**Section 1.4 Other Effective Sampling Methods**

**Objectives**

* 1. Obtain a Stratified Sample
  2. Obtain a Systematic Sample
  3. Obtain a Cluster Sample

Introduction, Page 1

What is the goal of sampling? **The goal of sampling is to collect as much information as possible about the population at the least cost. Cost includes monetary outlays, time, and other resources. With this goal in mind, we may find it advantageous to use sampling techniques other than simple random sampling.**

***Objective 1: Obtain a Stratified Sample***

Objective 1, Page 1

Explain how to obtain a stratified sample. **STRATIFIED SAMPLE**

A **stratified sample** is obtained by dividing the population into nonoverlapping groups called *strata* and then obtaining a simple random sample from each stratum. The individuals within each stratum should be homogenous (similar) in some way.

An advantage of stratified sampling over simple random sampling is that it may allow fewer individuals to be surveyed while it obtains the same or more information. This result occurs because individuals within each subgroup have similar characteristics, so opinions within the group are not as likely to vary much from one individual to the next. In addition, a stratified sample guarantees that each stratum is represented in the sample

In this example, we'll go over the procedure

for obtaining a stratified sample.

The president of DePaul University

wants to conduct a survey to determine the community's

opinion regarding campus safety.

The president divides the DePaul community

into three groups-- resident students,

nonresident or commuting students, and staff-- including

faculty-- so that he can obtain a stratified sample.

Suppose there are 6,204 resident students, 13,304

nonresident students, and 2,401 staff

members for a total of 21,909 individuals in the population.

The resident students make up approximately 28%

of the population, the non-resident students make up

approximately 61% of the population,

and staff makes up approximately 11% of the population.

We want a stratified sample to represent this percentage

breakdown so that it looks just like the population, only

smaller.

The president wants to obtain a sample of size 100

with the number of individual selected

from each stratum weighted by the population size.

So we'll multiply the desired percentage

by the sample size of 100.

That would provide us with 28 resident students

in our sample, 61 non-resident students, and 11 staff members.

We simply multiply the percentage

by the sample size for each group.

To obtain the stratified sample, we'll

construct a simple random sample within each group.

We'll select 28 of the 6,204 resident students,

61 out of the 13,304 nonresident students,

and 11 out of the 2,401 staff members.

If you're going to do this with technology-- and we

will-- be careful not to use the same seed for all the groups

in the stratified sample because we

want the simple random samples within each stratum

to be independent of each other.

STEPS

DATA

SIMULATE

DISCRETE UNIFORM

**Again, I need 28 values, but I'm going to ask for 5 more**

**than that in case there are duplicates.**

**So 33 rows, 1 column, the minimum number**

**is 1, the maximum number 6,204, and I'm**

**going to use the dynamic seed.**

**Press compute.**

**So I'm going to use StatCrunch to select the random numbers.**

**I'm going to begin with the resident students.**

**We needed 28 students out of 6,204 students.**

**To select those numbers, press on Data, Simulate, Discrete**

**Uniform.**

**Again, I need 28 values, but I'm going to ask for 5 more**

**than that in case there are duplicates.**

**So 33 rows, 1 column, the minimum number**

**is 1, the maximum number 6,204, and I'm**

**going to use the dynamic seed.**

**Press compute.**

**Let's go ahead and generate the 61 non-resident students.**

**Data, Simulate, Discrete Uniform.**

**I needed 61.**

**I'm going to ask for 66, again, in case there are repeats.**

**The minimum value is 1.**

**The maximum value is 13,304.**

**I'll use another dynamic seed.**

**Press Compute.**

**Now to generate the 11 staff.**

**Data, Simulate, Discrete Uniform.**

**I need 11, so I'll ask for 16 in one column.**

**The minimum value is 1.**

**The maximum value, 2,401.**

**I'm going to save myself some grief and where it says Prefix,**

**I'm going to put the column name.**

**This one's going to be called Staff,**

**and I won't have to change it later.**

**Again, I'll use a dynamic seed, let the computer select it.**

**Press Compute, and that column,**

And I'm going to get rid of this header

and replace it with Resident.

Now I will only use the first 28 of those values that

do not repeat.

I'll check them for repeats in a moment.

Let's go ahead and generate the 61 non-resident students.

Data, Simulate, Discrete Uniform.

I needed 61.

I'm going to ask for 66, again, in case there are repeats.

The minimum value is 1.

The maximum value is 13,304.

I'll use another dynamic seed.

Press Compute.

And I'm going to change that header to Non-resident.

I'll check that for repeats in a bit as well.

We'll only use the first 61 that are non-repeating.

Now to generate the 11 staff.

Data, Simulate, Discrete Uniform.

I need 11, so I'll ask for 16 in one column.

The minimum value is 1.

The maximum value, 2,401.

I'm going to save myself some grief and where it says Prefix,

I'm going to put the column name.

This one's going to be called Staff,

and I won't have to change it later.

Again, I'll use a dynamic seed, let the computer select it.

Press Compute, and that column, Staff 1-- I'll just get rid

of the 1-- has been added.

And I'm only going to use the first 11 that do not repeat.

It looks like the first 11 there are clean.

OK, here are the 33 random numbers

we generated for resident students.

We needed 28, which means that we're supposed to end here.

And we've scoured the first 28 looking for any repeats,

and in this case, there are none,

so we can discard the 5 extra values that I requested.

And there are the 28 resident students

that will be included in the study.

Now your 28 and my 28 will be different

because they're selected at random.

Let's look at the non-resident students.

Selected 66 values.

I only need the first 61.

And in the first 61, I go looking for repeats,

and there actually are repeats.

12,731 was elected twice, so I can discard the second one,

and my first 61 values-- that was 61.

Add one more and discard the rest like this.

Now again, your 61 will be different than my 61

because we selected them randomly.

Finally, let's take a look at the staff.

Here are the 16 values I selected.

I only need 11 of them.

So as I look at those first 11 staff members,

there are no repeats, so I can get rid of the extra 5

I asked for.

And there are the 11 staff members

that will be included in the study.

So we used the first 28 non-repeating values

that we found for the resident, the first 61

non-repeating values for the non-residents,

and the first 11 non-repeating values for the staff.

You could have done this using Microsoft Excel,

some other software package.

You also could have done this with a TI calculator.

Just keep going until you get your 28 unique residents, 61

unique non-residents, and so on.

An advantage of stratified sampling

over simple random sampling is that the researcher

can determine characteristics within each strata.

This allows an analysis to be performed on each stratum

to see if any significant differences among them exist.

For example, we could analyze the data obtained in example 1

to see if there's a difference in the opinions of students

versus staff.

Objective 1, Page 2

**Example 1** ***Obtaining a Stratified Sample***

The president of DePaul University wants to conduct a survey to determine the community’s opinion regarding campus safety. The president divides the DePaul community into three groups: resident students, nonresident (commuting) students, and staff (including faculty) so that he can obtain a stratified sample.

Suppose there are 6,204 resident students, 13,304 nonresident students, and 2,401 staff, for a total of 21,909 individuals in the population. What percent of the DePaul community is made up of each group?

The president wants to obtain a sample of size 100, with the number of individuals selected from each stratum weighted by the population size. How many individuals should be selected from each stratum?

To obtain the stratified sample, construct a simple random sample within each group.

***Objective 2: Obtain a Systematic Sample***

Objective 2, Page 1

Explain how to obtain a systematic sample. **SYSTEMATIC SAMPLE**

In both simple random sampling and stratified sampling, a [frame](https://xlitemprod.pearsoncmg.com/assignment/containerassignmentplayer.aspx#xln-lb-lnk_obj2_1_c7c47617-8634-539a-37e1-6d2a73e7cae4) must exist. Therefore, these sampling techniques require some preliminary work before the sample can be found. A sampling technique that does not require a frame is *systematic sampling*.

A **systematic sample** is obtained by selecting every kth individual from the population. The first individual selected corresponds to a number between 1 and k.

Because systematic sampling does not require a frame, it is a useful technique when you cannot gather a list of the individuals in the population. Also, systematic samples typically provide more information for a given cost than does simple random sampling. In addition, systematic sampling is easier to employ; so there is less likelihood of interviewer error occurring, such as selecting the wrong individual to be surveyed.

**Note:** Because systematic sampling does not require a frame, it is a useful technique when you cannot gather a list of the individuals in the population.

Objective 2, Page 2

**Example 2 *Obtaining a Systematic Sample without a Frame***

The manager of Kroger Food Stores wants to measure the satisfaction of the store’s customers. Design a sampling technique that can be used to obtain a sample of 40 customers.

STATCRUNCH

APPLET

RANDOM NUMBERS

7th client

1

7

Compute

3,7,14…….1 less than sample of 40 3 + 39(7) times every 7th

Objective 2, Page 4

*Answer the following after watching the video.*

1. that can result from choosing a value of *k* that is too small

If the size of the population is unknown,

there's no mathematical way to determine k

for systematic sampling.

The value of k must be small enough

to achieve our desired sample size.

But it also must be large enough to obtain a sample that's

representative of the population.

For example 2, suppose that k was equal to 30.

We said that we would start with customer 3,

and then add 30 to that, and 30 to that,

until we reach our 40th customer, which we could find

by multiplying 39 times 30 added to the original 3,

that's customer number 1,173.

If Kroger does not have 1,173 shoppers,

the desire sample size will not be achieved.

Now suppose that k equals 4, starting again at 3,

the second person selected would be customer 7, and then

customer 11, until we reach our 40th customer selected,

that's 39 times 4 plus 3, that's customer number 159.

The 159 shopper might leave the store at 3:00 in the afternoon,

so our survey would not include any of the evening shoppers.

An estimate of the size of the population

would help to determine an appropriate value of k.

#### W **Choosing a Value for**k

When using systematic sampling, how would we select the value of k?

If the size of the population is unknown, there is no mathematical way to determine k. The value of k must be small enough to achieve our desired sample size and large enough to obtain a sample that is representative of the population. The following video illustrates the importance of k

What can result from choosing a value of *k* that is too large? The value of k must be small enough to achieve our desired sample size

Objective 2, Page 5 

 *Answer the following after watching the second video after Example 2.*

Explain how to determine the value of *k* if the population size *N* is known.

Now we'll take a look at how to determine k

when the population size n is known.

So if possible, approximate the population size n.

Then determine the sample size desired-- lowercase n.

Divide the population size by the sample size

and round down to the nearest integer.

This value is k.

So let's suppose that the population size was 20,325.

And we desire a sample size of n equals 100.

If we divide 20,325 by 100, that's equal to 203.25.

And rounding this down-- that gives us a value of k of 203.

Now let's further suppose that we

start with the 90th individual.

So our sample would include the 90th individual.

Then I add 203.

That's 293.

I add 203 again.

That's 496.

And we keep going until we reach our desired sample size.

The last value will be 1 less than the sample

size-- 99-- times the value of k-- 203--

added to the starting value of 90.

And that equals 20,187.

So the 20,187th individual will be the last individual

in the survey.

Objective 2, Page 7

1. List the five steps in obtaining a systematic sample.

#### **Steps in Systematic Sampling**

1. If possible, approximate the population size, N.
2. Determine the sample size, n.
3. Compute Nn and round down to the nearest integer. This value is k.
4. Randomly select a number between 1 and k. Call this number p.
5. The sample will consist of the following individuals:
6. p,p+k,p+2k,...,p+(n−1)k

***Objective 3: Obtain a Cluster Sample***

Objective 3, Page 1

A fourth sampling method is called *cluster sampling*. Like the previous three sampling methods, this method has benefits under certain circumstances.

**DEFINITION**

A **cluster sample** is obtained by selecting all individuals within a randomly selected collection or group of individuals.

Suppose a school administrator wants to learn the characteristics of students enrolled in online classes. Rather than obtaining a simple random sample based on the frame of all students enrolled in online classes, the administrator treats each online class as a cluster and then finds a simple random sample of these clusters. The administrator then surveys *all* students in the selected clusters. This reduces the number of classes that get surveyed

Objective 3, Page 2

In this example, we'll explore the process

of obtaining a cluster sample.

A sociologist wants to gather data regarding

household income within the city of Boston.

Obtain a sample using cluster sampling.

The city of Boston can be set up so

that each city block is a cluster.

Recall that a cluster is a geographic location.

Once the city blocks have been identified,

we obtain a simple random sample of the city blocks,

and then we survey all households

on the block selected.

Let's suppose there are 10,493 city blocks in Boston.

First, the sociologist has to number these blocks from 1 up

through 10,493.

Suppose the sociologist only has enough time and money

to survey 20 clusters.

We'll randomly select 20 numbers between 1 and 10,493

to tell us which clusters, or city blocks,

will be included in the survey.

We have many options to select 20 blocks from the 10,493

blocks.

**I'm going to use StatCrunch.**

**Press Data, Simulate, Discrete Uniform.**

**I only need 20 values, but I'm going**

**to ask for 25 in case there are repeats.**

**I'm going to put them all in one column.**

**The minimum value is 1.**

**The maximum value is 10,493.**

**I'm going to use a dynamic seed.**

**Let the computer pick it, and press compute.**

**And my values are here in the first column.**

**I'm going to check for repeats, and if there are none,**

**I'll use the first 20 values.**

**So here are the 25 random numbers that we selected.**

**I only need 20, which will end above the line.**

**I want to take a look at those first 20**

**and see if there are any repeats.**

**And since there are not, I can discard the extra 5**

**random numbers I generated.**

**So these are the 20 blocks that will be included in the study.**

There are a couple of advantages of cluster

sampling we should discuss.

First, cluster sampling reduces the travel time

that would likely be required with stratified

sampling or simple random sampling in that many

of the households will be clustered together

rather than driving across the city.

Second, there's no need to obtain

a frame of all the households.

All we really need is one frame that

provides information regarding city blocks.

**Example 3 *Obtaining a Cluster Sample***

A sociologist wants to gather data regarding household income within the city of Boston. Obtain a sample using cluster sampling.

Objective 3, Page 3

1. If the clusters have homogeneous individuals, is it better to have more clusters with fewer individuals in each cluster or fewer clusters with more individuals in each cluster?
2. If the clusters have heterogeneous individuals, is it better to have more clusters with fewer individuals in each cluster or fewer clusters with more individuals in each cluster?

**Issues to Consider in Cluster Sampling**

The following questions arise in cluster sampling:

* How do I cluster the population?
* How many clusters do I sample?
* How many individuals should be in each cluster?

First, we must determine whether the individuals within the proposed cluster are *homogeneous* (similar individuals) or *heterogeneous* (dissimilar individuals).

In Example 3, city blocks tend to have similar households. Survey responses from houses on the same city block are likely to be similar. This results in duplicate information. We conclude that **if the clusters have homogeneous individuals, it is better to have more clusters with fewer individuals in each cluster.**

1. If the clusters have homogeneous individuals, is it better to have more clusters with fewer individuals in each cluster or fewer clusters with more individuals in each cluster?

**If the clusters have homogeneous individuals, it is better to have more clusters with fewer individuals in each cluster.**

1. If the clusters have heterogeneous individuals, is it better to have more clusters with fewer individuals in each cluster or fewer clusters with more individuals in each cluster?

What if the cluster is heterogeneous? Under this circumstance, the heterogeneity of the cluster likely resembles the heterogeneity of the population. In other words, each cluster is a scaled-down representation of the overall population.

For example, a quality control manager might use shipping boxes that contain 100 lightbulbs as a cluster. The manager does this because the rate of defects within the cluster resembles the rate of defects in the population, assuming that the bulbs are randomly placed in the box. Thus,

**When each cluster is heterogeneous, fewer clusters with more individuals in each cluster are appropriate.**

Objective 3, Page 5

1. Define: Convenience sampling

Objective 3, Page 6

**Note:** The most popular convenience samples are those in which the individuals in the sample are self-selected**,** meaning the individuals themselves decide to participate in the survey. Self-selected surveys are also called voluntary response samples.

Objective 3, Page 7

1. Define: Multistage sampling
2. List the two stages Nielsen Media Research uses to investigate TV viewing habits.

Objective 3, Page 8

1. How many stages does the Census Bureau use for the Current Population Survey? What are those stages?

Objective 3, Page 9

Researchers need to know how many individuals they must survey to draw conclusions about the population within some predetermined margin of error. They must find a balance between the reliability of the results and the cost of obtaining these results. The bottom line is that time and money determine the level of confidence researchers will place on the conclusions drawn from the sample data. The more time and money researchers have available, the more accurate the results of the statistical inference.

Objective 3, Page 10

 *Watch the animation for a summary of simple random sampling, systematic sampling, stratified sampling, and cluster sampling.*

**Convenience Sampling**

In the four sampling techniques just presented (simple random sampling, stratified sampling, systematic sampling, and cluster sampling), the individuals are selected randomly. Often, however, inappropriate sampling methods are used in which the individuals are *not* randomly selected.

Have you ever been stopped in the mall by someone holding a clipboard? These folks are responsible for gathering information, but their methods of data collection are inappropriate, and the results of their analysis are suspect because they collect data using a *convenience sample*.

**DEFINITION**

In a **convenience sample,** the individuals are easily obtained and not based on randomness.

**Examples of Convenience Samples**

The most popular convenience samples are those in which the individuals in the sample are **self-selected**, meaning the individuals themselves decide to participate in the survey. Self-selected surveys are also called **voluntary response** samples. One example of self-selected sampling is phone-in polling—a radio personality will ask his or her listeners to phone or text the station to submit their opinions. Another example is the use of the Internet to conduct surveys. For example, a TV news show will present a story regarding a certain topic and ask its viewers to "tell us what you think" by completing an online questionnaire or tweeting an opinion with a hashtag.

Both of these samples are poor designs because the individuals who decide to be in the sample generally have strong opinions about the topic. A more typical individual in the population will not bother phoning, texting, or tweeting to complete a survey. Any inference made regarding the population from this type of sample should be made with extreme caution.

**Multistage Sampling**

**In practice, most large-scale surveys obtain samples using a combination of the techniques just presented.**

As an example of multistage sampling, consider Nielsen Media Research. Nielsen randomly selects households and, through a People Meter, monitors the television programs these households are watching. The meter is an electronic box connected to each TV within the household. The People Meter measures what program is being watched and who is watching it. Nielsen selects the households with the use of a two-stage sampling process.

**Stage 1:** Using U.S. Census data, Nielsen divides the country into geographic areas (strata). The strata are typically city blocks in urban areas and geographic regions in rural areas. About 6000 strata are randomly selected.

**Stage 2:** Nielsen sends representatives to the selected strata and lists households within the strata. The households are then randomly selected through a simple random sample.

Nielsen sells the information obtained to television stations and companies. These results are used to help determine prices for commercials.

**An Example of Multistage Sampling**

Consider the sample used by the Census Bureau for the Current Population Survey. This survey requires five stages of sampling:

**Stage 1:** Stratified sample  
  
**Stage 2:** Cluster sample  
  
**Stage 3:**Stratified sample  
  
**Stage 4:** Cluster sample  
  
**Stage 5:** Systematic sample

This survey is very important because it is used to obtain demographic estimates of the United States in noncensus years. Details about the Census Bureau's sampling methods can be found in *The Current Population Survey: Design and Methodology*, Technical Paper No. 40

**Sample Size Considerations**

Throughout our discussion of sampling, we did not mention how to determine the sample size. Researchers need to know how many individuals they must survey to draw conclusions about the population within some predetermined margin of error.

Researchers must find a balance between the reliability of the results and the cost of obtaining these results. Time and money determine the level of confidence researchers will place on the conclusions drawn from the sample data. The more time and money researchers have available, the more accurate the results of the statistical inference will be.

Later in the course, we will discuss techniques for determining the sample size required to estimate characteristics regarding the population within some margin of error. (For a detailed discussion of sample size considerations, consult a text on sampling techniques such as *Elements of Sampling Theory and Methods* by Z. Govindarajulu, Pearson, 1999.)

VIDEO -SUMMARY OF SAMPLING TECHNIQUES

STUDENT: Can you help me distinguish the four

different sampling methods?

**Simple random sampling** is like drawing names from a hat.

We use a frame to assign numbers to each individual

in the population.

Then random numbers are generated

to identify the individuals in the sample.

For example, suppose we have a population

of all introductory statistics students with 100 individuals.

We would like to know the average GPA

of an introductory statistics student,

so decide to obtain a simple random sample of 5 students.

We number the individuals from 1 to 100

and randomly generate 5 unique numbers to obtain the sample.

So in this particular simple random sample,

we obtained student 9, 74, 90, 61, and 63,

and obtained an average GPA of 3.312.

**systematic sampling**

INSTRUCTOR: With systematic sampling,

we sample every kth individual.

Again, consider the 100 introductory statistics

students.

Suppose we want a sample of size 5 from this population.

Because the population size is 100, we divide 100 by 5

and obtain 20.

Therefore, randomly select a number between 1 and 20

and survey every 20th individual in the population.

The key with systematic sampling is

that the individuals in the population

are arranged in some order.

Even if you do not have a frame, a systematic sample

could be obtained, such as selecting every 10th can

from an assembly line.

So in this particular sample, we randomly selected the number 7

between 1 and 20, and so we survey

the 7th, 27th, 47th, 67th, and 87th student,

and obtain a mean GPA of 3.476.

**stratified sampling**,

we divide the population into non-overlapping homogeneous

groups.

That is, each group must have some common characteristic.

For example, suppose we divide the class

into on-campus students in red and off-campus students

in blue.

Then within each of these strata,

we might obtain a simple random sample of 5 students.

For example, for the on-campus students,

we selected student 25, 91, 61, 65, and 63.

For the off-campus students, we selected students 64, 34, 49,

6, and 30.

The mean GPA of the on-campus student is 2.941.

The mean GPA of the off-campus student is 2.719.

**cluster sampling**

INSTRUCTOR: Cluster sampling is obtained

by selecting all the individuals within a collection

of individuals.

Again, consider our classroom of 100 statistics students.

We could treat each row students as a cluster.

Then we might randomly select two rows.

Each student in the selected rows is then surveyed.

So here we randomly selected row five, and each student in row

five is surveyed for their GPA.

We also randomly selected row four.

And again, each student in row four

this surveyed to determine their GPA.

We find that the mean GPA of the 20 students surveyed is 3.019.

**How does stratified sampling**

**differ from cluster sampling?**

:

INSTRUCTOR: With stratified sampling,

we divide the population into non-overlapping groups,

such as freshmen, sophomores, juniors, seniors.

Then within each strata, we obtain a simple random sample.

This is great if we wish to compare

the value of the variable, such as GPA,

between the various strata.

Do freshman have a different GPA from sophomores?

With cluster sampling, the population is also divided,

but we randomly select clusters and then survey all individuals

within the cluster.

For example, suppose we wanted to find out

why students enroll in 8:00 AM courses on Monday morning.

You might divide the population of all 8:00 AM Monday courses

into clusters.

Then we obtain a simple random sample of four courses,

and survey all students in the four selected