# RESEARCH METHODS 2RM - Sampling

Author:   Ed Nelson
Department of Sociology M/S SS97
California State University, Fresno
Fresno, CA 93740
Email:  ednelson@csufresno.edu

**Note to the Instructor:** This is the second in a series of 13 exercises that were written for an introductory research methods class.  The first exercise focuses on the research design which is your plan of action that explains how you will try to answer your research questions.  Exercises two through four focus on sampling, measurement, and data collection.  The fifth exercise discusses hypotheses and hypothesis testing.  The last eight exercises focus on data analysis.  In these exercises we’re going to analyze data from one of the [Monitoring the Future Surveys](http://monitoringthefuture.org/) (i.e., the 2017 survey of high school seniors in the United States).  This data set is part of the collection at the Inter-university Consortium for Political and Social Research at the University of Michigan.  The data are freely available to the public and you do not have to be a member of the Consortium to use the data.  We’re going to use SDA (Survey Documentation and Analysis) to analyze the data which is an online statistical package written by the Survey Methods Program at UC Berkeley and is available without cost wherever one has an internet connection.  A weight variable is automatically applied to the data set so it better represents the population from which the sample was selected.  You have permission to use this exercise and to revise it to fit your needs.  Please send a copy of any revision to the author so I can see how people are using the exercises. Included with this exercise (as separate files) are more detailed notes to the instructors and the exercise itself.  Please contact the author for additional information.

I’m attaching the following files.

* [This page](http://ssric.org/files/Research_Methods_2RM.docx) in MS Word (.docx) format.
* [An Excel spreadsheet](http://ssric.org/files/exercise_spreadsheet_for_exercise_2RM.xlsx) for use in parts of this exercise.

**Goal of Exercise**

The goal of this exercise is to provide an introduction to sampling which is an integral part of any research design.  The other elements of your research design are measurement, data collection, and data analysis and will be discussed in future exercises.

**Part I—Populations and Samples**

Populations are the complete set of individuals that we want to study.[[1]](%22%20%5Cl%20%22_ftn1%22%20%5Co%20%22)  For example, a population might be all the individuals that live in the United States at a particular point in time.  The U.S. does a complete enumeration of all individuals living in the United States every ten years (i.e., each year ending in a zero).  We call this a census.

Another example of a population is all high school students in the United States.  The research study that we’ll be using in these exercises is the Monitoring the Future Survey of high school seniors in the United States that has been conducted yearly since 1975.  There is a [website](http://monitoringthefuture.org/) that will give you a lot of information about this study.  Here’s a brief description from the website’s home page.

“Monitoring the Future is an ongoing study of the behaviors, attitudes, and values of American secondary school students, college students, and young adults. Each year, a total of approximately 50,000 8th, 10th and 12th grade students are surveyed (12th graders since 1975, and 8th and 10th graders since 1991). In addition, annual follow-up questionnaires are mailed to a sample of each graduating class for a number of years after their initial participation.”

A major focus of these surveys is students’ drug use.  But the surveys include a lot more information than just drug use.  The website describes the range of questions asked.

“Questions include drug use and views about drugs, delinquency and victimization, changing roles for women, confidence in social institutions, concerns about energy and ecology, and social and ethical attitudes.”

These are only a few of the areas that students are asked about.  Other areas include, for example, their educational goals, religion, politics, the military, race, health, and background information.

Populations are often large and it’s too costly and time consuming to carry out a complete enumeration.  So, what we do is to select a sample from the population where a sample is a subset of the population.  That’s what the Monitoring the Future Survey did.  It selected a sample of all 12th graders in the United States.  Students in this sample were given a questionnaire to fill out and that became the data for the study.

A statistic describes a characteristic of a sample while a parameter describes a characteristic of a population.  The percent of all high school students (i.e., our population) that drink alcoholic beverages is a parameter.  However, the percent of high school students in the sample that drink is an example of a statistic. We use statistics to make inferences about parameters.  In other words, we use the percent of students in the sample who drink to make an inference about the percent who drink in the population.  Notice that the percent of the sample (our statistic) is known while the percent of the population (our parameter) is usually unknown.

**Part II – Probability and Non-Probability Sampling**

There are many different ways to select samples.  Probability samples are samples in which every object in the population has a known, non-zero, chance of being in the sample (i.e., the probability of selection).  This isn’t the case for non-probability samples.  An example of a non-probability sample is an instant poll which you hear about on radio and television shows.  A show might invite you to go to a website and answer a question such as whether you favor or oppose same-sex marriage.  This is a purely volunteer sample and we have no idea of the probability of selection.

In this exercise we’re going to focus on probability sampling.  We’re going to discuss three different types of probability samples – simple random samples, stratified random samples, and cluster samples.

**Part III – Simple Random Samples**

There are many ways of selecting a probability sample but the most basic type of probability sample is a simple random sample in which everyone in the population has the same chance of being selected in the sample.  If you have a list of all the individuals in your population, it’s easy to select a simple random sample.  There is a [data base](http://ssric.org/files/exercise_spreadsheet_for_exercise_2RM.xlsx) provided with this exercise.  In this hypothetical population there are 100 individuals numbered 1 to 100 (i.e., ID).  Individuals in the population are also listed by sex (M for male and F for female)and whether they favor (F for favor) or oppose (O for oppose) same-sex marriage.

To select a simple random sample, all you need to do is to follow these easy steps.

* Number all the individuals from 1 to n where n is the total number of individuals in the population.  If your population consists of 100 individuals, then number them from 1 to 100.  This is done for you in the data file.
* Select m random numbers where m is the number of individuals in your sample.  A set of random numbers has no discernable pattern to it.  There are many random number generators on the internet.  One of those generators can be found at the [Stat Trek website](http://stattrek.com/statistics/random-number-generator.aspx).   All you have to do is to enter the minimum value (i.e., 1 for the example above), the maximum number (i.e., 100), and the number of random numbers you want (e.g., 10 if you want a sample of 10 individuals).  Note that it also asks if you want to allow duplicate entries.  Most of the time you do not, so select “False.”  Ignore the “Seed” box.  Click on “Calculate” to generate the random numbers.

Write down the 10 random numbers that the generator produced and label this sample 1.  Now calculate the percent of respondents in this sample that favored and opposed same-sex marriage.

Repeat this process.  All you have to do is to click on “Calculate” again.  Write down the 10 random numbers and label this sample 2 and calculate the percent of respondents in this sample that favored and opposed same-sex marriage.  Notice that the two samples will consist of different individuals although there may be some overlap.

Now repeat this process again and label this sample 3 and again calculate the percent of respondents in this sample that favored and opposed same-sex marriage.

Were the percent of respondents in the three samples that favored and opposed same-sex marriage all the same or different?  What does this tell you about sampling?

**Part IV – Stratified Random Samples**

We know that no sample is ever a perfect representation of the population from which the sample is drawn.  This is because every sample contains some amount of sampling error.  Sampling error is inevitable.  There is always some amount of sampling error present in every sample.

Since we can’t eliminate sampling error, what we do is try to minimize sampling error.  One way to do that is to stratify the sample.  Notice that in the exercise data base 50% of the population is male and 50% is female.  When we select a simple random sample of 10 individuals from this population, sometime the sample has 50% male and 50% female and sometimes there are more males than females and other times there are more females than males.  Go back and check the three samples that you selected in Part 3 and calculate how many males and females there were in each sample.  Were there the same number of males and females or were there more males or more females?  You probably didn’t get exactly 50% males and 50% females in all three samples.  Although it is possible, it’s not likely.

We can stratify our sample by sex and ensure that the sample has the same percent males and females as does the population.  How would we do that?  Divide the sample into two groups – all males and all females.  Since the population is 50% males and 50% females, we want our sample to be 50% males and 50% females.  For a sample of 10 individuals, that means we want our sample to have 5 males and 5 females.  That’s easy to do in our exercise data base since the 50 males are listed first (id’s 1 to 50) and the 50 females are listed next (id’s 51 to 100).

Use the same [random-number generator](http://stattrek.com/statistics/random-number-generator.aspx) that we used in Part 3.   For the males, all you have to do is to enter the minimum value (i.e., 1), the maximum number (i.e., 50), and the number of random numbers you want (e.g., 5).  For females, just change the minimum value to 51 and the maximum value to 100 and leave the number of random numbers at 5.

Select three stratified random samples and write down the random numbers for each of the three samples.  Each sample should have 5 males and 5 females for a total sample of 10.  Calculate how many males and females there were in each sample and write that after the random numbers for each sample.  This time there should be exactly 5 males and 5 females in each sample.  These are stratified random samples.  Since we have made sure that the population and the samples have the same proportion males and females, they are often called proportional stratified random samples.[[2]](%22%20%5Cl%20%22_ftn2%22%20%5Co%20%22)

Stratification will decrease sampling error if the variable that is used to stratify the sample is related to what you want to estimate.  In this case, we want to estimate the proportion of the population that favor and oppose same-sex marriage (i.e., the parameter).  To do that we select a sample from the population and use the percent of the sample that favors and opposes same-sex marriage as an estimate of the population parameter.[[3]](%22%20%5Cl%20%22_ftn3%22%20%5Co%20%22)  Since sex is related to how people feel about same-sex marriage[[4]](%22%20%5Cl%20%22_ftn4%22%20%5Co%20%22), sampling error will be reduced.  In order to stratify a sample, the stratifying variable must be known for each case in the population as it is in this exercise.

**Part V – Cluster Samples**

Notice that simple random samples and stratified random samples assume that we have a list of the population from which to select our sample. But what if we don’t have such a list?  For example, how would we get a sample of high school seniors?  There is no list available.  But there is a list of all high schools in the United States.  So, we could select a sample of high schools and then within each high school in our sample select a sample of seniors.  This is called a cluster sample because high schools are the clusters where you find seniors.

This is similar to how the Monitoring the Future Survey selected its sample of high school seniors in the United States although their sampling design is more complex.  Information about this study is archived at the Inter-university Consortium for Political and Social Research (ICPSR) located at the University of Michigan.  Start by going to their [website](http://www.icpsr.umich.edu/icpsrweb/).  In the upper-right corner of the home page click on “Log In/Create Account.”  Scroll down and click on “Create Account” below “New User.”  Fill in the requested information and click on “Submit.”  It will create your account and give you access to the ICPSR archive.  You can use your account from anywhere you have internet access.  If you don’t use your account for six months, your account will go away.

If you are a student, faculty member or staff at a university or college that belongs to the ICPSR, you will have access to all the archive’s data holdings.  If you are not, then you will only have access to public-use data.  Fortunately, the Monitoring the Future Surveys were funded for public access, so you have access to this study regardless of your status.

Once you have created your account, click on “Find Data” in the menu bar at the top of the screen.  Then type "ICPSR 37182” in the “Find Data” box.  That's the ICPSR identification number for the 2017 Monitoring the Future survey of 12th graders. That's the survey we'll be using in these exercises.

Click on the link for this study.  Read through the study description to get an overview of the research design.  Click on the links for "Scope of Project" and "Methodology" to get more information about this survey.

Under "Methodology" you will see a brief description of the sampling design. Read through it. Don't get lost in the details.  You'll see that the sampling design is a multistage cluster sample consisting of three stages.  Primary sampling units refers to geographic areas.  The clusters are the high schools because that’s where you find high school seniors.  Write three paragraphs describing each of the three stages.  Don’t just cut and paste or repeat word for word what is on the website..  Rather summarize in your own words how the sample was selected.  Show that you understand what a multistage cluster sample is.

**Part VI – Sampling Error**

As we said earlier, no sample is ever a perfect representation of the population from which the sample is drawn.  That’s because every sample contains some amount of sampling error.  Sampling error is inevitable.  The question then is how can we reduce sampling error?

As we discussed in Part 4, stratifying a sample is one way that you can reduce sampling error.  This assumes that the variable you are using to stratify the sample is related to whatever you are studying.  For example, if you are trying to explain why some people favor same-sex marriage and others oppose it, then you could stratify your sample by sex.  Assuming that sex is related to how people feel about same-sex marriage (and it is), this will reduce sampling error.

Another way is to increase the sample size.  The larger the sample size, the less the sampling error.  A simple random sample of 400 will have half the sampling error that a simple random sample of 100 has.  To reduce the amount of sampling error by half for a simple random sample, you have to quadruple the sample size.

We also need to think about the effect of sampling design on sampling error. Let’s compare three samples of size 1,000 from the same population.  Sample A is a simple random sample of size 1,000.  Sample B is a stratified random sample of size 1,000 where we have stratified on a variable that is related to whatever we are trying to estimate.  Sample C is a multistage cluster sample of size 1,000.  Stratification will reduce sampling error while cluster sampling will increase sampling error. So, sample C will probably have the most sampling error and sample B will probably have the least.  Sample A will probably be between the other two samples.

Cluster sampling is important because sometimes it is the only way we can get a sample when there is no list of the population available.  There’s no list of all high school seniors in the United States but there are lists of all high schools in the United States.  So, we use a multistage cluster sample.  One way we can control sampling error when using cluster sampling is to select as many clusters as we can afford.  So, the Monitoring the Future Survey includes a large number of high schools and a large number of seniors per high school.  That makes for a very expensive study but it’s necessary to control sampling error

[[1]](%22%20%5Cl%20%22_ftnref1%22%20%5Co%20%22) We need not limit ourselves to studying individuals.  We could also study objects like businesses or nations.  So it might be better to define a population as the complete set of objects that we want to study.  But in these exercises our focus is on individuals so we’ll define a population as the complete set of individuals we want to study.

[[2]](%22%20%5Cl%20%22_ftnref2%22%20%5Co%20%22) We’re not going to discuss disproportional stratified random samples in this exercise.  That would be a sample that is selected such that some segments are oversampled and other segments are undersampled.  For example, we might undersample whites and oversample non-whites so that our sample is 50% whites and 50% non-whites.  This would be useful if we wanted to compare whites and non-whites and wanted to have a larger sample of non-whites for comparison purposes.

[[3]](%22%20%5Cl%20%22_ftnref3%22%20%5Co%20%22) In this example, we can calculate the population percent since we know the values for each person in the population.  However, we real life situations, we won’t know the population parameter.

[[4]](%22%20%5Cl%20%22_ftnref4%22%20%5Co%20%22) You can determine from the exercise data base that 60% of females favor same-sex marriage compared to only 40% of males.