# Research Methods Exercises Using the 2018

# Monitoring the Future Survey of High School Seniors

# and SDA (Survey Documentation and Analysis)

# Edward Nelson

# California State University, Fresno

# Table of Contents

Page

Preface 3

Exercise 1 – Research Design 4

Exercise 2 – Sampling 11

Exercise 3 – Measurement 17

Exercise 4 – Data Collection 21

Exercise 5 – Hypotheses and Hypothesis Testing 27

Exercise 6 – Introduction to Data Analysis 31

Exercise 7 – Central Tendency and Dispersion 37

Exercise 8 – Graphs and Charts 45

Exercise 9 – Crosstabulation 48

Exercise 10 – Chi Square 53

Exercise 11 – Measures of Association 59

Exercise 12 – Spuriousness 65

Exercise 13 – Writing Research Reports 70

Appendix A – Introduction to SDA 74

Appendix B – Notes to Instructors 79

# Preface

# These exercises were written for an introductory research methods course although they could be used in any class that has a research component. They could also be used by individuals who want practice with the different components of a research design – sampling, measurement, data collection, and data analysis. The exercises on data analysis use percentages, Chi Square, and measures of association. They do not discuss all aspects of these statistics nor do they describe how to compute most of these statistics.

These exercises use SDA (Survey Documentation and Analysis) which is an online statistical package written by the Survey Methods Program at UC Berkeley. SDA can be used without cost wherever one has an internet connection. Students can be shown how to use SDA in approximately ten minutes making it unnecessary to spend valuable class time learning how to use a statistical package. There is also an extensive help menu available to users of SDA.

The data used in this series of exercises is the 2018 Monitoring the Future (MTF) survey of high school seniors. The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been an annual survey ever since. More information about the MTF survey is available on their [website](http://monitoringthefuture.org/).

The 2018 survey of high school seniors is freely available through the Inter-university Consortium for Political and Social Research (ICPSR) located at the University of Michigan. All you have to do is to go to their [website](http://www.icpsr.umich.edu/icpsrweb/) and follow the instructions in Appendix A at the end of these exercises. 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 MTF Surveys were funded for public access so you have access to this study.

In these exercises variable names appear in italics and SDA commands are in all capitals to make them easily recognizable. You can modify this if you wish.

The exercises were written so that each exercise is independent of the other exercises. That means that there is some redundancy across the exercises. If you choose to use several exercises you may want to remove some of the redundant material.

You have permission to edit the exercises in whatever way you desire. You can freely delete and add materials of your own. Please cite the original source of the exercises. I would like to hear from you about your experiences using them. If you find any errors, please let me know and I’ll correct them.

## About the Author

Ed Nelson is Professor Emeritus of Sociology at California State University, Fresno. Before retiring he taught courses in research methods, statistics, and critical thinking. After retiring he continues to teach a course in critical thinking. He can be reached by email at [ednelson@csufresno.edu](mailto:ednelson@csufresno.edu). Please contact him with any questions you might have.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 1 -- Research Design

## Goals of Exercise

The goals of this exercise are to introduce the idea of a research design and to explore the basic elements of any research design – sampling, measurement, data collection, and data analysis.

## Part I—Research Questions and Research Design

All research starts with one or more research questions.  These are the questions that you want to answer in your research study.  For example, you might want to find out why some people vote Democrat and others vote Republican.  Or you might want to find out why some people don’t vote at all.  Another question you might want to try to answer is why some favor same-sex marriage and others oppose it.

There are lots of ways that we might go about trying to answer these questions.  Some might rely on what their friends or family tell them.  Others might rely on what people in authority like their religious leaders tell them.  Still others might use what is often called common sense to answer these questions.  But we’re going to use the scientific approach to try to answer these questions.  Thomas Sullivan defined science as a “method of obtaining knowledge about the world through systematic observations.”[[1]](#footnote-1)  Notice that science is empirical; it’s based on observations.  Also, notice that we’re talking about a particular type of observations – systematic observations.

A research design is your plan of action.  It lays out how you plan to go about answering your questions.  The research design includes how you plan to select the cases for analysis (sampling), how you will measure concepts, how you plan to collect your data, and how you will analyze the data.  Exercises two through five focus on the components of a research design and exercises six through thirteen deal with data analysis.

First, we have to learn how to formulate good research questions.  Let’s start by looking at some examples of poor questions.  Why are these poor questions?

* Women are more likely than men to vote Democrat in presidential elections.  This one is easy.  It’s not a question.  It’s actually a hypothesis which we will discuss in Exercise 5.
* Why are women more likely than men to vote Democrat in presidential elections?  This one is a little more difficult.  We want to start with the more general question such as why some people vote Democrat and others vote Republican?  Then we would consider possible answers to this question.  One of these answers might be that gender influences voting.  Since science is empirical, we would start by looking at data to see if, in fact, gender does influence voting and we would discover that in most recent presidential elections women are more likely to vote Democrat.  This would lead us to ask why women are more likely than men to vote Democrat.  But we would start our study with the more general question.
* Why do dogs bark?  This is certainly a question and perhaps an interesting question.  But it’s not a question that social scientists would be interested in.  Social scientists focus on questions that involve behavior, attitudes, and opinions.[[2]](#footnote-2)

What are the characteristics of a good research question?

* We start by looking at general questions such as what influences voting or why do some people favor same-sex marriage and others oppose it.  As our study progresses, we move to more focused questions such as why women are more likely to vote Democrat than men.
* We focus on questions that ask about behavior, attitudes, and opinions.
* Good questions are clearly stated.  Questions such as what about voting aren’t clear and therefore aren’t useful.
* As with everything we write, we want to make sure that we use correct spelling and good grammar.  So proofread everything you write including your questions.

## Part II – Now It’s Your Turn

Write three research questions that could guide the beginning of a research study.  They can deal with any subject matter that asks about the behavior, attitudes, and opinions of people.  Be sure to follow the guidelines for writing good questions discussed in part 1.

## Part III – Sampling

Populations are the complete set of individuals that we want to study.[[3]](#footnote-3)  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 the students in a particular school or all college students in your state.  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 and then use the sample data to make an inference about the population.

There are many different ways to select samples.  Probability samples are samples in which every individual 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.

There are a number of different ways of selecting a probability sample.

* The most basic type of probability sample is the simple random sample where every individual in the population has the same chance of being in the sample.
* Samples can also be stratified.
  + A proportional stratified random sample is one in which the sample is selected such that the sample has the same proportion on key variables as does the population.  For example, 51% of the nation is female and 49% is male.  The sample could be stratified on sex in such a way that 51% of the sample is female and 49% is male.
  + A disproportional stratified random sample is one in which the sample is selected such that we oversample some segments and undersample other segments of the population.  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.

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.

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 in inevitable.  There is always some amount of sampling error present in every sample.  The question then is how can we reduce sampling error?

* One 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, you have to quadruple the sample size.
* Stratifying a sample is another way that you can reduce sampling error assuming that the variable you use 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.

Sampling is an important component of any research design.  You need to carefully think about how you plan to select the cases for your research study.  Exercise 2 will explore sampling in more detail and give you practice in constructing a sampling design.

## Part IV – Measurement

Let’s say that we want to explain support or opposition to same-sex marriage and that we think religion might be related to how people feel about same-sex marriage.  We can distinguish between two different dimension of religion – religious preference and religiosity.  That means that we’re dealing with three different concepts.  Our concepts are:

* support or opposition to same-sex marriage,
* religious preference, and
* religiosity.

Concepts can be defined as the abstract ideas that we want to use in our study.  Another way to think about concepts is to view them as the tools we’re going to use to try to answer our research questions.  Imagine that you go to the dentist.  The dentist has a lot of tools to take care of your teeth but not all tools are appropriate.  A chain saw is a tool but you wouldn’t want to see a chain saw in your dentist’s office.

Concepts have to be defined.  There are two different ways to define concepts.

First, there is the theoretical definition.  This answers the question – what do we mean by these concepts.

* Religious preference refers to the religion with which a person identifies.  For example, some people identify themselves as Roman Catholic, others as Lutheran, others as Jewish, and still others as Muslim.
* Religiosity refers to how religious a person is.  Two individuals could identify themselves as Roman Catholic but one might be much stronger in their religion than the other.
* Opposition or support for same-sex marriage is obvious.  Do people define themselves as favoring or opposing same-sex marriage?

Second, there is the operational definition.  How do we measure these concepts?  What are the operations we go through to measure the concepts?

* Religious identification could be measured by asking people a question such as the following:  “What is your present religion, if any? Are you Protestant, Roman Catholic, Mormon, Orthodox such as Greek or Russian Orthodox, Jewish, Muslim, Buddhist, Hindu, atheist, agnostic, something else, or nothing in particular?”[[4]](#footnote-4)
* Concepts can be measured in different ways.  Religiosity could be measured by asking people how often they attend religious services, how often they pray, and how important their religion is to them.  Here are some questions that have been used in different surveys?
  + "Aside from weddings and funerals, how often do you attend religious services... more than once a week, once a week, once or twice a month, a few times a year, seldom, or never?”[[5]](#footnote-5)
  + “About how often do you pray?”  Categories are several times a day, once a day, several times a week, once a week, less than once a week, never.[[6]](#footnote-6)
  + “Would you call yourself a strong [insert religious preference] or a not very strong [insert religious preference]?[[7]](#footnote-7)
* Here’s a question from the 2014 Pew Political Polarization Survey that was used to measure how people feel about same-sex marriage.  “Do you strongly favor, favor, oppose, or strongly oppose allowing gays and lesbians to marry legally?”

Your research design should identify the concepts that you want to use in your study and both your theoretical and operational definitions of these concepts.  Exercise 3 will explore measurement in more detail and give you practice in developing measures of various concepts.

## Part V – Data Collection

Science is an empirical enterprise.  That means that it is data based.  There are two ways that we collect data[[8]](#footnote-8):

* we observe people and use our observations as data, and
* we ask people questions and use what they tell us as data.

We’re going to focus on survey research in these exercises.[[9]](#footnote-9) Sometimes surveys are referred to as sample surveys because we select a sample of individuals from the population and ask them questions.  Then we use their answers to these questions as our data.  Surveys can take various forms:

* in-person interviews,
* mailed questionnaires,
* telephone surveys, and
* web-based surveys.

Error is inevitable whenever we study something.  Since we can’t eliminate all error our goal is to minimize error.  Error can enter into a survey in various ways.

* Sampling error occurs when we select a sample from a population.  No sample is a perfect representation of a population.
* Coverage error occurs when the list of the population from which we select our sample does not perfectly match the population.  For example, about 98% of all households in the United States have a telephone (either landline or cell or both).  So when we do a telephone survey we fail to cover about 2% of all households.
* Nonresponse error occurs when we fail to reach the entire sample.  This type of error can occur in two ways – refusals and the failure to contact some individuals in the sample.
* Measurement error occurs when our measures of some concept fall short in some way.  For example, the way we word our survey questions can introduce error.  It turns out that it matters a great deal whether we refer to global warming or climate change when we ask people questions.

Our survey design should clearly describe how we plan to collect our data.  We should consider the different ways that error might enter into the data and how we will try to minimize that error.  Exercise 4 will explore survey design in more detail and give you practice in constructing questions.

## Part VI – Data Analysis

Once we have our data, then we want to analyze the data in such a way that we can begin to answer our research questions.  Exercises 6 through 13 will explore data analysis and give you practice in analyzing survey data.  We’ll have much more to say about data analysis in these exercises.

Typically we start by looking at variables one at a time (i.e., univariate analysis).  We can use various statistical tools such as frequency distributions, measures of central tendency, and measures of dispersion to help us describe variables.

Then we look at relationships between pairs of variables (i.e., bivariate analysis).  We’re going to use crosstabulation and Chi Square to help us explore these relationships.  From here, we look at sets of variables (i.e., multivariate analysis) to see what they can tell us.

One very important point to consider is the question of causality.  Survey design can never give us a complete picture of the causal patterns in our data but it can help us begin to tease out what these causal patterns might look like.

We’ll come back to data analysis in exercises 6 through 13.

## Part VII – Research Study We’ll Be Using

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 including their family.

Questions about drug use include a variety of questions including several questions about alcohol use.  These include questions about:

* whether the respondents had ever consumed alcohol,
* how often they drank over their lifetime, in the last twelve months, in the last 30 days,
* how often they had consumed alcohol to the point of feeling “pretty high,” and
* how often during the last two weeks they had consumed “five or more drinks in a row” which is a common definition of binge drinking.

## Part VIII – Now It’s Your Turn Again

In Part 2 you were asked to write three research questions that could guide the beginning of a research study.  This time write three questions that relate specifically to drinking alcoholic beverages.  Think about what you want to find out about drinking by high school students.  For example, we might ask whether males or females are more likely to binge drink.  (Don’t use this example as one of your three questions.)

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 2 -- Sampling

## 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.[[10]](#footnote-10)  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.[[11]](#footnote-11)

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.[[12]](#footnote-12)  Since sex is related to how people feel about same-sex marriage[[13]](#footnote-13), 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. For more information, see Appendix A – Introduction to SDA.

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 37416” in the “Find Data” box.  That's the ICPSR identification number for the 2018 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.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 3 -- Measurement

## Goal of Exercise

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

## Part I—Concepts

We use concepts all the time.  We all know what a book is.  But when we use the word “book” we may not be talking about a particular book we’re reading. We could be talking about books in general.  In other words, we’re talking about the concept to which we have given the name “book.”  There are many different types of books – paperback, hardback, small, large, short, long, and so on.  But they all have one thing in common – they all belong to the category “book.”

Let’s look at another example.  Religiosity is a concept which refers to the degree of attachment that individuals have to their religious preference.  It’s different than religious preference which refers to the religion with which they identify.  Some people say they are Lutheran; others say they are Roman Catholic; still others say they are Muslim; and others say they have no religious preference.   Religiosity and religious preference are both concepts.

In other words, a concept is an abstract idea.  There are the abstract ideas of book, religiosity, religious preference, and many others.  Since concepts are abstract ideas and not directly observable, we must select measures or indicants of these concepts.  We’ll call this process measurement.

Assume that we’re interested in the following research question:  Why do some people favor same-sex marriage and others oppose it?  In other words, we’re trying to explain support or opposition to same-sex marriage.  Let’s think about the concepts that might help us explain such support.  If we think that females are more likely to support same-sex marriage then men, then sex would be a concept that we would want to include in our study.  If we think that people with more education are more likely to support same-sex marriage, then education would be a concept to include.  But not all concepts would be relevant to our study of same-sex marriage.  Hair and eye color are concepts but they wouldn’t be relevant.

## Part II – Now It’s Your Turn

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.  A major focus of this survey is students’ drug use.  Questions about drug use include a variety of questions about alcohol.  One of these questions asked how often during the last two weeks students had consumed “five or more drinks in a row” which is a common definition of binge drinking.  Let’s assume that we want to explain why some students engage in binge drinking and others do not.  List three concepts that you think might help us understand this behavior.  For each concept write a paragraph indicating why you think this concept would be useful in explaining binge drinking.

## Part III – Measures

A concept is an abstract idea.  Abstract ideas can’t be directly observed.  We have to find some piece of empirical data that we can use as a measure (or indicant) of the concept.  Let’s see if we can make this clearer by looking at some examples.

Let’s start by looking at how the Monitoring the Future Survey measured students’ drug use.  Our concept is drug use.  First, the survey recognized that there are many different types of drugs that students might use (e.g., alcohol, marijuana, cocaine, LSD, etc.).  Use of each of these drugs needs to be measured separately.  For each drug the survey asked how often students had used this drug over their lifetime, during the last twelve months, and during the last 30 days.  In other words, students’ answers to these questions become our measures of this particular type of drug use.

Now let’s consider political views or outlook.  This refers to whether respondents are politically conservative, moderate, or liberal.  How could we measure this concept?

One way to measure political views is to ask respondents the following question – “In general, would you describe your political views as conservative, moderate, or liberal?”  In other words, respondents are asked to self-report their own political views.[[14]](#footnote-14)

Another way is to use a series of questions that asks respondents whether they thought government was wasteful, whether government regulation was necessary, whether the government should do more to help needy Americans, and whether the best way to ensure peace was through military strength.[[15]](#footnote-15)  Each of these questions had a clear conservative response and a clear liberal response.  You could count the number of conservative and the number of liberal responses and use this to measure political views.

Finally, let’s think about how we could measure voter turnout.  The Pew Research Center has shown that voter turnout in U.S. presidential elections is relatively low (i.e., relative to other developed countries).[[16]](#footnote-16)  Voter turnout is the concept.  But how do we measure voter turnout?  One alternative would be to divide the number of votes cast by an estimate of the voting-age population.  Another alternative is to divide the number of votes cast by the number of registered voters.  In 2016, 55.7% of the voting age population voted but 86.8% of registered voters actually voted.  What accounts for this difference is that in the United States a lot of people who are eligible to vote don’t actually register and therefore cannot vote.  The way in which we measure concepts often makes a difference so it’s critical to be clear in explaining how we measure our concepts.

## Part IV – It’s Your Turn Again

Here are three concepts that you might want to include in any study that focuses on political behavior.  Write a paragraph for each concept indicating how you might measure that concept.  Include the questions that you would want to ask in your survey.

* Political party preference – the political party that the person identifies with.  Be sure to allow for the possibility that a person might identify with a party other than the Republican and the Democratic parties and the possibility that the person has no political party preference.
* Voting behavior – the candidate that the person is most likely to vote for if the election were held today.  You can choose any future political election.
* Likelihood of voting – the likelihood that the person will vote in an upcoming election.

## Part V – Characteristics of Good Measures

You might be wondering how we distinguish between good and bad measures.  Good measures are reliable, valid, and precise enough.[[17]](#footnote-17)

Reliability means that they are consistent.  Consistency refers to two different aspects of measures.

* Imagine that you weighed yourself in the morning.  Then you took a shower and when you got out of the shower you weighed yourself again.  You would expect the scale to give you very similar results.  If you weighed much less the second time you weighed yourself, you would question the reliability of your scale.  This refers to consistency over time.  The way you check for consistency over time is test and then retest.  For example, if you wanted to check on the reliability of your question about same-sex marriage, ask a sample of respondents at time one and then retest by asking the same respondents the same question again at time 2 (e.g., one month later).  A few respondents might change their answer but most should give you the same response.
* Most of us have taken the written test for our driver’s license.  It’s a series of multiple-choice questions.  You have to get a certain number of questions correct to pass.  There isn’t just one test.  If there was, everyone would know the questions before they took the test and probably pass.  There’s a whole set of tests with different questions.  However, we wouldn’t want some of the tests to be harder than others.  We want the tests to be equivalent.  We refer to this type of consistency as equivalence.

We also want our measure to be valid.  A measure is valid if it measures what we say it measures.  We can test for validity in several different ways.  Here are some of these ways.

* Face validity means that the measure is valid on the face of it.  There are some measures that are obviously valid.  For example, we might ask how old a person was on their last birthday.  We would probably all agree that this measure of age is obviously valid.  Notice by the way that there are some situations where we wouldn’t consider it obviously valid.  A bartender wouldn’t accept a patron’s answer to this question as obviously valid.  The bartender would want to see a photo identification such as a driver’s license.
* When you apply for a credit card, a company must have some way to decide whether to give you that credit card.  But how do they measure your credit risk?  Are you a good risk or a poor risk?  Let’s say they select a sample of 1,000 clients from all their past clients.  Then they randomly divide this into two subsamples of 500 clients each. They use the data from the first subsample to develop their credit risk measure.  This measure scores applicants on a measure of credit risk that varies from 1 (very poor risk) to 10 (very good risk).  Then they use that measure to see if it accurately predicts whether the second random subsample paid their bills or not.  If there is a strong correlation between the credit risk measure and their credit history, then they have validated their measure.  We call this predictive validity because we can determine whether our measure accurately predicts this clearly valid measure of one’s credit risk.  If it is valid, then they can use this measure to make future decisions about who should be given credit cards because they know it is a valid measure of credit risk.
* Let’s say that we want to measure a person’s mathematical skill.  Mathematics encompasses many content areas – simple arithmetic, algebra, geometry, trigonometry, calculus, and beyond.  So we decide to develop a test to measure mathematical skill.  Our test should include all these areas and not just some of them.  If it does, we can say that it has content validity since it covers the entire range of mathematical skills.

We also want our measure to be precise enough.  If you want to know what the temperature currently is, you probably only need to know it to the nearest degree.  It’s sufficient to know that it is 84 degrees Fahrenheit.  You don’t need to know that it is 84.2 degrees.  Similarly, if you want to know a person’s family income, you probably only need to know it to the nearest thousand dollars.  More precision would probably be unnecessary.  On the other hand, you might want more precision than just knowing income to the nearest tens of thousands of dollars.  You want your measure to be precise enough.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 4 – Data Collection

## Goal of Exercise

The goal of this exercise is to provide an introduction to data collection which is an integral part of any research design.  In this exercise we’re going to focus on survey research as a method of data collection.  The other elements of your research design are sampling, measurement, and data analysis which are discussed in other exercises.

## Part I—Inevitability of Error

Error is inevitable in any research study.  It’s impossible to eliminate all sources of error.  What we do is to try to identify all sources of error and then minimize error to the extent possible.  There are a number of different types of error.  In this exercise we’ll discuss four types or sources of error in survey research – sampling, coverage, nonresponse, and measurement.

## Part II – Sampling Error

No sample is ever a perfect representation of the population from which the sample is drawn.  Some error is always introduced when you take a sample from a population and this is called sampling error.  Imagine that your population is all high school seniors at a large urban high school that has 3,000 seniors.  We’re interested in the percent of seniors that engage in binge drinking.[[18]](#footnote-18)  We decide to select a simple random sample of 300 seniors from this population and ask each of these 300 seniors if they ever had five or more drinks in a row.[[19]](#footnote-19)  The percent of the sample that has engaged in binge drinking is our estimate of the population percent.  Now imagine selecting a second simple random sample of 300 seniors from this same population and asking them the same question.  You can immediately see that our two estimates of the percent of the population that binge drink would not be identical because the two samples would consist of different high school seniors.[[20]](#footnote-20)

Assuming that we are using probability sampling, sampling error depends on three factors.

* Size of the sample.  The larger the sample, the less the sampling error. That’s why we prefer a large sample to a small sample.  But there is a point of diminishing returns.  Once we have a sample of between 1,000 and 2,000 there isn’t much of a reduction in sampling error when we further increase the size of our sample.  For example, national election polls rarely use a sample much larger than 1,500 to estimate the percent of the population that intend to vote for a particular candidate.
* Amount of variability in the population.  The more variability there is in the population, the more the sampling error.  If the entire population intended to vote for the same candidate, there wouldn’t be any sampling error.  If the population was evenly divided as to whether they were going to vote for candidate A or candidate B that would represent the maximum amount of population variability and we would have more sampling error.
* The way we select the sample.  In exercise 2 we discussed stratification.  When we stratify our sample, sampling error decreases.  This assumes that our stratification variable is related to whatever we are trying to estimate.

## Part III – Coverage Error

Coverage error occurs when the list of the population from which we select our sample does not perfectly match the population.  Think about the example from Part 2 where we selected a sample of 300 high school seniors from the population of 3,000 seniors at a large urban high school.  We would expect the high school to have an accurate list of all seniors from which we could select our sample.  In this case the list of the population would perfectly match the population and there would be no coverage error.[[21]](#footnote-21)

Here are some examples of coverage error.

* The General Social Survey is a large national probability sample of adults in the United States conducted biannually by the National Opinion Research Center at the University of Chicago.  Prior to 2006 the sample consisted of adults living in non-institutionalized settings who spoke English.  Starting in 2006, Spanish-speaking adults were included in the sample.  While the exclusion of non-institutionalized adults and those who don’t speak English or Spanish introduces two sources of coverage error, both represent small proportions of the adult population.  The exclusion of these two groups offers both cost savings and greater ease of survey administration.  Since the coverage error is relatively small, the advantages outweigh the small increase in sampling error.  However, as more non-English and non-Spanish speaking individuals immigrate to the U.S., coverage error might increase in the future.
* The Monitoring the Future Survey of high school seniors excludes seniors in Alaska and Hawaii for both cost reasons and ease of administration.  This introduces a small amount of coverage error.  Since the survey is administered in the spring of each year, it also excludes seniors who dropped out prior to the survey administration.  This too introduces a small amount of coverage error.

These examples demonstrate that small amounts of coverage error may be tolerated for cost reasons and the ease of survey administration.  But sometimes coverage error can be quite large as you will see in the next part of this exercise.

## Part IV – Now It’s Your Turn to Consider Coverage Error

We’re going to consider two hypothetical surveys and think about the types of coverage error that might occur.

* Our research center has been asked to do a survey of adults in our community regarding quality of life.  Specifically, we want to determine how satisfied respondents are with different areas of their local community including the local economy, level of crime, road conditions, health services, and education.  We decide to do a phone survey of households in our community.
  + Suppose we select a sample of phone numbers from the local phone directory.  What types of coverage error would that introduce?  Do you think the amount of coverage error would be fairly small or quite large?  Why?
  + Someone points out to us that published phone directories do not typically include cell phones or people with unlisted numbers so we contact a reputable research service that provides samples of both cell phone and landline numbers as well as people with unlisted numbers.  What would that do to our coverage error?  Would there be any remaining sources of coverage error?  What would they be?  Do you think they would be fairly small or quite large?  Why?
* Your research center has been asked to do a survey of all members of a particular religious group (e.g. Roman Catholics, Lutherans, Buddhists, Muslims) in a large metropolitan city (e.g., Los Angeles, Chicago, New York).  The problem is that there is no list of the population for any of these religious groups.  So you decide to do a multi-stage cluster sample.  First, you select a sample of centers of worship (e.g., churches or temples or mosques).  Then you go to each of the centers in your sample and request a membership list.[[22]](#footnote-22)  But you notice another problem.  A number of these centers have inadequate membership lists.   A close examination of these lists indicates that some former members who have died or moved away are still on the list.  To make matters worse, newer members have sometimes not been added to the list.  It’s even possible that some members appear twice on the list under different names.  What types of coverage error might these problems create?  What would be a possible solution to these problems?

## Part V – Nonresponse Error

Nonresponse error occurs when part of your sample does not respond to your survey.  This can occur in two ways.  Sometimes we are not able to locate or contact members of the sample.  Other times sample respondents refuse to take part in the survey.  Response rates can be calculated in various ways but typically it is computed by dividing the number of respondents who complete the survey by the number of eligible respondents.  For example, if your sample consisted of 1,000 eligible individuals and 200 completed the survey, the response rate would be 20%.  If 700 completed the survey, the response rate would be 70%.

Let’s consider how nonresponse error might occur in the two examples from part 3.

* The General Social Survey reports response rates in the range of 60% to 82% with a response rate of 61% in 2016.  The GSS conducts in-person interviews although a few interviews are conducted on the phone when an in-person interview cannot be scheduled.  It’s easy to imagine that some potential respondents are never home or never available to be interviewed and that others simply refuse to be interviewed.  We also know that some types of respondents are more difficult to reach or convince to be interviewed.  Young males are especially difficult to reach particularly if we do not have their cell phone number.  Women tend to have a higher response rate than men.  Let’s say that our survey includes questions on same-sex marriage.  We know from other surveys that women are more likely to favor same-sex marriage than men.  If our sample underrepresents males, this could introduce a bias in our survey results.
* The Monitoring the Future Survey reports that recent surveys have a response rate in the 80% to 84% range.  Some students are absent on the day that the survey is administered in schools and other students refuse to participate.  Since the survey deals in part with drug use, it’s possible that students who use illegal drugs are more likely to stay home that day or refuse to participate.  This could introduce a bias in the estimate of the percent of high school seniors who use particular drugs.

## Part VI – Now It’s Your Turn to Consider Nonresponse Error

For each of the surveys mentioned in Part 4, discuss the ways that nonresponse might occur and the type of biases that such nonresponse might introduce into the survey results.

## Part VII – Measurement Error

Measurement error occurs when the way we measure some concept (e.g., religious preference or religiosity) introduces error into the measurement process.  Imagine that a person goes into a bar to order an alcoholic drink.  The bartender asks the person how old he or she is.  If the individual is not of legal drinking age, the person might give an inaccurate answer because they have a self-interest in appearing older and obtaining an alcoholic drink.  Therefore, the bartender asks for proof of age.

Measurement error occurs in different ways.  The way we ask questions often influences what people tell us.  One of the areas that has been in the news lately has been global warming and climate change.  Research has shown that “Republicans were less likely to endorse that the phenomenon is real when it was referred to as ‘global warming’ … rather than ‘climate change’ … whereas Democrats were unaffected by question wording.”[[23]](#footnote-23)  Other research has shown that global warming is more likely than climate change to be interpreted as caused by people.[[24]](#footnote-24)

Research has also shown than question order can influence what people tell us.  In 1997 the Gallup Poll asked, “Do you generally think that [Bill Clinton/Al Gore) is honest and trustworthy?”  Half of the sample (randomly selected) was asked the question with Clinton’s name first and the other half was asked with Gore’s name first.  When Clinton’s name appeared first he was much less likely to be perceived as honest and trustworthy than Gore when his name appeared first.  But when Clinton’s name appeared second there was only a small difference compared to Gore’s name appearing second.[[25]](#footnote-25)

Another way that measurement error can occur is when one of the responses is perceived as more socially desirable than the other responses.  Respondents might be reluctant to choose the less socially desirable response.  For example, cheating is generally seen as socially undesirable.  If we asked students if they had ever cheated, they might be more likely to say no because cheating is seen as socially undesirable.

## Part VIII – Now It’s Your Turn to Consider Measurement Error

In this section we’re going to consider 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 focus of these surveys is students’ drug use.  There were three questions asked to measure marijuana and hashish use.

* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . in your lifetime?
* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . during the last 12 months?
* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . during the last 30 days?”

Write a paragraph discussing the types of measurement error might have occurred when students were asked these questions.

## Part IX – Conclusions

In this exercise we have focused on surveys as our method of data collection.  Clearly there are other ways we might collect data.  Instead of asking people questions, we could observe their behavior and listen to what people say to each other.  We could also use data that others have collected.  The U.S. Census is a frequently used data set.  We could use diaries and letters as another source of data.  But regardless of the way we collect data we encounter possible sources of error.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 5 – Hypotheses and Hypothesis Testing

## Goal of Exercise

The goals of this exercise are to teach students how to write hypotheses and to provide an introduction to hypothesis testing.[[26]](#footnote-26)

## Part I—Hypotheses

All research starts with a question.  We could ask why some high school seniors decide to go on to college and others decide not to.  Or we could ask why some people vote Republican in presidential elections and other vote Democrat.  Our goal is to answer these questions.

Theories are systematic attempts to answer these questions.  Theories are made up of the concepts that we think will be helpful in answering these questions and propositions that indicate how these concepts are interrelated.  For example, the French sociologist Emile Durkheim wanted to know why some groups had higher suicide rates than other groups.  He argued that the concepts of integration and regulation helped explain variation in suicide rates.

Concepts are abstract ideas.  There are many different concepts used in sociology including integration and regulation as well as others such as religious preference and religiosity.  Since concepts are abstract ideas, we must find ways to measure these concepts.  Measurement is the process of finding pieces of empirical data that become our measures or indicants of the concepts.  For example, we might ask people how often they attend worship services or how often they pray and use their answers to these questions as measures of their religiosity.

We test our theory by deriving hypotheses from the theory which should be true if the theory is true and then testing these hypotheses.  For example, one of the questions that Durkheim asked was why some religious groups had higher suicide rates than other groups.  The degree of integration of the group provided one possible answer to this question.  Durkheim suggested that extremes of integration lead to higher suicide rates.  Thus, groups that were very high or very low in integration should have higher suicide rates that groups that were more in the middle on integration.  Since Protestants have lower levels of integration than Catholics or Jews, he hypothesized that Protestants should have higher suicide rates.  His hypothesis specified the relationship that he expected to find between religion and suicide rates.  In other words, a hypothesis specifies the relationship that you expect to find between your measures.  These measures are often referred to as variables.  A hypothesis must be testable.  In other words, it must be capable of being shown to be false.

Let’s think about another example.  Suppose our research question is why some people vote Republican in presidential elections and other vote Democrat.  Our theory is that those with less economic power are more likely to support political parties that attempt to change the status quo, while those with more economic power are more likely to support parties that attempt to maintain the status quo.  Our concepts are economic power and support for political parties.  In the United States, family income is one possible measure or indicant of economic power and the Democratic Party is more likely to want to change the status quo than is the Republican Party.  If our theory is true, then those with less income will be more likely to vote Democrat, while those with more income will be more likely to vote Republican.

Now we need to either collect data or find existing data to test our hypothesis.  If our hypothesis is false, then there is something wrong with the theory from which it was derived.  If our hypothesis is true, then we have support for our theory.  It’s important to keep in mind that we can’t say we have proven our theory.  Instead, we say we have support for our theory.  That’s a big difference.

## Part II – Now It’s Your Turn

Our research question is why some children do better academically than other children.  Our theory suggests that media use might help explain academic performance.  Our reasoning is that the more time children spend accessing different types of media, the less time they will have available for academics, and the more likely they will be to do poorer academically.  We recognize that there are different types of media – television, radio, newspapers, magazines, and the internet.  We select a random sample of children in our state and ask their parents to complete a survey that asks several questions about the number of hours per week their children spend accessing various types of media.  The survey also asks questions about the grades their children get in school.  We decide to use grades as our measure of academic performance even though we realize it is not a perfect measure.

Write one possible hypothesis that specifies the relationship you would expect to find between the number of hours per week that children watch television and their grades in school.  Now write another hypothesis that specifies the relationship you would expect to find between the number of hours per week that children spend on the internet and their grades.

## Part III – Monitoring the Future Survey of High School Seniors

In this exercise we’re going to consider 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 focus of these surveys is students’ drug use.  There were three questions asked to measure marijuana and hashish use.

* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . in your lifetime?
* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . during the last 12 months?
* “On how many occasions (if any) have you used marijuana (grass, pot) or hashish (hash, hash oil) . . . during the last 30 days?”

Our research question is why some high school seniors are more likely to use marijuana or hashish than others.  That’s what we’re trying to explain.  What concepts do you think might help us answer this question?  Your list probably includes some of the following variables.

* Respondent’s gender – male, female
* Where respondents grew up – farm, small city or town, medium-sized city, suburb, large city
* Religiosity
  + ​How often attended religious services – never, rarely, once or twice a month, about once a week or more
  + How important religion is in their lives – not important, a little important, pretty important, very important
* Family factors such as parent’s education
  + Highest grade father completed – less than high school, completed high school, some college, completed college, graduate or professional school after college
  + Highest grade mother completed – same categories
* Political factors
  + Political party preference – strongly Republican, mildly Republican, mildly Democrat, strongly Democrat, independent, no preference
  + Political beliefs – very conservative, conservative, moderate, liberal, very liberal
* Number of days of school missed because respondent skipped or “cut” school
* High school grades

Choose two of these variables and for each variable write a hypothesis that specifies the relationship that you expect to find between that variable and marijuana and hashish use.  For each hypothesis explain why you expect to find that relationship.  In other words, if someone asked why you expected this hypothesis to be true, what would you say?

## Part IV – Hypothesis Testing

Let’s review.  We start with our research question.  A theory is a systematic attempt to answer that question.  From our theory we derive hypotheses that should be true if the theory is true.  These hypotheses must be empirically testable.  We collect or find data that we can use to test these hypotheses.  So now we turn to the process of hypothesis testing.

How do we go about testing our hypotheses?  We can distinguish between quantitative and qualitative data analysis.  Our focus in these exercises is on quantitative analysis.  In no way does this suggest that quantitative analysis is better than qualitative analysis.  It also doesn’t suggest that it’s an either/or proposition.  Researchers often combine quantitative and qualitative analysis in a research study.

Quantitative analysis creates measures or indicants of the concepts in our theory.  These measures can then be related to each other.  For example, we can create a measure of religiosity based on how often respondents say they attend worship services and how important they say their religion is to them.  We can also create a measure of how often respondents say they use marijuana or hashish based on the questions that we discussed in Part 3.  Based on these measures we can then determine if these two measures are related to each other.  In other words, are respondents who are more religious more or less likely to use marijuana or hashish than those who are less religious?

Statistics are tools that we use to answer these questions.  The next eight exercises will give you practice using various statistical techniques such as crosstabulation, tests of significance, and measures of association.  Statistical programs such as SDA (Survey Documentation and Analysis) carry out the computation of these different statistics so you don’t have to worry about computing them.  SDA is easy to learn and use.  You don’t have to be a computer programmer to use it. There are a few simple rules to follow which we’ll discuss in these later exercises.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 6 – Introduction to Data Analysis

## Goal of Exercise

The goals of this exercise are to provide an introduction to data analysis and specifically to discuss the different levels of analysis (univariate, bivariate, multivariate) and levels of measurement (nominal, ordinal, interval, and ratio).[[27]](#footnote-27)

## Part I—Introduction to Data Analysis

All research starts with one or more research questions.  These are the questions that you want to answer in your research study.  For example, you might want to find out why some people vote Democrat and others vote Republican or you might want to find out why some people don’t vote at all.  Another question you might want to try to answer is why some favor same-sex marriage and others oppose it.

A research design is your plan of action.  It lays out how you plan to go about answering your questions.  The research design includes how you plan to select the cases for analysis (sampling), how you will measure concepts, how you plan to collect your data, and how you will analyze the data.  In this exercise we’re going to discuss two important components of data analysis – levels of analysis (univariate, bivariate, and multivariate) and levels of measurement (nominal, ordinal, interval, and ratio).

The research study that we’ll be using in these exercises is the Monitoring the Future (MTF) 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 including their family.

## Part II – Levels of Analysis

Data analysis is always cumulative.  It starts at the simplest level and then builds to more complex levels.  Analysis is often described as univariate, bivariate, and multivariate.

Univariate or one-variable analysis focuses on single variables.  For example, most data sets include age as a variable.  Univariate analysis starts by looking at the frequency distribution for age.  Frequency distributions tell us how many respondents are at each age.  Measures of central tendency (e.g., mean, median, mode) provide us with a measure of the center of the distribution.  Measures of dispersion (e.g., range, standard deviation, variance, index of qualitative variation) tell us how spread out the values are.  Measures of skewness describe how much distributions deviate from a symmetrical, bell-shaped pattern.  Measures of kurtosis tell us how peaked or flat distributions are.  Graphs or charts (e.g., pie charts, bar graphs, line graphs) provide us with a visual picture of the distribution.  It’s important to have an understanding of what each variable looks like before we begin looking at relationships between variables.

Bivariate or two-variable analysis focuses on pairs of variables.  The goal is to discover the relationship between variables.  For example, we might be interested in the relationship between age and voting behavior.  Statistical techniques include crosstabulation, correlation, and regression among others.

Multivariate analysis focuses on sets of three or more variables.  The goal is to extend our analysis to answer questions such as the following.

* Could the relationship between two variables be due to another variable?
* Why is there a relationship between two variables?
* Does the relationship vary for different types of individuals?

Statistical techniques used in multivariate analysis are an extension of crosstabulation, correlation, and regression.

Exercises seven through thirteen explore some of these statistical techniques.  We’re not going to consider all these statistical tools but we are going to cover enough to give you the tools you need to begin data analysis.  Before we begin we need to talk about levels of measurement – nominal, ordinal, interval, and ratio.  It’s important that we choose the right statistical tools in our analysis.  Our choice depends, in part, on the level of measurement of the variables we are using.

## Part III – Levels of Measurement

We use concepts all the time.  For example, religiosity is a concept which refers to the degree of attachment that individuals have to their religious preference.  It’s different than religious preference which refers to the religion with which they identify.  Some people say they are Lutheran; others say they are Roman Catholic; still others say they are Muslim; and others say they have no religious preference.   Religiosity and religious preference are both concepts.

A concept is an abstract idea.  So there are the abstract ideas of religiosity, religious preference, and many others.  Since concepts are abstract ideas and not directly observable, we select measures or indicants of these concepts.  Religiosity can be measured in a number of different ways – how often people attend church, how often they pray, and how important they say their religion is to them.

In these exercises we’re going to use the MTF Surveys of high school seniors in the United States that has been conducted yearly since 1975.  Information about these surveys is archived at the Inter-university Consortium for Political and Social Research (ICPSR) located at the University of Michigan.  You will need to get an account at ICPSR before you can use this data set.  All accounts are free.  To find out how to create an account, go to the Appendix A in these exercises – Introduction to SDA. This will also give you a brief overview of SDA.

MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of about 14,500 high school students.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.  Often, we want to describe respondents in terms of social characteristics such as age, sex, and race. These are all variables in the MTF Survey.  A weight variable is automatically applied to the data.  This will weight the data so the sample better represents the population from which the sample was selected.

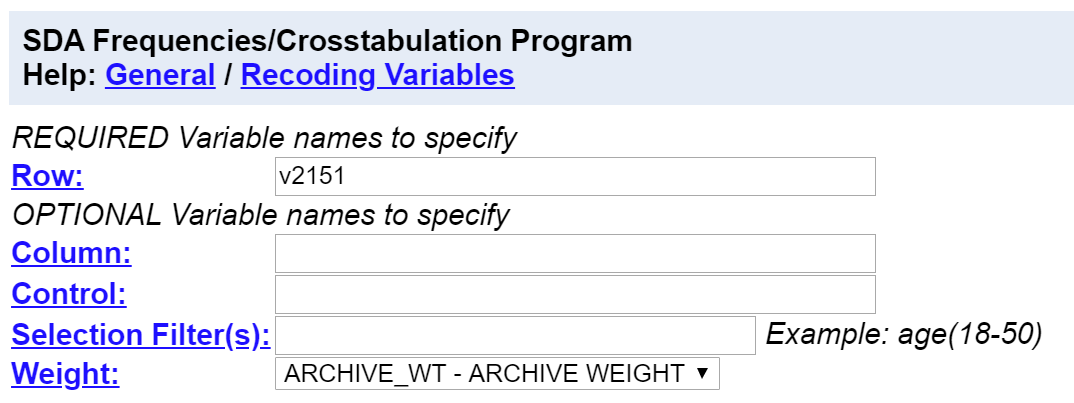
These measures are often classified in terms of their levels of measurement.  S. S.  Stevens described measures as falling into one of four categories – nominal, ordinal, interval, or ratio.[[28]](#footnote-28) Here’s a brief description of each level.

A **nominal measure** is one in which objects (i.e. in our survey, these would be high school seniors) are sorted into a set of categories which are qualitatively different from each other.  For example, we could classify individuals by their sex or race.  Students are classified as being either male or female and also classified as being white, black, or Hispanic.  There is also a separate category for students whose information is missing because they skipped that question.   Our categories should be mutually exclusive and exhaustive.  Mutually exclusive means that every individual can be sorted into one and only one category.  Exhaustive means that every individual can be sorted into a category.

The categories in a nominal level measure have no inherent order to them.  This means that it wouldn’t matter how we ordered the categories.  They could be arranged in any number of different ways.

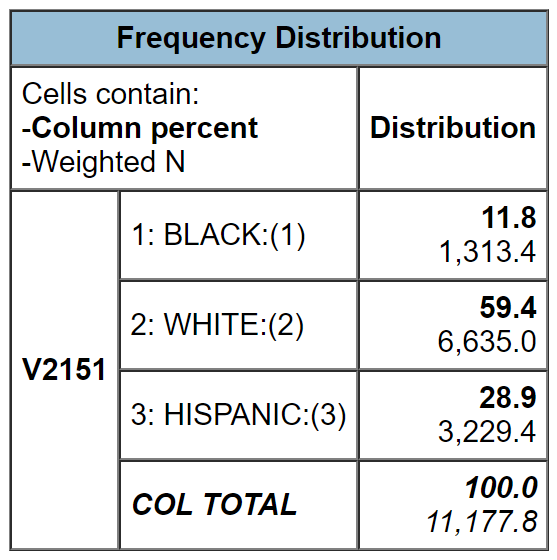
Open the 2018 Monitoring the Future survey of high school seniors. If you’re not sure how to do this, refer to Appendix A located at the end of this series of exercises. Appendix A is a brief introduction to SDA which will tell all you need to know about using SDA.

Run FREQUENCIES in SDA for the variable *V2151* which is the variable name for race so you can see the frequency distribution for a nominal level variable. To run the frequency distribution, enter the variable name, *V2151*, in the ROW box.  Your screen should look like the following.

​

Now click on RUN THE TABLE at the bottom.  Notice that the WEIGHT box is filled in for you.  This will weight the data so the sample better represents the population from which the sample was selected.[[29]](#footnote-29)

SDA will display the output in a separate window which should look like the following.  In this distribution the categories are black, white, and Hispanic.  But they could also be ordered as Hispanic, white, black or in still other ways.  It doesn’t matter how you order the categories in a nominal variable.

​

An **ordinal measure** is a nominal measure in which the categories are ordered from low to high or from high to low.  Two of the questions in the survey asked about the highest level of schooling for both the respondent’s father and mother.  Some parents attended only grade school; others attended high school but did not complete it; others graduated from high school but didn’t go on to college.  Still other individuals attended college but did not graduate; others graduated from college; and still others completed their bachelor’s degree and went on to graduate work.  These categories are ordered from low to high.

But notice that while the categories are ordered they lack an equal unit of measurement.  That means, for example, that the differences between categories are not necessarily equal.  Run FREQUENCIES in SDA for *V2163* and *V2164* which are father’s and mother’s education.  Look at the categories.  MTF assigned values (i.e., numbers) to these categories in the following way:

* 1 = grade school,
* 2 = some high school,
* 3 = high school graduate,
* 4 = some college,
* 5 = college graduate, and
* 6 = graduate school.

The difference in education between the first two categories is not the same as the difference between the last two categories.  We might think they are because 1 minus 2 is equal to 5 minus 6 but this is misleading.  These aren’t really numbers.  They’re just symbols that we have used to represent these categories. We could just as well have labeled them a, b, c, d, e, and f.  They don’t have the properties of real numbers.  They can’t be added, subtracted, multiplied, and divided.  All we can say is that b is greater than a and that c is greater than b and so on.

An **interval measure** is an ordinal measure with equal units of measurement.  For example, consider temperature measured in degrees Fahrenheit.  Now we have equal units of measurement – degrees Fahrenheit.  The difference between 20 degrees and 40 degrees is the same as the difference between 70 degrees and 90 degrees.  Now the numbers have the properties of real numbers and we can add them and subtract them.  But notice one thing about the Fahrenheit scale.  There is no absolute zero point. There can be both positive and negative temperatures.  That means that we can’t compare values by taking their ratios.  For example, we can’t divide 80 degrees Fahrenheit by 40 degrees and conclude that 80 is twice as hot as 40.  To do that we would need a measure with an absolute zero point.

A **ratio measure** is an interval measure with an absolute zero point.  For example, consider the number of siblings.  This variable has an absolute zero point and all the properties of nominal, ordinal, and interval measures and therefore is a ratio variable.

Notice that level of measurement is itself ordinal since it is ordered from low (nominal) to high (ratio).  It’s what we call a cumulative scale.  Each level of measurement adds something to the previous level.

Why is level of measurement important?  One of the things that helps us decide which statistic to use is the level of measurement of the variable(s).  For example, we might want to describe the central tendency of a distribution.  If the variable was nominal, we would use the mode.  If it was ordinal, we could use the mode or the median.  If it was interval or ratio, we could use the mode or median or mean.  Central tendency will be the focus of Exercise 7.

## Part IV – Now It’s Your Turn

Run FREQUENCIES in SDA for the following variables.

* *V2152* – where the respondent grew up
* *V2166* – political party preference
* *V2167* – political views
* *V2169* – how often respondent attends religious services
* *V2178* – number of days of school in the last four weeks that respondent skipped or cut
* *V2108* – number of times in the last two weeks that respondent had five or more drinks in a row
* *V2116* – number of times that respondent used marijuana or hashish in the last twelve months

For some of these variables, you will want to see the wording of the question to know what the response categories mean.  To do that find the question in the mini codebook on the left and click on that variable.  That will insert that variable in the VARIABLE SELECTION box.  All you have to do is click on the VIEW button and the question wording will be displayed.  See the Introduction to SDA in Appendix A for more instructions on how to do this.

For each variable, decide which level of measurement it represents and write a sentence or two indicating why you think it is that level.  Sometimes respondents fail to answer the question either because they don’t know the answer or they refuse.  These responses need to be designated as missing values.  Indicate in your answer for each variable which value(s) should be designated as missing values.  Do not take them into account when you decide on the level of measurement for that variable.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 7 – Central Tendency and Dispersion

## Goal of Exercise

The goal of this exercise is to explore measures of central tendency (mode, median, and mean) and dispersion (range, standard deviation, and variance). The exercise also gives you practice in using FREQUENCIES in SDA.

## Part I – Measures of Central Tendency

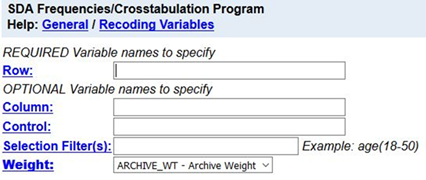
Data analysis always starts with describing variables one-at-a-time.  Sometimes this is referred to as univariate (one-variable) analysis.  Central tendency refers to the center of the distribution.

There are three commonly used measures of central tendency – the mode, median, and mean of a distribution.  The mode is the most common value or values in a distribution.[[30]](#footnote-30)  The median is the middle value of a distribution.[[31]](#footnote-31)  The mean is the sum of all the values divided by the number of values.

We’re going to use the Monitoring the Future (MTF) Survey of high school seniors for this exercise.  The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been an annual survey ever since.

Open the 2018 MTF survey of high school seniors in 2018. You’ll need to refer to Appendix A located at the end of this series of exercises. Appendix A is a brief introduction to SDA which will tell all you need to know about using SDA.

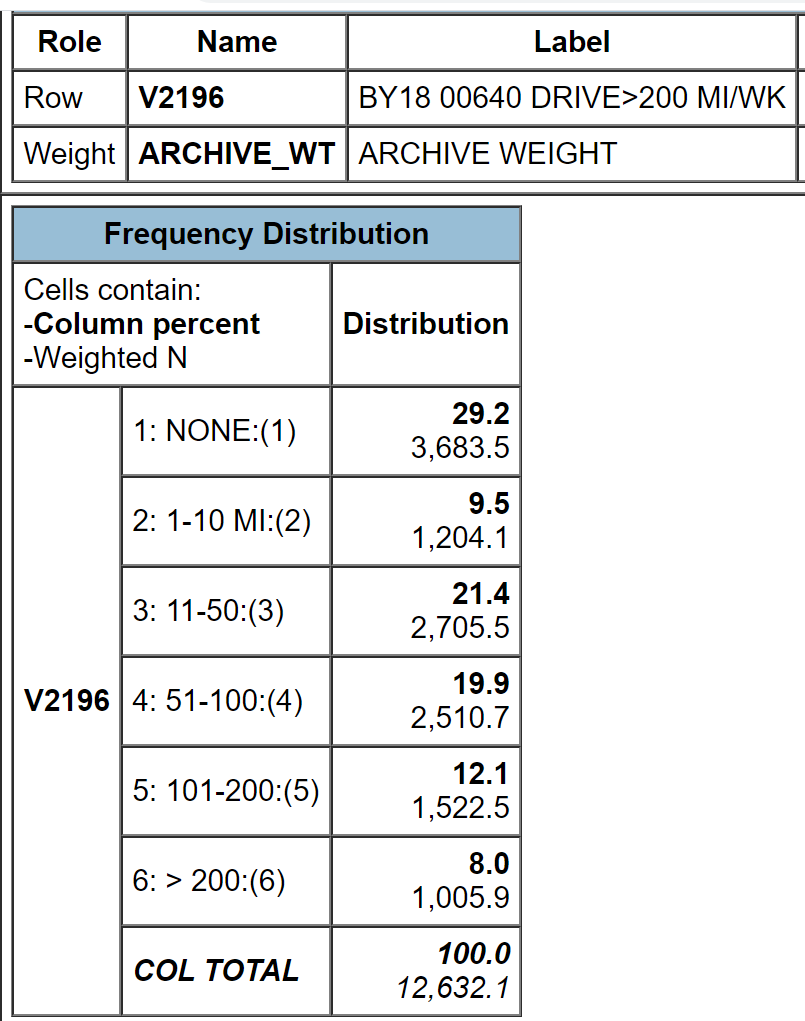
Your screen should look like the following.



Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.

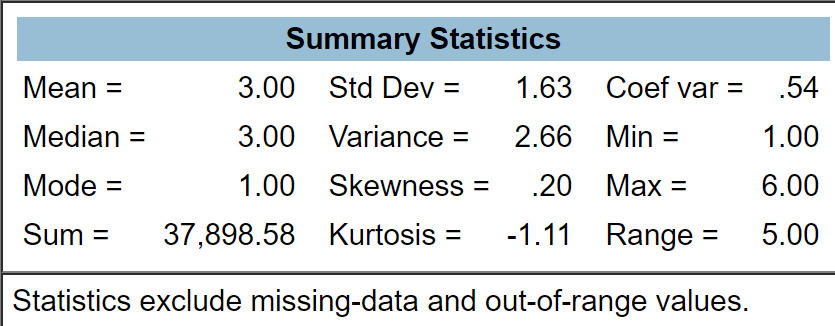
MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of about 14,500 students.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.

Run FREQUENCIES in SDA for the variable *v2196*.  This variable is the number of miles per week that students drive.  Here’s the question from the survey – “During an average week, how much do you usually drive a car, truck, or motorcycle?”  To run the frequency distribution, enter the variable name, *v2196*, in the ROW box.  The WEIGHT box is already filled in.  Click on RUN THE TABLE to get the frequency distribution.  Your screen should look like the following.



The responses to this question were divided into a set of six categories – none, 1 to 10, 11 to 50, 51 to 100, 101 to 200, and more than 200.  This was done to make the question easier to answer.  It’s difficult for respondents to remember the precise number of miles they drove per week.  It’s a lot easier to select one of these categories.  But this means that we don’t have the exact number of miles driven.  Keep that in mind as we think about measures of central tendency.

Rerun the table but this time check the box for SUMMARY STATISTICS under TABLE OPTIONS and click on the drop-down arrow next to TYPE OF CHART and select BAR CHART.  Below the frequency distribution you should see the statistics that SDA computes for you and the bar chart.  The summary statistics should look like the following.



Your output will display a number of summary statistics.  Three of these statistics are commonly used measures of central tendency – mode, median, and mean.

* Mode = 1 meaning that the first category, none, was the most common answer (29.2%) from the 12,632 respondents who answered this question.  However, not far behind are those who drove 11 to 50 miles (21.4%).  So while technically the first category (none) is the mode, what you really found is that the most common values are one (none) and three (11 to 50 miles).  Another part of your output is the bar chart which is a chart or graph of the frequency distribution.  The bar chart clearly shows that categories one and three are the most common values (i.e., the highest bars in the bar chart) with category 4 not far behind.  So we would want to report that these two categories are the most common responses.
* Median = 3 which means that the third category, 11 to 50 miles, is the middle category for this distribution.  The middle category is the category that contains the 50th percentile which is the value that divides the distribution into two equal parts.   In other words, it’s the value that has 50% of the cases above it and 50% of the cases below it.  If you added up the percents for all values less than 3 and the percents for all values less than or equal to  3, you would find that 38.7% of the cases drove 10 miles or less and that 60.1% of the cases drove 50 miles or less.  So the middle case (i.e., the 50th percentile) falls somewhere in the category of 11 to 50 miles.  That is the median category.
* Mean = 3.00.  Clearly this is wrong.  The mean number of miles driven can’t be 3.00 miles.  SDA has computed the mean of the categorical values for this variable.  In other words, SDA has added up all the 1’s, 2’s, 3’s, 4’s, 5’s, and 6’s and divided that sum by the total number of cases.  Notice that SDA gives you the sum which is 37,898.58.  To get the mean, SDA divided that sum by 12,632.1 which equals 3.00 which is the mean of the categorical values.[[32]](#footnote-32)  Let’s see if we can figure out a way to get SDA to compute the actual mean and not just the mean of the categorical values.

We can do this be changing the categorical values so they are the midpoint of the miles driven for each category.  That would mean we would have to do the following.

* We would change the categorical value of 1 to 0 which is the number of miles driven for this category.
* Change 2 to 5.5 which is the midpoint of category 2. To find the midpoint, add the smallest value in this category (1) and the largest value (10) and divide that sum by 2.
* Change 3 to 30.5 following the same procedure as above.
* Change 4 to 75.5.
* Change 5 to 150.5.
* Change 6 to 250.5.  Notice we have a problem here.  There is no upper limit to this category.  It’s simply more than 200.  We’re going to assume that nobody drives more than 300 miles per week and use 300 as our upper limit.  We have no way of knowing what the upper limit is so we’ll make what seems like a reasonable guess.

How are we going to tell SDA to make these changes?  By the way, this is called recoding.  We’re recoding the categorical values of 1, 2, 3, 4, 5, and 6 into the values above.  Follow these steps to recode in SDA.

* Enter the variable name in the row box.  The variable name in this example is *v2196*.  (Don’t enter the period.)
* After the variable name, enter (r: where r stands for recode.
* Enter the new value you want to assign to the first recode followed by the original value.  In our case we want to assign the new value 0 to the old value 1 so this would be 0=1.  (Don’t enter the period.)
* If you want to assign a label to each category, enter the label in double quotation marks. For example, our recode for the first category would look like this – *v2196* (r:0=1”none”;.  (Don’t enter the period.)  We’re going to omit the labels in this exercise for simplicity sake.
* One problem is that SDA won’t allow you to recode using a fractional value so we’re going to drop the .5 in the midpoints.  That means that we will enter the midpoints as 0, 5, 30, 75, 150, and 250.  This will give us a slight underestimate of the mean but by a very small amount.
* Separate the recodes by a semicolon.
* Repeat this process for each recode.  For example, for the second category it would look this – 5=2.  (Don’t enter the period.)
* After the last recode, end the statement with a right parenthesis.
* This is what our recode statement would look like – v2196(r:0=1;5=2;30=3;75=4;150=5;250=6).  (Don’t enter the period.)

Now tell SDA to compute the summary statistics for the recoded variable.  The mean should be 59.79 this time.  Notice that the mode is now 0 since that is the value for the first category and the median is 30 which is in the third category.  Remember that this is based on the assumption that all the cases in each category fall at the midpoint of that category.

## Part II – Now it’s Your Turn

One of the variables in the data set is *v2197* which is the number of driving tickets respondents received in the last twelve months.  The response categories are 0, 1, 2, 3, and 4 or more.  The only problem is the last open-ended category.  Let’s assume that no one received more than six tickets.  So the last category would be 4 to 6 with a midpoint of 5.  Follow the procedure described in Part I and compute the mode, median, and mean.  Write a paragraph discussing what these measures of central tendency mean.

## Part III – Deciding Which Measure of Central Tendency to Use

The first thing to consider is the level of measurement (nominal, ordinal, interval, ratio) of your variable (see Exercise 6).

* If the variable is nominal, you have only one choice.  You must use the mode.
* If the variable is ordinal, you could use the mode or the median.  You should report both measures of central tendency since they tell you different things about the distribution.  The mode tells you the most common value or values while the median tells you where the middle of the distribution lies.
* If the variable is interval or ratio, you could use the mode or the median or the mean.  Now it gets a little more complicated.  There are several things to consider.
  + How skewed is your distribution?  Go back and look at the bar chart for *v2196*. Notice that there is a long tail to the right of the distribution.  The category with the largest number of cases is the first category which represents those who did not drive at all.  But there are quite a few respondents who report driving quite a bit.  For example, 12.1% report driving between 101 and 200 miles and 8.0% said they drive more than 200 miles per week.  That’s what we call a positively skewed distribution where there is a long tail towards the right or the positive direction. Now look at the median and mean for the recoded variable.  The mean (59.79) is larger than the median (30.0).  The respondents that drove a lot of miles pull the mean up.  That’s what happens in a skewed distribution.  The mean is pulled in the direction of the skew.  The opposite would happen in a negatively skewed distribution.  The long tail would be towards the left and the mean would be lower than the median.  In a heavily skewed distribution the mean is distorted and pulled considerably in the direction of the skew.  So consider reporting only the median in a heavily skewed distribution.  That’s why you almost always see median income reported and not mean income.  Imagine what would happen if your sample happened to include Bill Gates.  The income distribution would have this very, very large value which would pull the mean up but not affect the median.
  + Is there more than one clearly defined peak in your distribution?   For example, consider a hypothetical distribution of 100 cases in which there are 50 cases with a value of two and fifty cases with a value of 8.  The median and mean would be five but there are really two centers of this distribution – two and eight.  The median and the mean aren’t telling the correct story about the center. You’re better off reporting the two clearly defined peaks of this distribution and not reporting the median and mean.

Run FREQUENCIES for the following variables.  Once you have entered the variable names in the ROW box, ask for the SUMMARY STATISTICS and a BAR CHART.  For each variable write a sentence or two indicating which measure(s) of central tendency (i.e., mode or median) would be appropriate to use to describe the center of the distribution and what the values of those statistics mean.  For some variables there will be more than one appropriate measure of central tendency.

* *v49* – number of siblings
* *v2108* – number of times had five drinks in a row in last two weeks
* *v2116* – number of times used marijuana or hashish in the last 12 months
* *v2151* – race
* *v2169* – how often attend religious services
* *v2173* – how rate self on school ability

## Part IV – Measures of Dispersion or Variation

Dispersion or variation refers to the degree that values in a distribution are spread out or dispersed.  The most commonly used measures – range, standard deviation, variance – are only appropriate for interval and ratio level variables (see Exercise 6). The variables in the MTF survey are entirely nominal and ordinal variables but as you have seen in this exercise, we can recode some of these variables so they are ratio variables.

The range is the difference between the highest and the lowest values in the distribution.  We don’t actually know the highest value for *v2196* since the last category is more than 200 miles.  Earlier in this exercise we assumed that the largest value was 300.  If that is the case, what would the range be for the recoded variable?

The range is not a very stable measure since it depends on the two most extreme values – the highest and lowest values.  These are the values most likely to change from sample to sample.

The variance is the sum of the squared deviations from the mean divided by the number of cases minus 1 and the standard deviation is just the square root of the variance.  Your instructor may want to go into more detail on how to calculate the variance by hand.  Look back at the summary statistics for *your recode of v2196*.  The variance equals 5,426.77.  What will the standard deviation equal?

The variance and the standard deviation can never be negative.  A value of 0 means that there is no variation or dispersion at all in the distribution.  All the values are the same.  The more variation there is, the larger the variance and standard deviation.

So what does the variance and the standard deviation for *v2196* mean?  That’s hard to answer because you don’t have anything to compare it to.  But if you knew the standard deviation for both men and women you would be able to determine whether men or women have more variation.  Instead of comparing the standard deviations for men and women you would compute a statistic called the Coefficient of Relative Variation (CRV).  CRV is equal to the standard deviation divided by the mean of the distribution.   A CRV of 2 means that the standard deviation is twice the mean and a CRV of 0.5 means that the standard deviation is one-half of the mean.  You would compare the CRV’s for men and women to see whether men or women have more variation relative to their respective means.

How do we get SDA to compute the means and standard deviations for both men and women?  Click on ANALYSIS and then on COMPARISON OF MEANS in the blue horizontal bar at the top of your screen.  Enter the variable for which you want to compute the mean and standard deviation in the DEPENDENT box.  We're going to use the same variable we used in Part I (v2196).  Be sure to enter the recode that you used in part 1.  Enter the variable (*V2150*) that you want to use to divide the sample into men and women in the ROW box.  SDA will automatically calculate the mean number of miles driven for both men and women.  To get the standard deviations, check the STD DEV box under TABLE OPTIONS.  The mean will be the top number in each box of your output and the standard deviation will be right below the mean.  Compute the Coefficient of Variation for both men and women and write a sentence or two discussing whether men or women have more variation.

By the way, you might also have wondered why you need both the variance and the standard deviation when the standard deviation is just the square root of the variance.  You’ll just have to take my word for it that you will need both as you go further in statistics.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 8 – Graphs and Charts

## Goals of Exercise

The goal of this exercise is to explore graphs and charts for frequency distributions. The exercise also gives you practice in using FREQUENCIES in SDA.

## Part I – Pie Charts

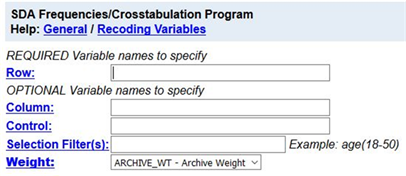
Data analysis always starts with describing variables one-at-a-time.  Sometimes this is referred to as univariate (one-variable) analysis.  Graphs and charts are useful tools for displaying visually what the distribution of responses to a question look like.

A pie chart is a chart that shows the frequencies or percents of a variable with a small number of categories.  It is presented as a circle divided into a series of slices.  The area of each slice is proportional to the number of cases or the percent of cases in each category.  It is normally used with nominal or ordinal variables (see Exercise 6).

We’re going to use the Monitoring the Future (MTF) Survey of high school seniors for this exercise.  The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been an annual survey ever since.

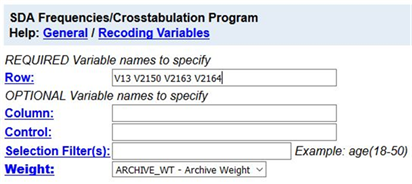
Open the 2018 MTF survey of high school seniors in 2018. You’ll need to refer to Appendix A located at the end of this series of exercises. Appendix A is a brief introduction to SDA which will tell all you need to know about using SDA

Your screen should look like the following.  Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected

​

MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of a little more than 14,500 seniors.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.  Often, we want to describe respondents in terms of characteristics such as the region of the country in which the respondent’s school is located (variable name is *v13*), respondent’s sex (*v2150*), father’s education (*v2163*), and mother’s education (*v2164*).  These are all variables in the 2018 MTF survey.

Run FREQUENCIES in SDA for the variables *v13*, v2150, *v2163*, and *v2164.* To run the frequency distributions, enter the variable names in the ROW box.  Separate the variable names by either a space or a comma.  Notice that the WEIGHT box is filled in. Your screen should like the following.

​  
  
Once you have selected the variables, click on the down arrow next to TYPE OF CHART and select PIE CHART.  Click also on the box to SHOW PERCENTS so SDA will print the percents on the pie chart.  If you want, you can check the box for SUPRESS TABLE which is under TABLE OPTIONS so SDA will not print out the frequency distribution.  Now click on RUN THE TABLE at the bottom.  SDA will draw the pie chart for each of these variables.  Write a sentence or two for each variable describing the distributions based on these pie charts.

## Part II – Bar Charts

A bar chart is a chart that shows the frequencies or percents of a variable and is presented as a series of vertical bars.  The height of each bar is proportional to the number of cases or the percent of cases in each category.  It is normally used with nominal or ordinal variables (see Exercise 6).

Run FREQUENCIES for the same variables that you used in part 1 (*v13*, v2150, *v2163*, and *v2164)*.  This time click on the down arrow next to TYPE OF CHART and select BAR CHART.  Click also on the box to SHOW PERCENTS so SDA will print the percents on the bar chart.  Now click on RUN THE TABLE to produce the bar charts.  Write a sentence or two for each variable describing the distributions based on the bar charts.

## Part III – Stacked Bar Charts

SDA will also produce what it calls a stacked bar chart.  To get a stacked bar chart, click on the down arrow next to TYPE OF CHART and select STACKED BAR CHART.  Click also on the box to SHOW PERCENTS so SDA will print the percents on the stacked bar chart.  Now click on RUN THE TABLE to get the stacked bar charts for the variables in Parts 1 and 2.  Write a short paragraph describing the stacked bar chart and how it is different from a bar chart.  Which do you prefer – bar charts or stacked bar charts?  Why?

## Part IV – Line Charts

The last type of chart that SDA will produce is called a LINE CHART.  To get a line chart, click on the down arrow next to TYPE OF CHART and select LINE CHART.  Click also on the box to SHOW PERCENTS so SDA will print the percents on the line chart.  Now click on RUN THE TABLE to get the line charts for the variables in Parts 1, 2, and 3.  Write a short paragraph describing the line chart and how it is different from the other types of charts.  Do you think a line chart is clearer than or not as clear as the other types of charts?  Why?

## Part V – Conclusions

We have talked about four different types of graphs – pie charts, bar charts, stacked bar charts, and line charts.  Are there limitations on when you should use a particular type of chart?  Why?

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 9 – Crosstabulation

## Goals of Exercise

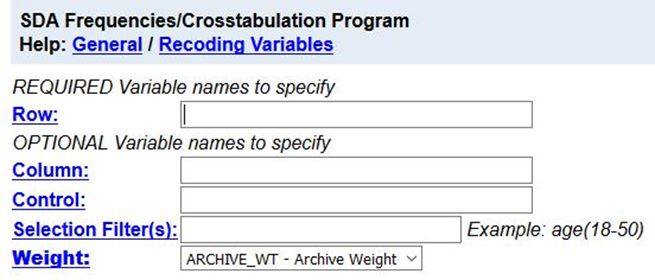
The goal of this exercise is to introduce crosstabulation as a statistical tool to explore relationships between variables.  The exercise also gives you practice in using CROSSTABS in SDA.

## Part I—Relationships between Variables

We’re going to use the Monitoring the Future (MTF) Survey of high school seniors for this exercise.  The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been an annual survey ever since.

Open the 2018 MTF survey of high school seniors. You’ll need to refer to Appendix A located at the end of this series of exercises which is a brief introduction to SDA and will tell all you need to know about using SDA.

Once you have opened SDA, your screen should look like the following.

​

Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected

MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of about 14,500 seniors.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.

In Exercises 6, 7, and 8 we described variables one-at-a-time which is typically referred to as univariate (i.e., one variable) analysis.  For example, we computed the percent of men and women in the sample.

But what if we wanted to explore the relationship between variables?  What if we wanted to know if sex was related to binge drinking?  (Binge drinking is often defined as having five or more drinks in a row.)  Crosstabulation can be used to look at the relationship between nominal and ordinal variables (see Exercise 6). When we look at the relationship between two variables, we often call this bivariate analysis.

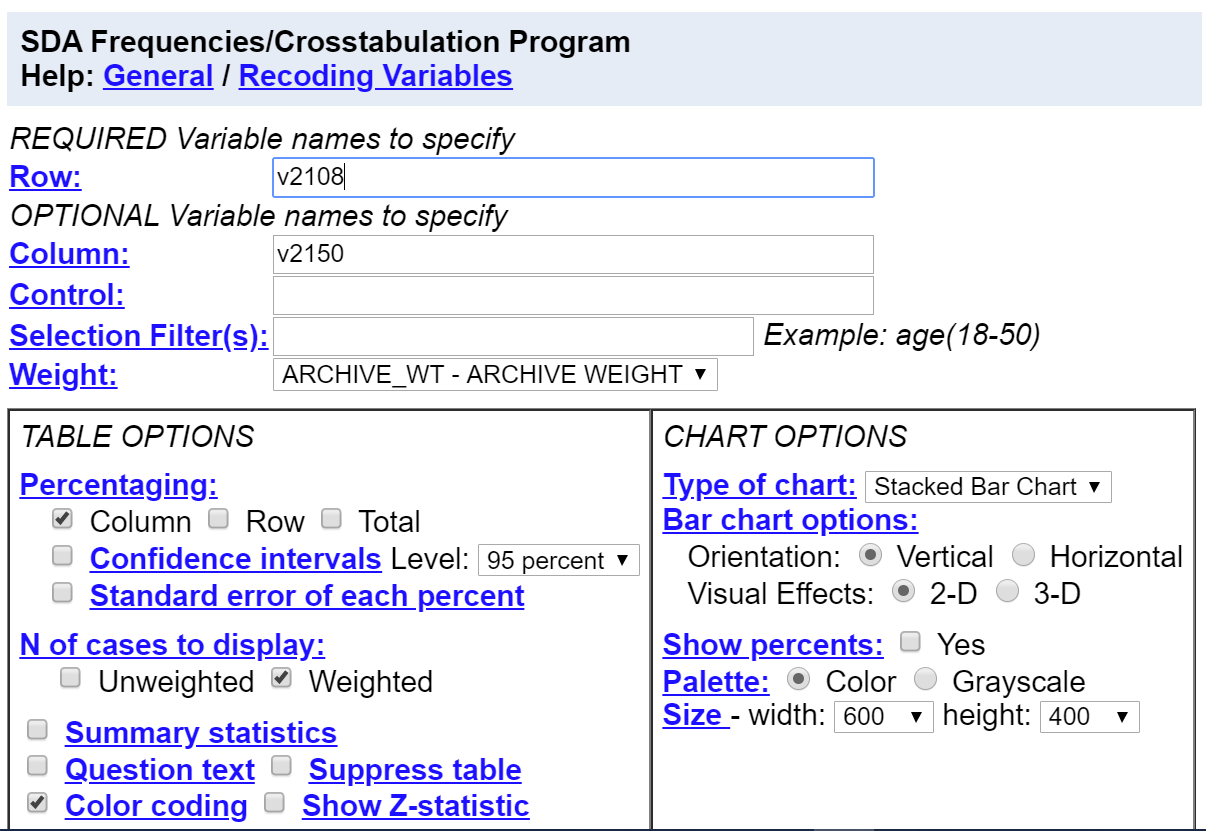
Before we look at the relationship between two variables, we need to talk about independent and dependent variables.  The dependent variable is whatever you are trying to explain.  In our case, that would be why some students engage in binge drinking and others don’t.  The independent variable is some variable that you think might help you explain why some students binge drink.  In our case, that would be sex.  Normally we put the dependent variable in the row and the independent variable in the column of our table.  We’ll follow that convention in this exercise.

First, we have to locate the variable names for sex and binge drinking.  On the left of the screen that you just opened you should see a mini codebook.  You can double click on any category in the codebook to see the list of variables in that category.  Double click on DEMOGRAPHIC/RESPONDENT CHARACTERISTICS and then on RESPONDENT CHARACTERISTICS  and one of the variables should be the respondent's sex.

Now let's click on SUBSTANCE USE and then on ALCOHOL.  Now you can see that *v2150* is the variable name for sex and that *v2108* is the name for binge drinking.

Since *v2150* is the independent variable it should go in the column of our table and since *v2108* is the dependent variable it will go in the row.  Enter these variable names in the appropriate boxes.

To interpret the table, you will need to compute percents.  SDA can compute row percents, column percents, and total percents.  Look at the TABLE OPTIONS section of your screen and you will see the boxes to check to indicate which percents you want SDA to compute.  By default, the box for column percents is already checked. Your screen should look like the following.

​

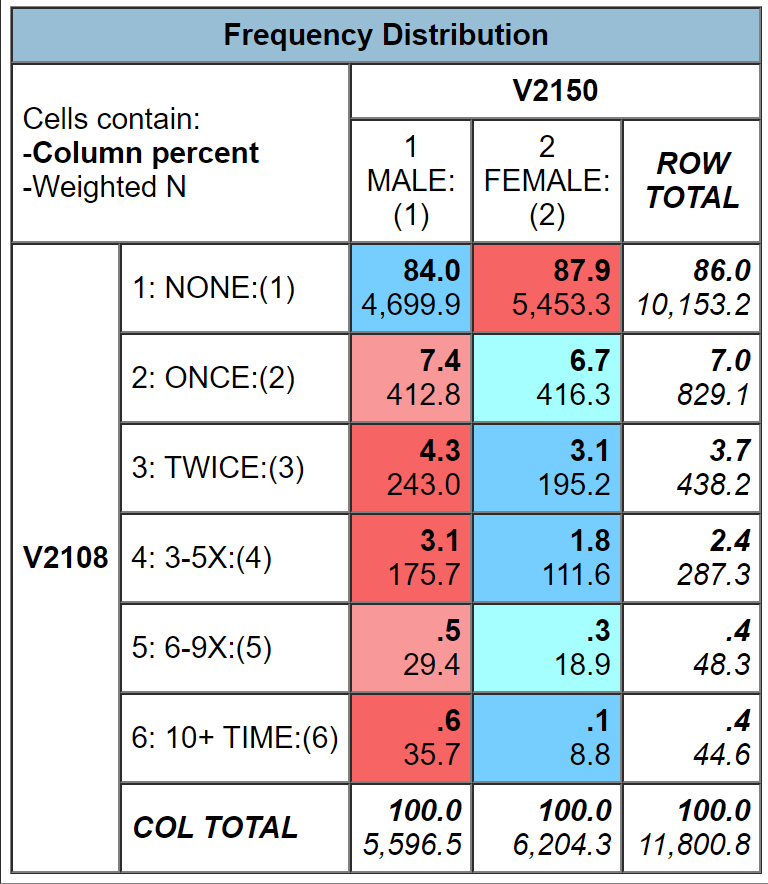
Your instructor will probably talk about how to compute these different percents.  But how do you know which percents to ask for?  Here’s a simple rule for computing percents.

* If your independent variable is in the column, then you want to use the column percents.
* If your independent variable is in the row, then you want to use the row percents.

Since you put the independent variable in the column, you want the column percents.  Click on RUN THE TABLE to get the crosstabulation.

## Part II – Interpreting the Percents

Your first table should look like this.[[33]](#footnote-33)

​

It’s easy to make sure that you have the correct percents.  Your independent variable (sex) should be in the column and it is.  Column percents should sum down to 100% and they do.

How are you going to interpret these percents?  Here’s a simple rule for interpreting percents.

* If your percents sum down to 100%, then compare the percents across.
* If your percents sum across to 100%, then compare the percents down.

Since the percents sum down to 100%, you want to compare across.

Look at the first row.  Approximately 84% of men have never engaged in binge drinking compared to approximately 88% of women.  There’s a difference of 4% which is fairly small.  We never want to make too much of small differences.  Why not?  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 in inevitable.  There is always some amount of sampling error present in every sample.  The larger the sample size, the less the sampling error and the smaller the sample size, the more the sampling error.  Since our sample size is so large, sampling error will be quite small.  That means that in this case we can conclude that females are a little less likely to engage in binge drinking than men.  If our sample size had been a lot smaller, we would have concluded there probably wasn’t much difference in the population between men and women for binge drinking.  Note that we’re using our sample data to make inferences about the population.

## Part III – Now it’s Your Turn

Choose two of the variables from the following list and compare men and women.

* how likely will attend a technical or vocational school after high school (*v2180*)
* how likely will graduate from a four-year college after high school (*v2183*)
* how likely will attend graduate or professional school after high school (*v2184*)
* how likely will serve in the armed services after high school (*v2181*)

Make sure that you put the independent variable in the column and the dependent variable in the row.  Be sure to ask for the correct percents.  What are values of the percents that you want to compare?  What is the percent difference?  Does it look to you that there is much of a difference between men and women in the population for the variables you chose?  How did you decide?

So far, we have only looked at variables two at a time.  Often, we want to add other variables into the analysis which is typically called multivariate analysis (i.e., analysis with more than two variables).  We’ll consider multivariate analysis in Exercise 12.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 10 – Chi Square

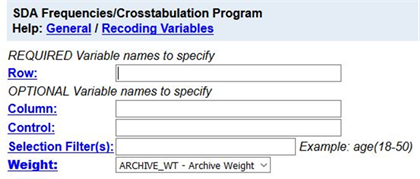
## Goals of Exercise

The goal of this exercise is to introduce Chi Square as a test of significance.  The exercise also gives you practice in using CROSSTABS in SDA.

## Part I—Relationships between Variables

We’re going to use the Monitoring the Future (MTF) Survey of high school seniors for this exercise.  The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been an annual survey ever since.

Open the 2018 MTF survey of high school seniors. You’ll need to refer to Appendix A located at the end of this series of exercises which is a brief introduction to SDA and will tell all you need to know about using SDA. After opening SDA, your screen should look like the following.

​

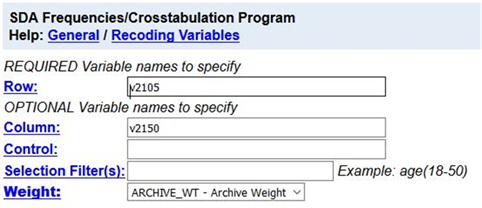
Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected.

MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of about 14,500 seniors.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.

In Exercise 9 we used crosstabulation and percents to describe the relationship between pairs of variables in the sample.  But we want to go beyond just describing the sample.  We want to use the sample data to make inferences about the population from which the sample was selected.  Chi Square is a statistical test of significance that we can use to test hypotheses about the population.  Chi Square is the appropriate test when your variables are nominal or ordinal (see Exercise 6).

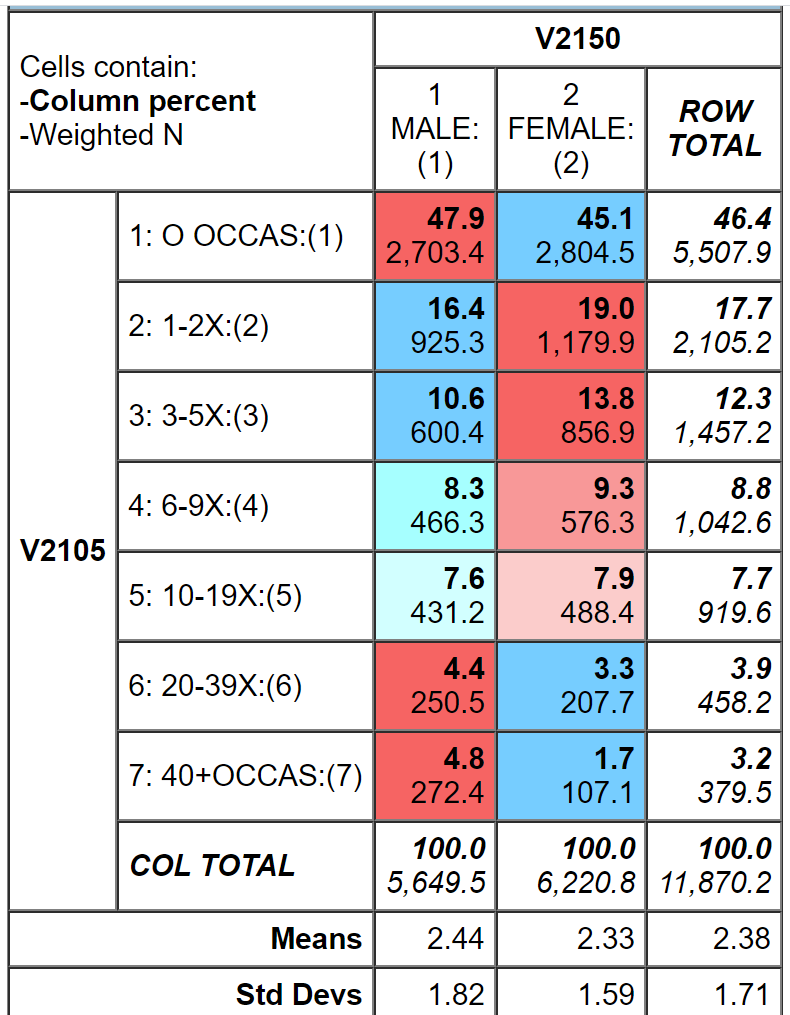
Before we look at the relationship between variables, we need to talk about independent and dependent variables.  The dependent variable is whatever you are trying to explain.  For example, we could be trying to explain why some students have consumed alcoholic beverages during the last year and others have not.  The independent variable is some variable that you think might help you explain why some students have consumed alcoholic beverages.  We’re going to use sex as our independent variable.  Normally we put the dependent variable in the row and the independent variable in the column of our table.  We’ll follow that convention in this exercise.

Run CROSSTABS in SDA to produce the crosstabulation of *v2105 (how often consumed alcoholic beverages in last year)* and *v2150 (sex)*.  Look at PERCENTAGING under TABLE OPTIONS.  Since your independent variable is in the column, you want to use the column percents.  By default, the box for column percents is already checked. Your screen should look like the following. Notice that the WEIGHT box is filled in.  Click on RUN THE TABLE to produce the crosstabulation.

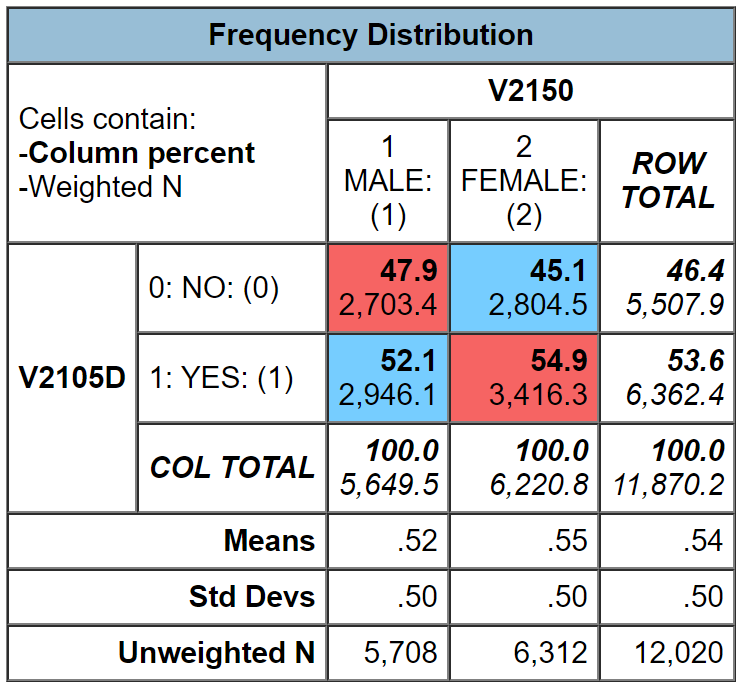
​

## Part II – Interpreting the Percents

Now your table should look like the following.[[34]](#footnote-34)

​

One of the first things you notice is that a large percent of students (46.4%) have not consumed alcoholic beverages in the last 12 months and a much smaller percent have consumed them a large number of times (7.1% engaged in drinking 20 or more times).  It might be easier to divide *v2105* into two categories – never and one or more times.  Fortunately, the data set has such a variable.  The dichotomous variable is *v2105d*.  (The d stands for dichotomy.)  Rerun the table you just ran replacing *v2015* with *v2015d*.  Your screen should look like the following.

​

Since your percents sum down to 100% (i.e., column percents), you want to compare the percents across.  Look at the first row. Approximately 45.1% of men have not consumed alcoholic beverages in the last year compared to 47.9% of women which is a difference of only 2.8 percentage points. We never want to make too much of small differences.  Why not?  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.  The larger the sample size, the less the sampling error and the smaller the sample size, the more the sampling error.

But what is a small percent difference?  If you think that three percent is a small difference, what about a four or five or six or seven percent difference?  Is that small?  Or is it large enough for us to conclude that there is a difference between men and women **in the population?**  Here’s where we can use Chi Square.

## Part III – Chi Square

Let’s assume that you think that males and females have different drinking patterns.  We’ll call this our research hypothesis.  It’s what we expect to be true.  But there is no way to prove the research hypothesis directly.  So, we’re going to use a method of indirect proof.  We’re going to set up another hypothesis that says that the research hypothesis is not true and call this the null hypothesis.  In our case, the null hypothesis would be that the two variables are unrelated to each other.[[35]](#footnote-35)   In statistical terms, we often say that the two variables are independent of each other. If we can reject the null hypothesis, then we have evidence to support the research hypothesis. If we can’t reject the null hypothesis, then we don’t have any evidence in support of the research hypothesis.  You can see why this is called a method of indirect proof. We can’t prove the research hypothesis directly but if we can reject the null hypothesis then we have indirect evidence that supports the research hypothesis.

Here are our two hypotheses.

* research hypothesis – the two variables, sex and drinking, are related to each other
* null hypothesis – the two variables, sex and drinking, are unrelated to each other; in other words, they are independent of each other

It’s the null hypothesis that we are going to test.

SDA will compute Chi Square for you.  Follow the same procedure you used to get the crosstabulation between *v2150 (sex)* and *v2105d (drinking)*.  Remember to get the column percents.  Check the box for SUMMARY STATISTICS under TABLE OPTIONS.  Finally, click on RUN THE TABLE.

In the SUMMARY STATISTICS part of the output, you’ll see two Chi Squares – Chisq-P and Chisq-LR.  We want to use the first one listed – Chisq-P.  This is usually referred to as the Pearson Chi Square.  The number in parentheses which in this case is 1 is the degrees of freedom.

The value of the Pearson Chi Square is 9.25.  Your instructor may or may not want to go into the computation of the Chi Square value but we’re not going to cover it in this exercise.

The degrees of freedom (df) is 1.  That’s the number inside the parentheses following the Chi Square value.  Degrees of freedom is the number of values that are free to vary.  In a table with two columns and two rows only one of the cell frequencies is free to vary assuming the marginal frequencies are fixed.  The marginal frequencies are the values in the margins of the table.  There are 5,649.5 males and 6,220.8 females in this table and there are 5,507.9 that have not consumed alcoholic beverages and 6,362.4 who have.  Try filling in any one of the cell frequencies in the table.  The other three cell frequencies are then fixed assuming we keep the marginal frequencies the same.

Now we have to decide if we should reject the null hypothesis that the two variables are unrelated (or statistically independent) based on the Chi Square value and the degrees of freedom.  Look at your output again and you’ll see that after the Chi Square value it says (p=0.00).  That is the probability that you would be wrong if you rejected the null hypothesis.  It looks like the probability is zero but it’s not.  This is actually a rounded value.  The probability is less than 0.005.  There is some very small chance that you would be wrong if you rejected the null hypothesis.  With odds like that, of course, we’re going to reject the null hypothesis.  A common rule is to reject the null hypothesis if the significance value is less than .05 or less than five out of one hundred.  Since <0.005 is smaller than .05, we reject the null hypothesis.  That means that there is support for our research hypothesis that sex is related to drinking.

You might be wondering how such a small percent difference (i.e., 2.8 percentage points) could be statistically significant.  It’s because the sample is so large. The larger the sample, the less the sampling error.  The smaller the sample, the more the sampling error.  With a sample of almost 12,000, there won’t be much sampling error and even a small percent difference could be significant.

You might be wondering why our sample is now a little less than 12,000 while before it was more than 12,000.  It's because of missing data. Some cases have missing data on one or both of the variables and these cases are omitted from the table.

## Part IV – Now it’s Your Turn

Choose any two of the tables from the following list and compare men and women using crosstabulation and Chi Square.

* ever smoked cigarettes (*v2101d*)
* ever used marijuana or hashish (*v2115d*)
* how often attend religious services (*v2169*)
* how important religion is in life (*v2170*)
* rate self on school ability (*v2173*)
* how often skipped classes in last four weeks (*v2178*)
* how likely to graduate from four-year college (*v2183*)

Make sure that you put the independent variable in the column and the dependent variable in the row.  Be sure to ask for the correct percents and Chi Square.  What are the research hypothesis and the null hypothesis?  Do you reject the null hypothesis?  How do you know?  What does that tell you about the research hypothesis?

## Part V – Expected Values

We said we weren’t going to talk about how to compute Chi Square, but we do have to introduce the idea of expected values.  The computation of Chi Square is based on comparing the observed cell frequencies (i.e., the cell frequencies that you see in the table that SDA gives you) and the cell frequencies that you would expect by chance assuming the null hypothesis was true.

Computing expected frequencies is not difficult. Look back at the crosstabulation of v2105d by v2150 and focus on the upper left cell of the table. This cell is for males who have not consumed alcoholic beverages in the last year. To get the expected frequency for this cell multiply the row marginal (5,507.9) and the column marginal (5,649.5) and divide the product by the total number of cases in the table (11,870.2). Your answer should be 2,621.4. This is the expected frequency for this cell. You would do this for all four cells in the table to get the expected frequencies.

Chi Square assumes that all the expected cell frequencies are greater than five.  Small expected frequencies might occur when you have a column or a row with a small number of cases in it.  What you’ll have to do is to combine rows or columns that have small marginal frequencies in order to increase the expected frequencies values.  You can do that in SDA but we're not going to go into it in this exercise.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 11 – Measures of Association

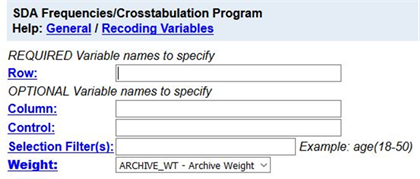
## Goals of Exercise

The goal of this exercise is to introduce measures of association.  The exercise also gives you practice in using CROSSTABS in SDA.

## Part I—Relationships between Variables

We’re going to use the Monitoring the Future (MTF) Survey of high school seniors for this exercise.  The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been an annual survey ever since.

Open the 2018 MTF survey of high school seniors. You’ll need to refer to Appendix A located at the end of this series of exercises which is a brief introduction to SDA and will tell all you need to know about using SDA. Your screen should look like the following.



Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected

MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2017 and is a relatively large sample of about 14,500 seniors.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.

In Exercise 9 we used crosstabulation and percents to describe the relationship between pairs of variables in the sample.  In Exercise 10we went beyond simple description.  We used the sample data to make inferences about the population from which the sample was selected.  Chi Square was used to test hypotheses about the population.  Chi Square is the appropriate test when your variables are nominal or ordinal (see Exercise 6).

Chi Square is a test of the null hypothesis that two variables are unrelated to each other.  Another way to put this is that the two variables are independent of each other.  If we can reject the null hypothesis then we have support for our research hypothesis that the two variables are related to each other.  But showing that two variables are related is not the same thing as determining the strength of the relationship.  The strength of a relationship is actually a continuum from very weak to very strong.  To measure the strength of a relationship we need to select and compute a measure of association.  In this exercise we’re going to focus on nominal and ordinal variables.

## Part II – What is a Measure of Association?

Before we discuss measures of association, we need to talk about independent and dependent variables.  The dependent variable is whatever you are trying to explain.  For example, let’s say we want to find out why some people think they will eventually graduate from a four-year college while others don’t.   The independent variable is some variable that you think might help you answer this question.  Perhaps we decide to use their grades in high school as our independent variable.

A measure of association is a numerical value that tells us how strongly related two variables are.  There are several characteristics of a good measure of association.

* They range from a value of 0 (i.e., no relationship) to 1 (i.e., the strongest possible relationship).
* For variables that have an underlying order from low to high they can be positive or negative.  A positive value indicates that as one variable increases, the other variable also increases.  A negative value indicates that as one variable increases, the other variable decreases.[[36]](#footnote-36)
* Some measures specify which variable is dependent and which is independent.  The independent variable is some variable that you think might help explain the variation in the dependent variable.  For example, if your two variables were education and voting you might choose education as the independent variable and voting as your dependent variable because you think that education will help you explain why some people vote Democrat and others vote Republican. Measures of association that specify which variable is dependent and which is independent are called asymmetric measures and measures that don’t specify which is dependent and which is independent are called symmetric measures.

## Part III – Choosing a Measure of Association

There are many measures of association to choose from. We’re going to limit our discussion to those measures that SDA will compute plus a couple others. When choosing a measure of association, we’ll start by considering the level of measurement of the two variables (see Exercise 6).

* If one or both of the variables is nominal, then choose one of these measures.
  + Contingency Coefficient – SDA doesn’t compute this but it’s easy to compute by hand.
  + Cramer’s V – SDA doesn’t compute this either but it’s also easy to compute by hand and we’ll show you how.
* If both of the variables are ordinal, then choose from this list.
  + Gamma
  + Somer’s d with the row variable as the dependent variable
  + Kendall’s tau-b
  + Kendall’s tau-c
* Dichotomies should be treated as ordinal. Most variables can be recoded into dichotomies. For example, marital status can be recoded into married or not married. Race can be recoded as white or non-white. All dichotomies should be considered ordinal.

## Part IV – Measures of Association for Nominal Variables

There are a few nominal level variables in the MTF survey.

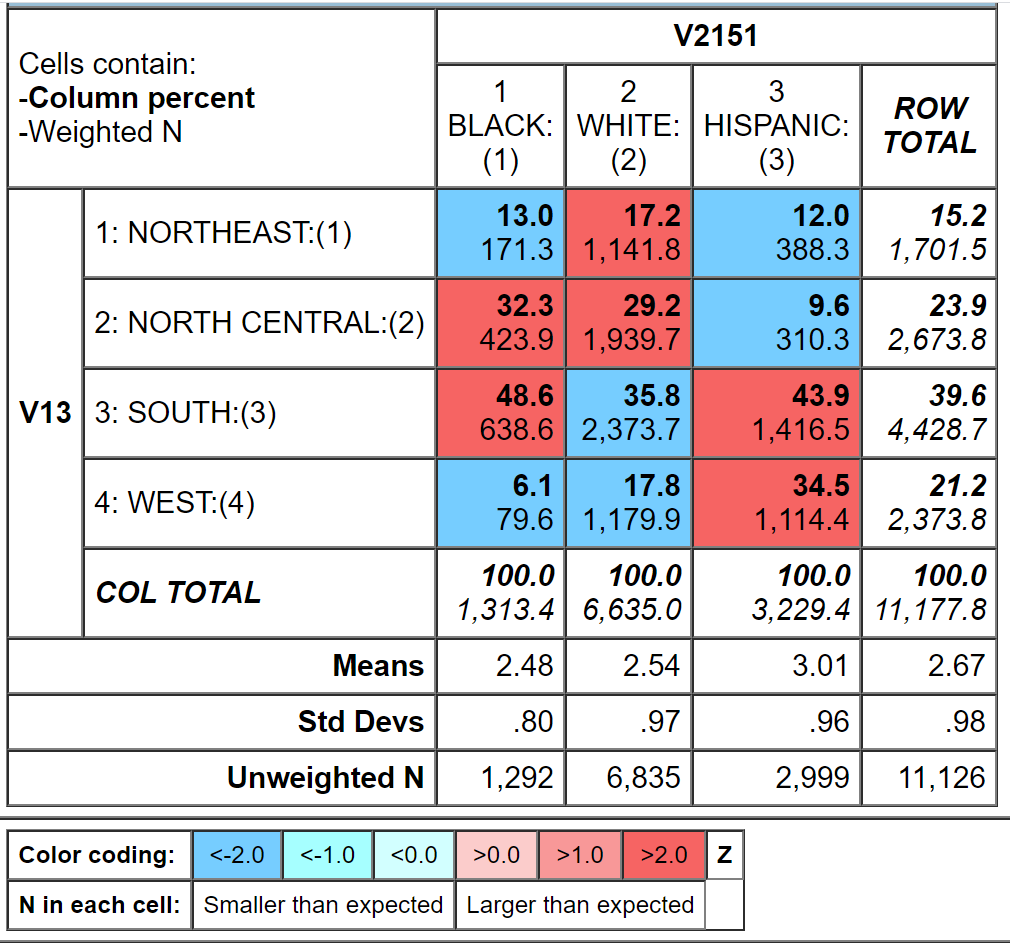
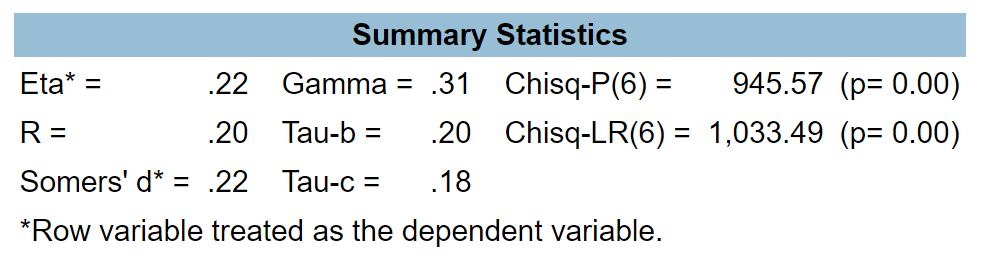
* *v13* – region of country where high school is located which we’ll refer to as the region where they currently live
* *v2151* – race
* *v2152* – type of community where respondent grew up

When one or both of your variables are nominal, you have a choice of the following measures[[37]](#footnote-37) – Contingency Coefficient and Cramer’s V.   Let’s start with the Contingency Coefficient (C).  One of the problems with this measure is that it varies from 0 to some value less than 1.  The larger the number of categories, the closer the maximum value is to 1.  For a table with two rows and two columns, the maximum value is .707 but for a table with three rows and three columns the maximum value is .816.  So, you can’t use C to compare the strength of the relationship unless the tables have the same number of rows and columns.

Cramer’s V is an extremely useful measure because it can vary between 0 and 1 regardless of the number of rows and columns.  Values of V can therefore be compared for tables with different number or rows and columns.[[38]](#footnote-38)

Let’s look at an example to help us better understand measures of association for nominal variables.  We’re going to use two variables – *v2151*and *v13*.  The first variable – *v2151* – is race and the second – *v13* – is the region of the country where the respondent currently lives.  It would make sense to think of region as the dependent variable since race might influence where they currently live.  Always put the dependent variable in the row and the independent variable in the column of your table.[[39]](#footnote-39)

Run CROSSTABS in SDA to produce the crosstabulation of *v2151* and *v13*.  Click on OUTPUT OPTIONS and look at PERCENTAGING.  Since your independent variable is always in the column, you want to use the column percents.  By default, the box for column percents is already checked. Also, check the box for SUMMARY STATISTICS.  You probably don’t want any of the charts so you could click on the drop-down arrow next to TYPE OF CHART and select NO CHART.  Click on RUN THE TABLE to produce the crosstabulation.  Your screen should look like the following.[[40]](#footnote-40)

Calculating C and V is easy. All you have to do is follow these simple steps.

* C equals the square root of the following: Chi Square divided by the sum of the number of cases in the table and Chi Square.
  + Chi Square is the Pearson Chi Square.  SDA expresses this as Chisq-P.  (See exercise 10RM)
  + Look at the SUMMARY STATISTICS that SDA gives you.  The Pearson Chi Square is 945.57 and the number of cases in the table is 11,177.8.
  + So, divide 945.67 by the sum of 945.67 and 11,177.8.  This equals 945.57 divided by 12,123.47 or 0.0780
  + Now take the square root of .0780 which equals 0.279.
* V equals the square root of the following: Chi Square divided by the product of the number of cases in the table and the smaller of two values – the number of rows minus 1 and the number of columns minus 1.
  + The Pearson Chi Square is 945.67 the number of cases in the table is 11,177.8, the number of rows minus 1 is 4-1 or 3, the number of columns minus 1 is 3 – 1 or 2.
  + The smaller of the number of rows minus 1 and the number of columns minus 1 is 2 since 3 -1 is smaller than 4 – 1.
  + So divide 945.67 by the product of 11,177.8 and 2. This equals 945.67 divided by 22,355.6 or .0423.
  + Now take the square root of .0423 which equals 0.206.

Notice that C and V are moderate in size.  C is 0.279 and V is 0.206.  You can see that C tells us that there is a moderate relationship between these two variables as does V.

## Part V – Now it’s Your Turn

Use CROSSTABS in SDA to give you the table for *v2150* and *v13*.  The variable *v2150* is the respondent’s sex.  We want to find out whether the respondent’s sex is related to where the respondent currently lives.  Decide which variable is independent and dependent.  Remember to put the dependent variable in the row and the independent variable in the column.  Get the correct percents and tell SDA to compute Chi Square. Then compute C and V by hand.  Use all this information to describe the relationship between these two variables.

## Part VI – Measures of Association for Ordinal Variables

There are a number of ordinal level variables in the MTF survey.  Here are a few examples.

* *v2173* – how respondents rates self on school ability compared to others of same age
* *v2174* – how respondent rates his or her intelligence compared to others of same age
* *v2179* – respondent’s report of his or her average grade in high school
* *v2183* – how likely the respondent thinks he or she will graduate from a four-year college.

You have a choice from four measures that SDA will compute for ordinal variables – Gamma, Somer’s d, Kendall’s tau-b, and Kendall’s tau-c.  Let’s start with Somer’s d.   This measure is the only one of the four that is an asymmetric measure.  That means that Somer’s d allows you to specify one of the variables as independent and the other as dependent.  Use CROSSTABS in SDA to get the crosstabulation of *v2173* and *v2183*.  If we think that the respondents’ evaluation of their school ability influences how likely it is that they think they will graduate from a four-year college, then graduation from college would be our dependent variable and would go in the row and evaluation of school ability would go in the column.  Be sure to get the column percents, Chi Square, and the four measures of association we listed above.

Chi Square tells us that we should reject the null hypothesis that the two variables are unrelated which provides support for our research hypothesis that the variables are related to each other.  Since *v2183* is our dependent variable, the appropriate value of Somer’s d is .27.  Tau-b and tau-c are very close to each other (0.30 and 0.27).  Gamma (0.44) is larger.  Gamma will always be larger because of the way it is computed.

You probably noticed that these measures for ordinal variables can be both positive and negative.  It’s often hard to interpret the sign.  We would like to be able to say that a positive value indicates that as one variable increases the other variable increases and a negative value indicates that as one variable increases the other variable decreases.  But that depends on how the values are coded.  So to determine whether a relationship is positive or negative it’s better to look at the percentages and let them tell you if it is positive or negative.  In our example, the relationship is positive since the higher respondents rate their school ability, the more likely they are to think they will graduate from college.

## Part VII – Now it’s Your Turn Again

Use CROSSTABS to give you a table for *v2179* and *v2183*.  We want to find out if the respondent’s grades in high school help us understand why some think they will graduate from college and others don’t.  Decide which variable is independent and dependent.  Get the correct percents and tell SDA to compute Chi Square and the four measures of association we discussed.  Use all this information to describe the relationship between these two variables.

## Part VIII – Using Measures of Association to Compare Tables

The primary use of measures of association is to compare the strength of a relationship in several tables.  You want to make sure that you compare the same measure of association across tables.  Compare Gamma values to Gamma values and V values to V values.  Rerun one of the tables that you created in parts 5 and 7 but this time hold sex constant.  Put *v2150* (sex) in the control box which is right below the COLUMN box in the crosstabs dialog box.  Now compare the appropriate measure of association to determine if the relationship is stronger for males or females or whether it doesn’t vary much by sex.  Remember not to make too much out of small differences in the measures.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 12 – Spuriousness

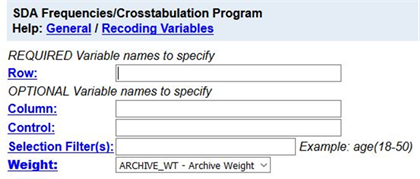
## Goals of Exercise

The goal of this exercise is to explore the concept of spuriousness.  We will consider the relationship between students’ high school grades and their expectation of graduating from a four-college in the future.  Then we will test for the possibility that this relationship is spurious.   The exercise also gives you practice in using CROSSTABS in SDA to explore relationships among variables and to test for spuriousness.

### Part I—Relationship of Grades in High School and Expectation of Graduating from Four-Year College

We’re going to use the Monitoring the Future (MTF) Survey of high school seniors for this exercise.  The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been an annual survey ever since.

Open the 2018 MTF survey of high school seniors. You’ll need to refer to Appendix A located at the end of this series of exercises which is a brief introduction to SDA and will tell all you need to know about using SDA. Your screen should look like the following.

​

Notice that a weight variable has already been entered in the WEIGHT box.  This will weight the data so the sample better represents the population from which the sample was selected

MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of about 14,500 seniors.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.

In previous exercises we looked at variables one at a time (i.e., univariate analysis) and at the relationship between two variables (i.e., bivariate analysis).  In this exercise we’re going to add a third variable into the analysis (i.e., multivariate analysis) and consider the possibility that our two-variable relationship might be spurious due to this third variable.  Spuriousness means that there is a statistical relationship between two variables, but it is not a causal relationship.  The statistical relationship is due to the third variable which we typically call the control variable.

To illustrate the idea of spuriousness, think about children in elementary, middle, and high school.  Every year children take standardized tests at the end of the school year to measure their achievement in areas such as mathematics, reading, and science.  Did you know that children with small feet score lower on these tests than children with big feet?  There is a relationship between children’s foot size and their test scores.  There is a clear statistical relationship between these two variables.  But is it a causal relationship?  Of course not!  No parent ever says I hope my kids have big feet so they will do better in school.  What we’re saying is that we think this relationship is spurious.  There is a statistical relationship but it’s not causal.

But why is this relationship spurious?  There must be some third variable that is creating this relationship.  One possibility is children’s grade level.  Children in lower grades have smaller feet and lower test scores.  Children in higher grades have bigger feet and higher test scores.  So, the relationship between foot size and test scores might be due to grade level.

How are we going to test this hypothesis?  What we do is to hold the third variable constant.  Let’s say that we have test scores for children in grades 6 through 12.  We’ll start with the sixth grade and look at the relationship of foot size and test scores for only the sixth graders.  Then we’ll repeat this for the seventh graders and for each successive grade level.  If the relationship is spurious, then we ought to find that the relationship between foot size and test scores goes away or is considerably reduced for **each** grade level. If the relationship is not spurious, then we ought to find that the relationship does not change much for the different grade levels.

Now let’s turn to an example from the MTF survey.  One of the questions in the survey asks students “How likely is it that you will graduate from a four-year college after high school?”  This is variable *v2183* in the data set.  The response categories are definitely won’t, probably won’t, probably will, and definitely will.  This will be our dependent variable.  In other words, we’re trying to explain why some students think they will graduate from college and others don’t.

It might be that students’ expectations about college are, in part, based on how they have done academically in high school.  Another question in the survey asks, “Which of the following best describes your average grade so far in high school?”  This is variable v2179 and will be our independent variable.  In other words, this is the variable that we think might explain why some students think they will graduate from college and others don’t.

Run CROSSTABS in SDA to see the relationship between students’ grades (v*2179*) and their expectation of graduating from a four-college in the future (*v2183*).[[41]](#footnote-41)  Make sure that you put the independent variable in the column and the dependent variable in the row.  Be sure to ask for the correct percents and the summary statistics.  Write a paragraph interpreting this relationship using the percents, Chi Square, and an appropriate measure of association.

Remember not to make too much out of small percent differences. The reason we don’t want to make too much out of small differences is because of sampling error.  No sample is a perfect representation of the population from which the sample was selected.  There is always some error present.  Small differences could just be sampling error.  So, we don’t want to make too much out of small differences.

## Part II—Adding a Third Variable into the Analysis

At this point we have only considered two variables (i.e., bivariate analysis).  We need to consider other variables that might be related to grades and expectations about college.  For example, sex may be related to both these variables.  Women may report higher grades and women may also be more likely to think they will graduate from college.  This raises the possibility that the relationship between grades and expectations for graduating from college might be due to sex.  In other words, it may be spurious due to sex.

Let’s check to make sure that sex is related to both our independent and dependent variables.  This is important because the relationship can **only** be spurious if the third variable (*v2150*) is related to both your independent and dependent variables.  Use CROSSTABS in SDA to get two tables – one table should cross-tabulate *v2150* and *v2179* and the other table should cross-tabulate *v2150* and *v2183*.  Be sure to get the SUMMARY STATISTICS so you will be able to use Chi Square and whichever measure of association you think is appropriate.  If sex is related to both variables, then we need to check further to see if the original relationship between grades and expectations for graduating from college is spurious as a result of sex.

## Part III—Checking for Spuriousness

How are we going to check on the possibility that the relationship between grades in high school and expectations about college is due to the effect of sex on the relationship?  What we can do is to separate males and females into two tables and look at the relationship between grades and expectations about college separately for men and for women. Sex is variable 2150 in our data set.

We can do that in SDA by getting a crosstab putting *v2183* in the ROW box (our dependent variable), *v2179* in the COLUMN box (our independent variable), and *v2150* in the CONTROL box.  In this case, *v2150* is the variable we are holding constant and is often called the control variable. You will get two tables – one for males and the other for females. Sometimes we call these partial tables since each partial table contains part of the sample.

We’re going to check to see what happens to the relationship between grades and expectations about college when we hold *sex* constant. If the original relationship is spurious then it either ought to go away or to decrease substantially for **both** males and females.  So, look carefully at the two tables – one for males and the other for females. But how can we tell if the relationship goes away or decreases markedly for both males and females?  One clue will be the percent differences between those who get high grades and those who get lower grades. Did the percent differences stay about the same or did they decrease substantially?  But there are so many percent differences that it’s hard to make sense of them. That’s where the summary statistics come in handy. Did the measures of association for males and females stay about the same or did they decrease substantially from that in the original two-variable table?

If the relationship had been due to sex, then the relationship between grades and expectations about college would have disappeared or decreased substantially for **both** males and females when we took out the effect of sex by holding it constant.  In other words, the relationship would be spurious.  Spurious means that there is a statistical relationship, but not a causal relationship.  It important to note that just because a relationship is not spurious due to sex doesn’t mean that it is not spurious at all.  It might be spurious due to some other variable.

The first thing we notice is that the Chi Square is significant for both males and females.  That tells us that there is probably a relationship between grades and expectations for graduating from college for **both** males and females.

Look at the pattern to the percents and the measures of association in the tables for males and females.  For both males and females, the higher the grades in high school, the more likely students are to feel they will graduate from college.  Additionally, the measures of association aren’t identical, but they aren’t very different.  Remember we don’t want to make too much of small differences because of sampling error.  So, in this analysis we would conclude that the relationship is not spurious due to sex.  But keep in mind that it might be spurious when we control for a different variable.

## Part IV – Now it’s Your Turn

Let’s stick with the same two-variable relationship – *v2179* and *v2183* – but this time let’s use a different control variable.  This time let’s use father’s education as our control variable.  Father’s education is coded into the following categories: 1=completed grade school or less, 2=some high school, 3=completed high school, 4=some college, 5=completed college, 6=graduate or professional school after college, and 7=don't know or does not apply. This is variable v2163 in our data set.

We’re going to make things simpler by recoding father’s education (*v2163*) into two categories – not college grad and college grad.  Recoding just means to combine categories of the variable.  Follow these steps to recode in SDA.

* Enter the variable name in the row box.  The variable name in this example is *v2163*.  (Don’t enter the period.)
* After the variable name, enter (r: where r stands for recode.
* Enter the new value you want to assign to the first recode followed by the values you want to combine.  In our case, we want to assign the new value 1 to the old values of 1 through 4.  So, this would be 1=1-4.  (Don’t enter the period.)
* We also want to assign a label to each category.  Enter the label in double quotation marks. For example, our recode for the first category would look like this – *v2163* (r:1=1-4"not college grad";.  (Don’t enter the period.)
* Separate the recodes by a semicolon.
* Repeat this process for each recode.  For the second category it would look this – 2=5-6"college grad".  (Don’t enter the period.)
* After the last recode, end the statement with a right parenthesis.
* This is what our recode statement would look like – v2163(r:1=1-4"not college grad";2=5-6"college grad").  (Don’t enter the period.)
* One more thing.  You’ll notice that we didn’t enter the value 7 (don’t know) into the recode.  That’s because we want to treat 7’s as missing data.  These respondents didn’t really answer the question.  They said they didn’t know.  Missing data are automatically excluded from the table.

Follow the same procedure that we used in parts 1, 2, and 3.  Interpret the tables and decide if the relationship is spurious or not.

## Part V—Conclusions

Summarize what you learned in this exercise.  What happened when you introduced *sex* into the analysis as a control variable?  What happened when you used father’s education as the control variable?  Were the original relationships spurious or not?  What does it mean to say a relationship is spurious?

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Exercise 13 – Writing Research Reports

## Goal of Exercise

The goal of this exercise is to discuss how we write research reports.  A research report describes what you discovered during your analysis of the data.

## Part I – An Outline of the Research Report

We used the Monitoring the Future (MTF) Survey of high school seniors for this series of exercises.  The MTF survey is a multistage cluster sample of all high school seniors in the United States.  The survey of seniors started in 1975 and has been done annually ever since.

MTF is an example of a social survey.  The investigators selected a sample from the population of all high school seniors in the United States.  This particular survey was conducted in 2018 and is a relatively large sample of about 14,500 seniors.  In a survey we ask respondents questions and use their answers as data for our analysis.  The answers to these questions are used as measures of various concepts.  In the language of survey research these measures are typically referred to as variables.

In previous exercises we started by looking at variables one at a time (i.e., univariate analysis) and then at the relationship between two variables (i.e., bivariate analysis).  Eventually we added a third variable into the analysis (i.e., multivariate analysis) and considered the possibility that our two-variable relationship might be spurious due to this third variable.

Now we need to think about how to write a research report so that others may read it and learn from our analysis.  This report might be for a class you are taking, or it might be a report that you are submitting to a research conference.  If you are going to submit your report to a journal for possible publication, you need to look carefully at the instructions that all journals provide on preparing a manuscript for publication.

Here's an outline for your report.  Don't think that this is the only way you can organize your report, but this is one way to do it.

* Title page including your name, the date, and the class it is for.
* Abstract – An abstract is a short summary of what you did in the paper and the major findings of your analysis.  Abstracts are really short, so you need to be succinct. It should be less than 200 words or even shorter depending on the requirements of your professor or the research conference to which you are submitting your paper.
* Table of contents (optional).
* Body of the paper.
  + An introduction to the paper which explains why you wrote the report and provides an introduction to the topic of the paper.
  + Your review of the literature that summarizes what others discovered about this topic.  Virtually everything you might do has been written about by others.  You should review the relevant literature and summarize what others have found.  You don't want to simply list the relevant literature and consider the articles and books one by one.  Rather you want to summarize what others have done and look for themes around which you can organize your literature review.  If you are having trouble finding relevant literature, go to the library at your university or a nearby university and talk with a reference librarian.  They are trained in searching for relevant literature and will be able to help you.
  + The methodology of your study.
    - If you collected your own data, discuss how you chose your sample, how you measured the concepts, and how you collected your data.
    - If you used an existing data set, discuss the sampling, measurement, and data collection used in that study.  Studies that are part of data archives such as the Inter-university for Political and Social Research at the University of Michigan and the Roper Center for Public Opinion Research at Cornell University provide good summaries for all data sets that are housed at their archive.
  + Theory and Hypotheses – If you are using a theoretical perspective and/or testing hypotheses, describe the theory and state the hypotheses you plan to test.  Be sure to cite supporting literature that forms the basis for your theory and hypotheses.
  + Empirical findings and interpretation – What are the empirical findings that came out of your data analysis and what did they tell you?  If you are testing hypotheses, did your analysis support your hypotheses?  Remember that you are telling a story.  Start simple and build up.  That means starting with univariate analysis, then proceeding to bivariate analysis, and then looking at sets of three or more variables (i.e., multivariate analysis) to consider such things as spuriousness.
  + Conclusions and summary. This is a little like your abstract but not as short.  What did you do, what did you find in your study and what did it mean?
* Tables.  You may choose to put your tables in the body of your paper, or you may decide to put them all at the end of your paper.
* References.  For every article or book that you cite, you need to provide a full bibliographic reference at the end of the report.

## Part II—Tables

There are advantages and disadvantages to putting your tables in the body of the report or at the end of the report.  Putting them in the body of the report keeps them front and center for the reader but they often are bulky and get in the way of reading your report.  Putting them at the end of the report gets them out of the way and allows the reader to spread them out and look at them as he or she is reading the paper.  Your instructor or the research conference will usually tell you where to put your tables.

If they are placed at the end of the paper, put a note in the body of the report that says something like "Table 1 about here."  That will let the reader know where the table fits into your report.

Constructing a good table is important.  Sometimes your instructor will tell you to copy tables from the program you are using for statistical analysis (e.g., SDA, SPSS, PSPP, and others) into your paper.  Other times you will construct the tables yourself.  A good reference on creating tables is *The Chicago Guide to Writing About Numbers* by Jane E. Miller.[[42]](#footnote-42)  Your word processing program (e.g., Word in Microsoft Office) will provide you with templates that you can choose for your tables.

## Part III – Footnotes or Endnotes?

Often you want the reader to be aware of something but you don't want to put it in the body of the paper.  It may be a technical issue such as how you recoded a variable or why you chose a particular statistic.  Or you may want to tell the reader that you will discuss something later in the paper.  You can put comments like these in either a footnote or an endnote.  A footnote goes at the bottom of the page and an endnote goes at the end of the paper.  Your word processing program will allow you to enter either footnotes or endnotes in your paper.  Which you use is up to you unless your instructor or the research conference tells you that one or the other is required.

## Part IV – Citing Articles, Papers, and Other Materials

There are many styles (e.g., American Psychological Association (APA) or Modern Language Association (MLA)) that you could use to cite materials that you refer to in your paper.  Remember that anytime you refer to someone else's work, you must acknowledge the source.  Your instructor or research conference will often specify which style you should use.

## Part V – Plagiarism

Plagiarism is using someone else's words or ideas without acknowledging the source.  If you are quoting from a document, you must cite the source.  Even if you are paraphrasing, you must acknowledge the source.  If you are using someone else's ideas, you must acknowledge the source.  There is a good review of plagiarism written by Earl Babbie that can be found on the Internet by clicking [here](http://www1.chapman.edu/~babbie/plag00.html).  Click on the red arrow at the top to go forward or backward in this review of plagiarism.

## Part VI– Proofreading

Be sure to proofread your paper several times before turning it in.  Use the spell and grammar checker in your word processor.  Another possibility is asking a friend to read it and tell you about errors or parts that are confusing.

## Part VII – Other Guides to Writing

There are many other guides to writing research reports.  One that is commonly used in Sociology is the Guide to Writing Sociology Papers.[[43]](#footnote-43)  You can find others on the internet by entering "writing research reports" in the search box.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Appendix A – Introduction to SDA

## Part I – Opening the Data Set

In these exercises we’re going to use the Monitoring the Future (MTF) Surveys of high school seniors in the United States that have 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.”

Information about these surveys 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.  All accounts are free.

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 MTF 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 37416 in the FIND DATA box and you will see the 2018 12th grade survey which is what we are using in these exercises. Click on the link to the survey which will take you to the home page for this survey.  Scroll down a little and you will see two links - one for downloading the data and the other for analyzing the data online.  Click on ANALYZE ONLINE and then on FULL ANALYSIS CAPABILITIES. That will take you to the home page for the National Addiction and HIV Data Analysis Program (NAHDAP) which is where the Monitoring the Future surveys are archived. Log into NAHDAP using your username and password for the ICPSR account you just created. Your username will be the email address that you used when you created your ICPSR account. You should see the TERMS OF USE which is a legal agreement you need to agree to in order to use the data. Read it and then click on I AGREE.

Survey Documentation and Analysis (SDA) was developed by the Survey Methods Program at UC Berkeley and is currently maintained by the Institute for Scientific Analysis. It’s an online statistical package that is freely available wherever you have internet access. SDA is easy to use and it doesn’t take very long to get up and running on SDA. It has a very good context-sensitive help menu that is easily accessible. Once you agree to the TERMS OF USE, you should see the following.

## This shows the SDA dialog box for frequencies and crosstabulation Part II – SDA Dialog Box

The image above shows you the SDA dialog box which is your interface to SDA.  Let’s spend a little time exploring the dialog box.

* In the very top line is the name of the data set you opened.  Always check to make sure you opened the correct data set.  It should say that you have opened the “Monitoring the Future: A Continuing Study of American Youth (12th-Grade Survey), 2018 Core Data.”
* Below that you will see five buttons - Analysis, Create Variables, Download, Codebook, and Getting Started.
* Getting started has more information than you will need. This introduction to SDA should provide enough information.
* Codebook has information about the variables in this data set and documentation for the survey.
* On the left you should see a list of all the variables in the core data set.  The variables are grouped in the following categories.
  + Survey/Dataset information which includes the weight variable
  + Geographic which provides information about where they live
  + Demographic variables describing both the respondent and the respondent’s family
  + Attitudes and beliefs about politics, religion, and school
  + Educational and military aspirations
  + Employment and income
  + Recreational information
  + Driving and substance abuse
  + Substance use
  + Substance use variables recoded into dichotomies

## Part III – Carrying Out the Statistical Analysis

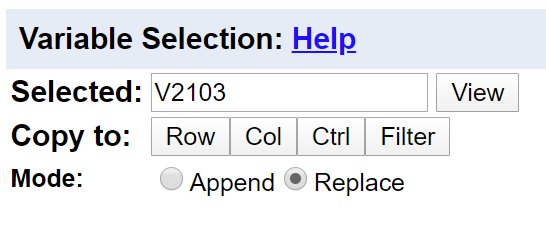
To see the statistical procedures you can use to analyze the data, click on ANALYSIS at the top of the page.  We’re only using frequencies and crosstabulation in these exercises which is the default (see image above).  We’re going to use frequencies as an example in this introduction.

Notice that there is a mini codebook on the left-side of the screen.  What you see is the categories of variables in the core data set.  Click on the + sign to the left of one of the categories, for example, SUBSTANCE USE.  This will expand the codebook and show you all the subcategories of this category.  You can also double-click on the category to expand it.  Your screen should look like the following.

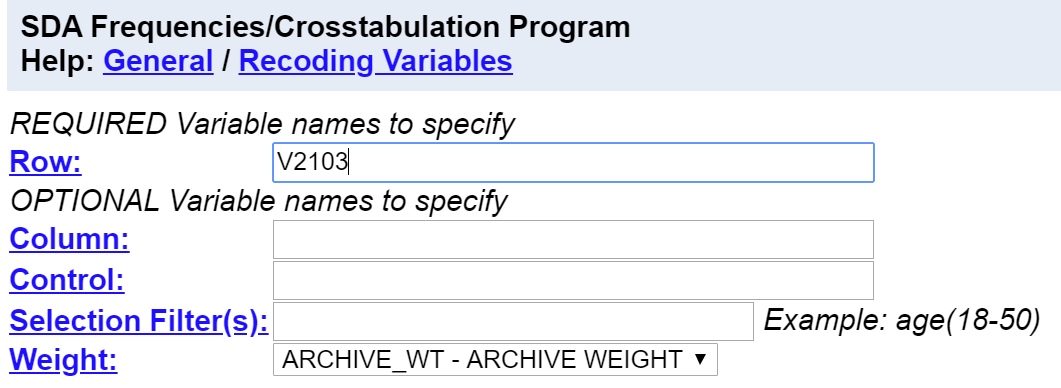
The second subcategory is ALCOHOL.  Click on the + sign to the left of ALCOHOL.  Your screen should look like the following.



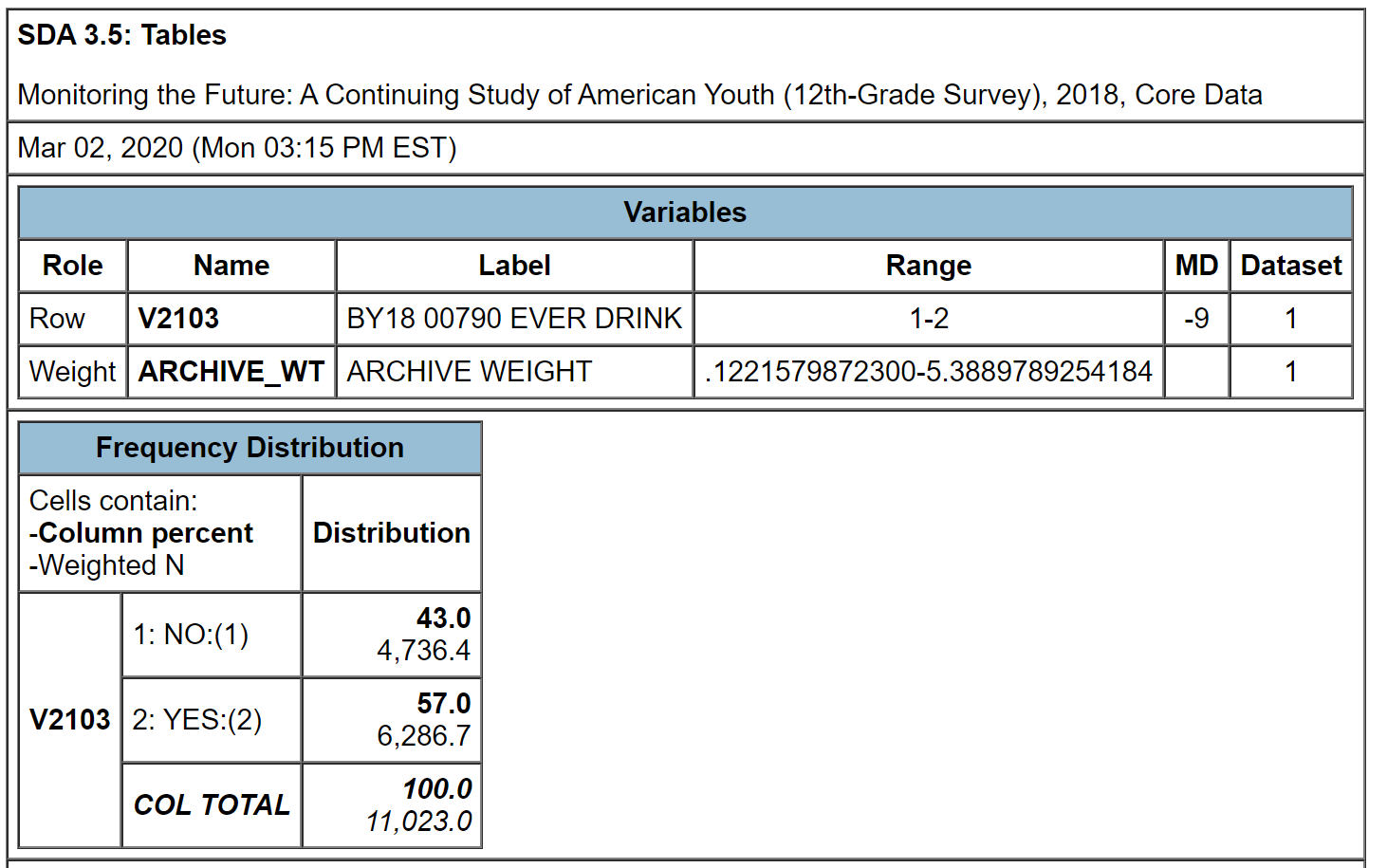
The variable names are on the far left, for example, V2103.  To the right of the variable name is the variable label.  Click on V2103.  The variable now appears in the VARIABLE SELECTION box at the top of the screen.  Your screen should look like the following.

  
Click on the VIEW button to the right of the variable name and you will see the wording of question, the frequency distribution, and the values that have been designated as missing values.  Then close the VIEW tab.

Now click on ROW to the right of COPY TO and the variable name will be entered in the ROW box.  You could also have typed V2103 in the box.  Either way works. Your screen should look like the following.



Click on RUN THE TABLE at the bottom of the page to get the frequency distribution for V2103.  Your screen should look like the following.



The table shows that 57% of high school seniors said they had an alcoholic drink while 43% said they had not.

You can copy and paste the tables from the SDA output into your paper.  Or you could create a screen capture using an application like Snipping Tool.

Instructions for further use of SDA will be included in the exercises of this module.

# Exercises for an Introductory Research Methods Class Using the 2018 Monitoring the Future Survey of High School Seniors and SDA Edward Nelson, California State University, Fresno Appendix B – Notes to Instructors

Survey Documentation and Analysis (SDA) was developed by the Survey Methods Program at UC Berkeley and is currently maintained by the Institute for Scientific Analysis. It’s an online statistical package that is freely available wherever you have internet access. SDA is easy to use and it doesn’t take very long to get up and running on SDA. It has a very good context-sensitive help menu that is easily accessible.

One of the limitations of SDA is that it’s not easy to develop data sets in SDA format. While SDA is freely available to use, it requires a site license to create your own SDA data sets. Developing SDA data sets is labor intensive and has a fairly steep learning curve. That means that one normally relies on data sets that others have developed. Fortunately, there are many high-quality SDA data sets available on the internet. Here are some places to go to find SDA data sets.

* [SDA Archive](http://sda.berkeley.edu/archive.htm) located at UC Berkeley including:
  + General Social Survey cumulative data file from 1972 through 2018
  + American National Election Studies from 1992 through 2016 and the ANES cumulative data file from 1952 through 2016
  + Census microdata
    - U.S. from 1990 through 2003
    - California for 1990 and 2000
  + Field Poll data from 1956 through 2006 .
* Inter-University Consortium for Political and Social Research (ICPSR) including:
  + many data set located in the ICPSR [general archive](http://www.icpsr.umich.edu)
  + [data-driven learning guides](http://www.icpsr.umich.edu/icpsrweb/instructors/biblio/resources) which are instructional exercises that can be used in the classroom
  + Investigating Community and Social Capital, an [instructional module](http://www.icpsr.umich.edu/icpsrweb/instructors/icsc/index.jsp) using SDA
  + Voting Behavior: the 2012 Election, another [instructional module](http://www.icpsr.umich.edu/icpsrweb/instructors/setups2012/index.jsp) using SDA
  + Voting Behavior: the 2016 Election, still another [instructional module](https://www.icpsr.umich.edu/icpsrweb/instructors/setups2016/) using SDA

The Monitoring the Future survey that we will be using in these exercises is freely available at the Inter-University for Political and Social Research (ICPSR). Users do not have to be at a university or college that belongs to the ICSPR. However, they will need an account at the ICPSR which are free. Appendix A describes how students should get an account and how to access the data.

.

It’s important to weight the data so the sample better represents the population from which the sample is selected. SDA automatically enters the weight variable in the WEIGHTS box.

There are a couple of things that need to be mentioned about SDA. SDA doesn’t have a command to run a t test. The t test is a special case of one-way analysis of variance when the independent variable is a dichotomy.

These exercises use the SDA 4.0 interface which is an update from the earlier 3.5 interface. The differences between these two interfaces is explained on the SDA [website](https://sda.berkeley.edu/archive.htm).

You have permission to use these exercises and to revise them to fit your needs. Feel free to revise the exercises in any way you want. Just recognize the source of the original exercises. I would be interested in hearing about your experiences when you use these exercises.

If you would like to contact me, please email me at [ednelson@csufresno.edu](mailto:ednelson@csufresno.edu). I’m Professor Emeritus at California State University, Fresno in the Sociology department. I taught research methods, statistics, and critical thinking before retiring and now teach a critical thinking course part time.

1. Thomas J. Sullivan.  1992. Applied Sociology. New York: Macmillan Publishing Company (p. 10). [↑](#footnote-ref-1)
2. These are not the only things that social scientists are interested in.  For example, we might want to answer questions that deal with businesses or nations.  We’re going to focus in these exercises on individuals and their behavior, attitudes, and opinions.  In other words, our unit of analysis will be the individual. [↑](#footnote-ref-2)
3. Just a reminder that 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.  [↑](#footnote-ref-3)
4. This is the question used in the Pew Research Center’s Political Polarization Survey conducted in 2014.  More information on this survey can be found on their [website](http://www.pewresearch.org/fact-tank/2014/06/12/7-things-to-know-about-polarization-in-america/).  [↑](#footnote-ref-4)
5. This is also from the 2014 Pew Political Polarization Survey referred to in footnote 4. [↑](#footnote-ref-5)
6. This question is from the 2014 General Social Survey conducted by the National Opinion Research Center.  More information about this survey can be found on their [website](https://gssdataexplorer.norc.org/).  [↑](#footnote-ref-6)
7. This question is also from the 2014 General Social Survey referred to in footnote 6.  [↑](#footnote-ref-7)
8. Matilda White Riley. Sociological Research I: A Case Approach. 1963. New York: Harcourt, Brace & World (pp. 184-190). [↑](#footnote-ref-8)
9. For more on survey research, see the following. Ernest L. Cowles and Edward Nelson. An Introduction to Survey Research, Volume I – The Basics of Survey Research and Volume II – Carrying Out the Survey. 2019. Business Expert Press. [↑](#footnote-ref-9)
10. 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.  [↑](#footnote-ref-10)
11. 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.  [↑](#footnote-ref-11)
12. 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.  [↑](#footnote-ref-12)
13. You can determine from the exercise data base that 60% of females favor same-sex marriage compared to only 40% of males. [↑](#footnote-ref-13)
14. This question is from the 2014 Pew Political Polarization survey.  More information can be found on their [website](http://www.pewresearch.org/fact-tank/2014/06/12/7-things-to-know-about-polarization-in-america/). [↑](#footnote-ref-14)
15. These questions are also from the 2014 Pew Political Polarization Survey referred to in footnote 13  They actually used a series of 10 questions.  I only reported four of them here. [↑](#footnote-ref-15)
16. See the Pew Research Center’s [report](https://www.pewresearch.org/fact-tank/2018/05/21/u-s-voter-turnout-trails-most-developed-countries/?utm_source=Pew+Research+Center&utm_campaign=95ba07b7a9-Weekly_August_4_20168_3_2016&utm_medium=email&utm_term=0_3e953b9b70-95ba07b7a9-399416193) on voting turnout. [↑](#footnote-ref-16)
17. A good introduction to validity can be found in Earl Babbie’s The Practice of Social Research. Wadsworth Cengage Learning, 2013, pp. 151-153.  Babbie discusses several different dimensions of validity – face validity, criterion-related validity, construct validity, and content validity.  We omitted construct validity in our brief discussion in this exercise.   [↑](#footnote-ref-17)
18. Binge drinking is often defined as having five or more drinks in a row. [↑](#footnote-ref-18)
19. A simple random sample is one in which every member of the population (i.e., all 3,000 seniors) has the same chance of being selected in the sample. [↑](#footnote-ref-19)
20. Some students might be in both samples but we would expect the degree of overlap to be small. [↑](#footnote-ref-20)
21. It’s easy to imagine that there could be some clerical error in our list which would create some coverage error but we would expect that error to be very small. [↑](#footnote-ref-21)
22. For the purposes of this exercise ignore the confidentiality issue which would be a serious concern. [↑](#footnote-ref-22)
23. Schuldt, J.P; S.H. Konrath, and N. Schwarz.  2011. “’Global Warming’ or ‘Climate Change’? Whether the Planet is Warming Depends on Question Wording.”  *Public Opinion Quarterly* 75, no. 1, p. 115. [↑](#footnote-ref-23)
24. L. Whitmarsh. 2009. “What’s in a Name” Commonalities and Differences in Public Understanding of ‘Climate Change’ and ‘Global Warming.’” *Public Understanding of Science* 18, no. 4, p. 410. [↑](#footnote-ref-24)
25. Moore, D.W. 2002. “Measuring New Types of Question-Order Effects: Additive and Subtractive.” *Public Opinion Quarterly* 66, no. 1, p. 83. [↑](#footnote-ref-25)
26. In these exercises we’re going to focus on survey research.  [↑](#footnote-ref-26)
27. In these exercises we’re going to focus on survey research.  [↑](#footnote-ref-27)
28. Stanley Smith Stevens, 1946, “On the Theory of Scales of Measurement,” Science 103 (2684), pp. 677-680. [↑](#footnote-ref-28)
29. We need to clear something up.  Why do the weighted data have decimal values?  When you weight the cases by the weight variable, you will get a fractional number of cases.  Don’t worry about this.  It’s a technical issue and not important to us in this discussion.  [↑](#footnote-ref-29)
30. Frequency distributions can be grouped or ungrouped.  Think of age.  We could have a distribution that lists all the ages in years of the respondents to our survey.  We could also divide age into a series of categories such as under 30, 30 to 39, 40 to 49, 50 to 59, 60 to 69, and 70 and older.  In a grouped frequency distribution the mode would be the most common category or categories.  [↑](#footnote-ref-30)
31. In a grouped frequency distribution the median would be the category that contains the middle value. [↑](#footnote-ref-31)
32. We need to clear something up.  Why do the weighted data have decimal values?  When you weight the cases by the weight variable, you will get a fractional number of cases.  Don’t worry about this.  It’s a technical issue and not important to us in this discussion.  [↑](#footnote-ref-32)
33. We need to clear something up.  Why do the weighted data have decimal values?  When you weight the cases by the weight variable, you will get a fractional number of cases.  Don’t worry about this.  It’s a technical issue and not important to us in this discussion.  [↑](#footnote-ref-33)
34. We need to clear something up.  Why do the weighted data have decimal values?  When you weight the cases by the weight variable, you will get a fractional number of cases.  Don’t worry about this.  It’s a technical issue and not important to us in this discussion.  [↑](#footnote-ref-34)
35. The null hypothesis is often called the hypothesis of no difference.  We’re saying that there is no relationship between these two variables.  In other words, there’s nothing there. [↑](#footnote-ref-35)
36. See Exercise 6 for a discussion of levels of measurement.  Nominal variables have no underlying order and ordinal variables have an underlying order.  Measures of association for nominal variables range from 0 to 1 while measures for ordinal variables range from -1 to +1. [↑](#footnote-ref-36)
37. There’s another popular measure called Lambda but SDA doesn’t compute it and it’s harder to compute by hand so we’re going to skip it. [↑](#footnote-ref-37)
38. If your table has two columns and two rows, V is equal to Phi which you might be familiar with.  Since Phi is a special case of V, we’re not going to discuss it. [↑](#footnote-ref-38)
39. That’s important because as we noted earlier some measures are asymmetric which means that it depends on which of the two variables is dependent and which is independent.  In these cases, SDA assumes that the row variable is the dependent variable.  [↑](#footnote-ref-39)
40. We need to clear something up.  Why do the weighted data have decimal values?  When you weight the cases by the weight variable, you will get a fractional number of cases.  Don’t worry about this.  It’s a technical issue and not important to us in this discussion.  [↑](#footnote-ref-40)
41. We need to clear something up.  Why do the weighted data have decimal values?  When you weight the cases by the weight variable, you will get a fractional number of cases.  Don’t worry about this.  It’s a technical issue and not important to us in this discussion.  [↑](#footnote-ref-41)
42. Jane E. Miller. 2015.  The Chicago Guide to Writing About Numbers.  Chicago: University of Chicago Press. [↑](#footnote-ref-42)
43. Sociology Writing Group.  A Guide to Writing Sociology Papers. 2013 (7th edition). Worth Publishers. [↑](#footnote-ref-43)