is the chance that any student in the population will be selected?
1/R where R is the number of rows in your classroom.
With 10 rows we have:
TI: **Row #1 receives label 1, Row #2 receives label 2, ... , Row #10 receives label 10.**
Table: **Row #1 receives label 1, Row #2 receives label 2, ... , Row #10 receives label 0.**

(c) Select one cluster at random. Use your calculator with seed value = 279 or Row 9, Column 21 of the random number table.
Which cluster (row) did you select? **TI: select Row #1.** **Table: select Row #1.**
How many students were in the selected cluster (that is, what sample size did you get)?
It will depend on the actual class.
Is the sample size fixed? Explain. **Yes if the number of students in each row is the same for all rows. No if the number of students in each row is not the same for all rows.**

LDI 2.12

How long? 7-8 minutes

How might it be done? We generally do this exercise together as a whole class. If the room arrangement in your classroom is different, you can modify this exercise accordingly, using location between students as a way to form the clusters. Sketch out your resulting map together, discuss and decide on a labeling scheme, and perform the cluster sampling.

How important? A nice exercise that directly uses the whole class and does not take too much time.

Let's Do It! 2.13 Which Sampling Method Could Have Been Used?

A population consists of 12 people, listed in the following table:

	Population			
	Stratum I		*Stratum II*	
Cluster 1	Ann (1)	Bowei (2)	George (3)	Juan (4)
Cluster 2	Carrie (5)	Donna (6)	Bob (7)	Steve (8)
Cluster 3	Latisha (9)	Fran (10)	Paul (11)	Tom (12)

The two columns in the table divide up the population into 2 strata, labeled I and II. The population is also divided into 3 clusters by row. Below each name in the table is the person's corresponding ID number. Thus,

Cluster 1 consists of Ann, Bowei, George, Juan.
Cluster 2 consists of Carrie, Donna, Bob, Steve.
Cluster 3 consists of Latisha, Fran, Paul, Tom.
Stratum I consists of Ann, Bowei, Carrie, Donna, Latisha, Fran.
Stratum II consists of George, Juan, Bob, Steve, Paul, Tom.

A sample of four people was obtained from this population. Listed next are three different samples. Consider the following sampling methods: (1) simple random sampling, (2) stratified random sampling, with equal sample sizes from each stratum, (3) cluster sampling *by rows*. For each sample determine which sampling method(s) could have generated that sample, by circling yes or no for each. *Hint*: more than one method is possible.

		Sampling Method		
		(1) Simple Random Sample?	*(2) Stratified?*	*(3) Cluster?*
(a)	Carrie, Donna, Bob, Steve	[✓ Yes] No	[✓ Yes] No	[✓ Yes] No
(b)	Ann, Fran, Carrie, Bowei	[✓ Yes] No	Yes [✓ No]	Yes [✓ No]
(c)	Carrie, Donna, George, Tom	[✓ Yes] No	[✓ Yes] No	Yes [✓ No]

(d) Take a systematic 1-in-3 sample from this population. Use the ID numbers as the ordered listing of the population items (that is, Ann = 1, Juan = 4, Steve = 5, etc.). Use your calculator with seed value = 78 and *N* = 3 or the random number table with Row 3, Column 3. List the names of the people in your sample.

TI: Starting point is #1, so the sample consists of Ann, Juan, Donna, Fran
Table: Starting point is #1, so the sample consists of Ann, Juan, Donna, Fran

LDI 2.13

How long? 8-10 minutes

How might it be done? Read through the scenario as a whole class. You might go through the possible sampling methods for sample (a) together as a class. Then have the students do sample (b), sample (c), and part (d).

How important? Very important -- a good exercise that tests students' knowledge about the sampling methods and their reasoning skills.

Let's Do It! 2.14 Name that Sampling Method

Read each scenario and identify the sampling method being described (simple random sample, convenience sampling, stratified random sampling, systematic sampling, or cluster sampling). Discuss your answers with your neighboring classmates.

(a) A shipment of 1000 3-ounce bottles of cologne has arrived to a merchant. These bottles were shipped together in 50 boxes with 20 bottles in each box. Of the 50 boxes, 5 boxes were randomly selected. The average content for these 100 selected bottles (that is, all 20 from each of the 5 selected boxes) was obtained.
Method: Cluster Sampling

(b) A faculty member wishes to take a sample from the 1600 students in the school. Each student has an identification number. A list of all identification numbers is available. The faculty member selects an identification number at random from among the first 16 identification numbers in the list, and then every sixteenth identification number on the list from then on.
Method: 1-in-16 Systematic Sampling

(c) A faculty member wishes to take a sample from the 1600 students in the school. The faculty member decides to interview the first 100 students entering her class next Monday morning.
Method: Convenience Sampling

LDI 2.14

How long?	3-5 minutes
How might it be done?	A basic 'name that sampling method' short quiz. Not directly a group exercise, but students enjoy trying out these short basic questions in class too.
How important?	Not crucial, but a quick check on some sampling knowledge.

Think About It (TAI) Solutions

Page 106:

Think about it

When will selecting a simple random sample be simple to do? Will it always be possible?
When will it be difficult to do? Why could it be difficult?
Some of the key ideas are that you need to be able to assign a label to all of the units in the population. Sometimes this is difficult to do or impossible to do. Imagine trying to list all of the adults in a country. We will discuss a multistage sampling method that helps in this area.

How would you label the units if the population size were 78? 292? 4000?
Would it be simpler with the random number table or with a calculator (or computer)?
In general it is easier to use a calculator or computer and have it produce a (random) list of actual labels from 1 to *N*. If you use a random number table, you may end up skipping over a lot of unassigned labels or resort to assigning multiple labels to each unit. Suppose the population size were *N*=292. With a calculator you simply label the units from 1 to 292. With a random number table you use a set of 292 three-digit labels, say 001, 002, ..., 292. The three-digit labels of 000 and 293 through 999 would not be assigned. If you came across any of them in the table, you would skip over them. Alternatively you could assign to each unit a total of 3 three-digit labels. The coding and decoding of this assignment scheme can take some time.

Page 113:

Think about it

When do you take a larger sample size from one stratum versus another?
In general, you take a larger sample from a stratum that has more variability in the responses.

When you form the strata, how should the variability of the units within each stratum compare to the variability between the strata?
Ideally, the variability of the units within each stratum should be small compared to the variability between the strata.

Is a stratified random sample a simple random sample? Explain.
A stratified random sample is not a simple random sample. A stratified random sample, using gender as the stratification variable, will always result in a sample with some females and some males. A simple random sample from the same population could result in a sample which contains only males (or only females). Such a sample could not occur under stratified random sampling. With stratified random sampling, all samples of size *n* do not have the same chance of being selected.

Page 119:

Think about it

In a systematic sample, every unit has the same chance of being selected. Does this imply that a systematic sample is a simple random sample?
A systematic sample is not a simple random sample. In Example 2.19, with systematic sampling the sample AEIMQ has a 1/4 chance of being selected, while the sample ABCDE has no chance of being selected. With systematic sampling, all samples of size *n* do not have the same chance of being selected.

In a systematic sample, the units are effectively divided into groups of size *k* and one unit from each group ends up being in the sample. Does this imply a systematic sample is a stratified random sample?
A systematic sample is not a stratified random sample. With stratified sampling you would select (at least) one unit from each stratum, not necessarily in the same position. In Example 2.19, the sample AEIMQ is possible under both systematic and stratified sampling, while the sample AFKOR could occur with stratified sampling but not with systematic sampling.

Page 125:

Think about it

Is a cluster sample a simple random sample?
A cluster sample is not a simple random sample. A cluster sample, using rows of students as the clusters, will result in a sample with students who sit in the same row. A simple random sample from the same population could result in a sample which contains students from many different rows (and not all of the students from these rows). Such a sample could not occur under cluster sampling. With cluster sampling, all samples of size *n* do not have the same chance of being selected.

Is a cluster sample a stratified random sample?
A cluster sample is not a stratified random sample. Although you could think of the population of students as being divided into strata by rows, with stratified sampling you would select (at least) one unit from each stratum, not necessarily all of the units from a stratum.

When you form the clusters, how should the variability of the units within each cluster compare to the variability between the clusters?
Ideally, the variability of the units within each cluster should be large compared to the variability between the clusters.

Is this criterion the same as in stratified random sampling?
No, in stratified random sampling, you want the variability of the units within each stratum to be small compared to the variability between the strata.

Chapter 2 Extra Examples

Example 2.A:

A survey is carried out at a college to estimate the proportion of undergraduates commuting to school from further than 10 miles for the current semester.

(a) What is the population of interest?
What is the parameter of interest?

The registrar keeps an alphabetical list of all undergraduates, with their current addresses. Suppose there are 20,000 undergraduates in the current semester. Someone proposes to choose a number at random from 001 to 100, count down that far from the top of the list, select that name and every 200th name after it for the sample.

(b) Is this a probability sampling method? (circle one): **YES** **NO**

(c) Is it the same as simple random sampling? Circle one: **YES** **NO**
If yes, explain why.
If no, name the type of method and explain why it is not the same as a simple random sample .

Example 2.B:

The Michigan Secretary of State Office regularly samples the population of cars registered in Michigan, that is, cars having Michigan license plates, to measure the level of and changes in various characteristics. Suppose the Office plans to take a simple random sample of 1,000 cars from the total of 534,322 cars. A list of the license plates registered to the 534,322 cars is available: 099XVD, 502PNB, ... , 567XZP. Explain precisely how you would label the cars, that is, the license plates, in this list in order to take the simple random sample. Use your labeling scheme and a calculator (seed = 62) or the random number table (row 3, column 16) to **list the labels** for the first 3 cars to be included in the sample.

Example 2.C:

A farmer has four orchards of apple trees which are located at different locations on his farm. Each orchard has 200 apple trees. He wishes to find out whether the apple trees are infested with a certain type of insect. If this is so, he would hire a crew to spray his trees. Instead of examining all 800 trees, he decides to select a sample of 80 trees and just examine these. There are three proposed sampling plans described below:

Plan 1: Randomly select 80 trees from the 800 trees.
Plan 2: Randomly select 20 trees from each of the 4 orchards.
Plan 3: Randomly select 2 orchards from the 4 orchards, and then randomly select 40 trees from each of these 2 selected orchards.

(a) For each of the above plans, identify the type of sampling method being proposed.

Plan 1: ____
Plan 2: ____
Plan 3: ____

(b) Which of the three proposed plans would you recommend in this situation? Explain.

Example 2.D:

The Provost of a large university is interested in determining whether the proportions of male and female students who passed proficiency tests in statistics differ significantly. A simple random sample of 250 male student records were obtained from among all 2000 male student records, and a simple random sample of 200 female student records were obtained from among all 1600 female student records. Among the 250 selected male students, 100 passed the proficiency test, for a proportion of 0.40. Among the 200 selected female students, 50 passed the proficiency test, for a proportion of 0.25.

(a) The proportion of 0.40 is a ... (circle one): parameter statistic
Because ....

(b) What type of sampling technique was used to obtain the 450 selected student records?

(c) Use your calculator and seed = 80 to list the labels for the first 5 male student records selected for this study.

If you do not have a TI calculator ...

(1) Use a random number table (or row 10, column 1, left to right) and show all needed steps:

(2) Use the random number generator in your calculator and state which calculator you are using:
(you may be asked to verify this at a later time)

(d) The hypotheses being tested are:

H_0 : Proportion of male students who pass the proficiency test is equal to the proportion of female students who pass the proficiency test.

H_1 : Proportion of male students who pass the proficiency test is *not* equal to the proportion of female students who pass the proficiency test.

Based on the results, the above test was found to be **statistically significant** at a significance level of 0.05.

(i) Which hypothesis was supported?

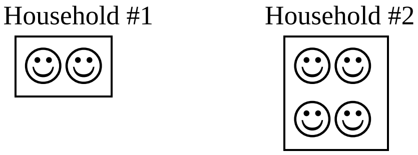
(ii) What can you say about the *p*-value for this test?

(iii) An error could have been made, which type? Type I Type II
What is the chance that this error was made?

Chapter 2 Projects

Project 2.A: Using your Judgment versus Simple Random Sampling Which is Better?

We need to estimate the average size of households for a given community. There are *N* = 100 households in this community. This project will use a picture map of the 100 households along with each household size. For example, shown below are two households. Household #1 has a size of 2, while Household #2 has a size of 4.



Part I. Using Your Best Judgment

1. The instructor will distribute the community map showing the households and their size. Keep the sheet covered until the instructor gives the signal to begin. Then, look at the sheet for ten seconds (the instructor will keep the time) and write down your guess as to the average size of households in this community.

   Your Initial Guess: _______________

2. Select 5 households that, in your judgment, are representative of the households in the community. Write down their numbers and their size. Compute the average size for the 5 households and compare it to your initial guess.

   Sizes for 5 Representative Households: ____

   First Judgment Average (n = 5): __________

3. Now, select 5 more households (all different from the first 5) that, in your judgment, represent the households in the community. Write down their numbers and their size. Compute the average size of the second 5 households and compare it to your initial guess.

   Sizes for Additional 5 Representative Households: ____

   Second Judgment Average (n = 5): __________

4. Compute the average size for the 10 households in your two judgmental samples combined. Compare it to your initial guess.

   Judgment Average (n = 10): __________

   Which estimate of the average size do you like best from among your initial guess, the two averages of 5, or the average of 10? Why?

5. As a class, collect the initial guesses, the first judgment average of 5, the second judgment average of 5, and the judgment average of 10 from each student. Plot each of the four data sets (use the same scale for all plots) and discuss the results.

Initial Guess:

First Judgment Average (n = 5):

Second Judgment Average (n = 5):

Judgment Average (n = 10):

Part II. Simple Random Sampling

1. Using the household numbers (1 though 100, if you use the random number table you may consider them 01 through 00), select 5 households at random. Use your calculator and a seed value of your choice or select a row and column of the random number table. Write down the sizes of the selected households and compute their average.

   Sizes for 5 Households selected at random: ___________________

   First Simple Random Sample Average (n = 5): __________

2. Now, repeat the above step 1 and obtain a second set of 5 households, all different from those found above.

   Sizes for 5 Additional Households selected at random: ________________________

   Second Simple Random Sample Average (n = 5): __________

3. Compute the average size for all 10 randomly selected households.

   Simple Random Sample Average (n = 10): __________

4. As in the case of your judgmental samples, collect the data from each student in the class and construct plots of the sets of averages. Compare your results between Part I and Part II, between *n* = 5 to *n* = 10.

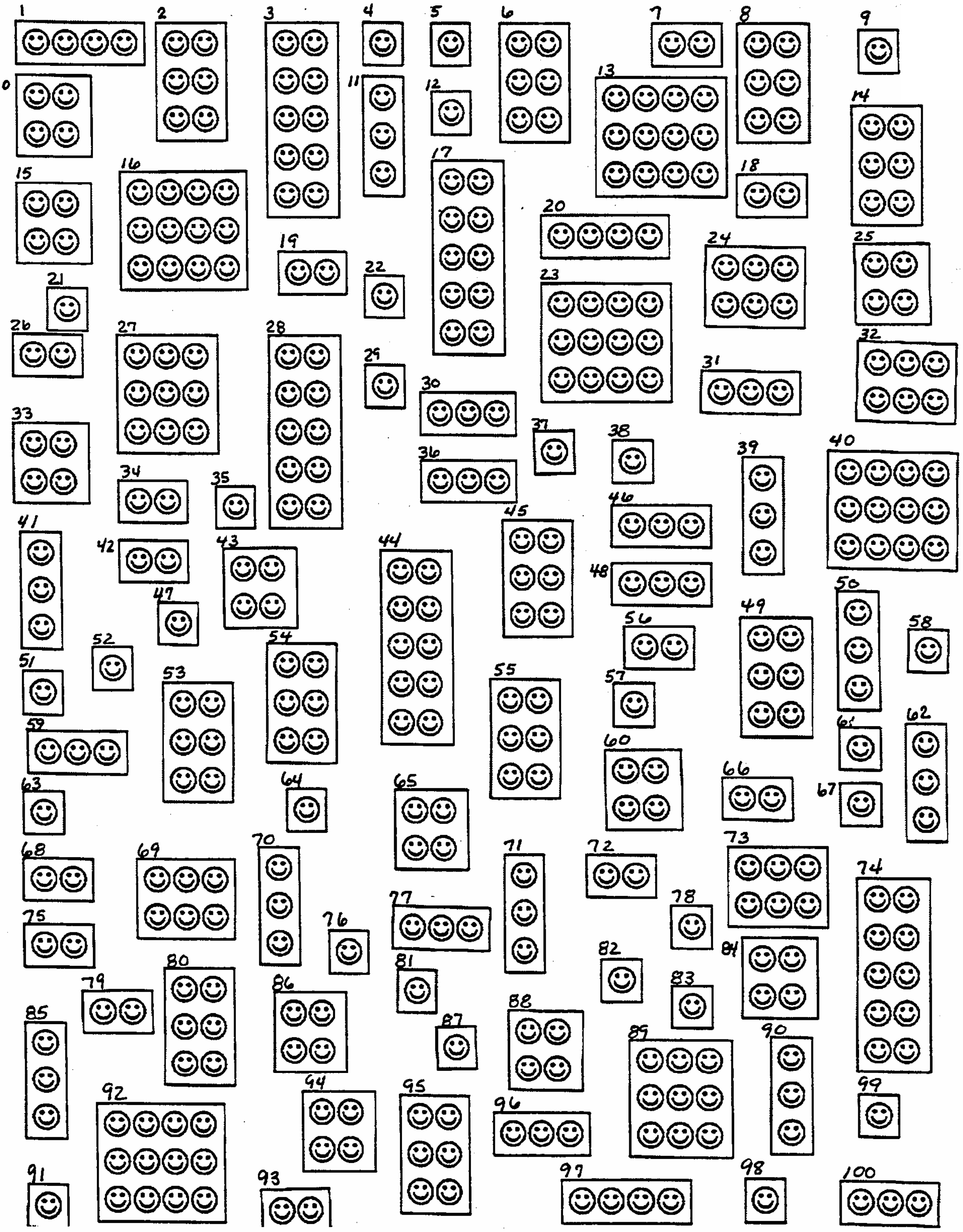
First Simple Random Sample Average (n = 5):

Second Simple Random Sample Average (n = 5):

Simple Random Sample Average (n = 10):

Part III. Write up an overall summary of your findings.

Community Map of Households



Project 2.B: Sampling People versus Sampling Households
Size or Length Bias

In Let's Do It! 2.3 we discussed a study that was conducted to estimate the average size of *households* in the U.S. A simple random sample of *people* were selected from the population and they were asked to report the number of people in their household. The average of these responses was used to estimate the average household size for the entire population. However, such an average would tend to be larger than the true average size of households in U.S. because the larger households have more people in the list, so members of a large household are more likely to be selected. This was called **size or length bias sampling**, the larger (or longer) units have a better chance of being included in the sample. To better estimate the average size of households in U.S., *the units* that should be labeled, and thus sampled from, are *not* the individual people, but rather the *households*.

Let's see this bias in action.

We need to estimate the average size of households for a given community. We have a list of all the 382 people living in this community. The list is provided on the next page along with the number of people in their household.

Part I. Simple Random Sample of People

1. Each person in the population has a label starting at 1 to 382.
   Take a simple random sample of n = 20 people and record their responses below.
   Use a seed value of your choice or pick a row and column of the random number table.

   Responses of the 20 selected people:

2. Compute the average response for your simple random sample of 20 people.

   Average = SUM/n =

3. Collect the averages from each student or group in the class. Plot the averages and describe it.

Average for Simple Random Sample of people:

Part II. Simple Random Sample of Households

1. There are 100 households in the community population which are labeled 1 to 100 on the community map. Take a simple random sample of n = 20 households and record the size of the selected households below. Use a seed value of your choice or select a row and column of the random number table.

   Responses of the 20 selected households:

2. Compute the average response for your simple random sample of 20 households.

   Average = $^{\text{SUM}}\!/_{n}$ =

3. Collect the averages from each student or group in the class. Plot the averages (using the same scale as in Part I) and describe it.

Average for Simple Random Sample of households:

Part III. How did we do?

The true average household size for this community population is 3.82.

Which method of sampling yielded estimate which were generally larger?

Sampling the people or Sampling the households

Which method of sampling do you like best if you are interested in estimating the average size?

Sampling the people or Sampling the households

Why?

Estimate the amount of bias if you were to sample the people and not the households.
(i.e. what is difference between the true average and the approximate average if you take a sample of people?)

ID = Label   and # = Number of People in their Household

ID	#	ID	#	ID	#	ID	#	ID	#	ID	#	ID	#	ID	#
1	1	49	2	97	3	145	4	193	6	241	6	289	10	337	12
2	1	50	2	98	3	146	4	194	6	242	6	290	10	338	12
3	1	51	2	99	3	147	4	195	6	243	6	291	10	339	12
4	1	52	2	100	3	148	4	196	6	244	6	292	10	340	12
5	1	53	2	101	3	149	4	197	6	245	6	293	10	341	12
6	1	54	2	102	3	150	4	198	6	246	6	294	10	342	12
7	1	55	3	103	3	151	4	199	6	247	6	295	10	343	12
8	1	56	3	104	3	152	4	200	6	248	6	296	10	344	12
9	1	57	3	105	3	153	4	201	6	249	6	297	10	345	12
10	1	58	3	106	3	154	4	202	6	250	6	298	10	346	12
11	1	59	3	107	3	155	4	203	6	251	6	299	10	347	12
12	1	60	3	108	3	156	4	204	6	252	6	300	10	348	12
13	1	61	3	109	4	157	4	205	6	253	6	301	10	349	12
14	1	62	3	110	4	158	4	206	6	254	6	302	10	350	12
15	1	63	3	111	4	159	4	207	6	255	9	303	10	351	12
16	1	64	3	112	4	160	4	208	6	256	9	304	10	352	12
17	1	65	3	113	4	161	4	209	6	257	9	305	10	353	12
18	1	66	3	114	4	162	4	210	6	258	9	306	10	354	12
19	1	67	3	115	4	163	4	211	6	259	9	307	10	355	12
20	1	68	3	116	4	164	4	212	6	260	9	308	10	356	12
21	1	69	3	117	4	165	6	213	6	261	9	309	10	357	12
22	1	70	3	118	4	166	6	214	6	262	9	310	10	358	12
23	1	71	3	119	4	167	6	215	6	263	9	311	10	359	12
24	1	72	3	120	4	168	6	216	6	264	9	312	10	360	12
25	1	73	3	121	4	169	6	217	6	265	9	313	10	361	12
26	1	74	3	122	4	170	6	218	6	266	9	314	10	362	12
27	1	75	3	123	4	171	6	219	6	267	9	315	10	363	12
28	1	76	3	124	4	172	6	220	6	268	9	316	10	364	12
29	2	77	3	125	4	173	6	221	6	269	9	317	10	365	12
30	2	78	3	126	4	174	6	222	6	270	9	318	10	366	12
31	2	79	3	127	4	175	6	223	6	271	9	319	10	367	12
32	2	80	3	128	4	176	6	224	6	272	9	320	10	368	12
33	2	81	3	129	4	177	6	225	6	273	10	321	10	369	12
34	2	82	3	130	4	178	6	226	6	274	10	322	10	370	12
35	2	83	3	131	4	179	6	227	6	275	10	323	12	371	12
36	2	84	3	132	4	180	6	228	6	276	10	324	12	372	12
37	2	85	3	133	4	181	6	229	6	277	10	325	12	373	12
38	2	86	3	134	4	182	6	230	6	278	10	326	12	374	12
39	2	87	3	135	4	183	6	231	6	279	10	327	12	375	12
40	2	88	3	136	4	184	6	232	6	280	10	328	12	376	12
41	2	89	3	137	4	185	6	233	6	281	10	329	12	377	12
42	2	90	3	138	4	186	6	234	6	282	10	330	12	378	12
43	2	91	3	139	4	187	6	235	6	283	10	331	12	379	12
44	2	92	3	140	4	188	6	236	6	284	10	332	12	380	12
45	2	93	3	141	4	189	6	237	6	285	10	333	12	381	12
46	2	94	3	142	4	190	6	238	6	286	10	334	12	382	12
47	2	95	3	143	4	191	6	239	6	287	10	335	12		
48	2	96	3	144	4	192	6	240	6	288	10	336	12		