

# Passion Driven Statistics



A Supportive, Multidisciplinary,  
Project-Based, Introductory Course

# Table of Contents

<b>Foreword:</b> .....	ii
<b>Chapter 01:</b> Introduction .....	3
<b>Chapter 02:</b> Data Sets & Code Books .....	4
<b>Chapter 03:</b> Literature Review .....	6
<b>Chapter 04:</b> Working with Data .....	8
<b>Chapter 05:</b> Data Management .....	10
<b>Chapter 06:</b> Univariate Graphing .....	16
<b>Chapter 07:</b> Bivariate Graphing .....	18
<b>Chapter 08:</b> Hypothesis Testing .....	22
<b>Chapter 09:</b> Analysis of Variance .....	24
<b>Chapter 10:</b> Chi-Square .....	26
<b>Chapter 11:</b> Correlation .....	28
<b>Chapter 12:</b> Moderation .....	30
<b>Chapter 13:</b> The Question of Causation .....	33
<b>Chapter 14:</b> Multivariate Models and Confounding .....	34
<b>Chapter 15:</b> Writing for Poster Presentations .....	36
<b>Acknowledgements:</b> .....	xxxvii

# Foreword

Statistical data analysis is central to research in the natural and social sciences, technology, engineering and mathematics (STEM). It is essential for scientific findings to be regarded as meaningful. In many ways, statistics is the shared language of the sciences. A solid understanding of how statistical methods are employed and reported is essential for communication both within and across disciplines.

This **Passion-Driven Statistics e-book** includes a step-by-step curriculum for answering questions with data. It is designed for students at all levels of experience, including complete beginners. You will use existing data to pose research questions, and you will learn a statistical software platform to answer those questions by turning raw data into useful information.

Passion-Driven Statistics “flips” the traditional course format. Rather than attending class sessions in order to listen to lectures, you will engage with online lectures as homework and use class time to work on a research project of your own design.

The e-book is broken into 15 chapters, each of which will provide a brief overview of what you have learned, a summary of the upcoming lesson, links to the video lessons, and a corresponding project assignment.

For each chapter, you should:

1. Read the overview and lesson summary.
2. Use the appropriate link to watch the video lesson for the software that your instructor has chosen.
3. Use the model code/instructions to complete the assignment.

Though model code is provided as part of several of the chapters, the code is not meant to reflect proper formatting. You can copy/paste relevant code into your own software program and make needed changes based on the variables you have selected and the formatting rules that your instructor will provide. Note that text listed in **bold** reflects the parts of the code that are required, while non-bolded text is meant to be changed to reflect the decisions that you have made.

If you are using SPSS, lessons include both a video and a print-screen tutorial. Also, Python video lessons are segmented. After watching each segment, select the next segment on the right-hand side of the YouTube screen (e.g. 1/4, 2/4, 3/4, 4/4).

We recommend keeping a journal or lab notebook to track the progress of your project. You will find this to be a valuable research tool that saves time and unnecessary work.



# Introduction

---

“Statistics is the grammar of science.”

- Karl Pearson

## Overview

This course is presented in the service of a project of your choosing and will offer you an intensive hands-on experience in the quantitative research process. You will develop skills in generating testable hypotheses, understanding large data sets, formatting and managing data, conducting descriptive and inferential statistical analyses, and presenting results for expert and novice audiences. It is designed for those who are interested in developing skills that are useful for working with data and using statistical tools to analyze them. No prior experience with data or statistics is required.

Our approach is “statistics in the service of questions”. As such, the research question that you choose (from data sets made available to you) is of paramount importance to your learning experience. It must interest you enough that you will be willing to spend many hours working to answer it.

## Lesson

In this lesson, you will learn about the resources available to you. This e-textbook will give you a brief introduction for each topic, a link to video lessons, example code (or point-and-click instructions) for the statistical platform you will be using, and basic instructions for project-based assignments. Your instructor will provide additional resources to help you. Click **HERE** to watch the introductory video.



# Data Sets and Code Books

---

“Numbers have life. They are not just symbols on paper.”

- Shakuntala Devi

## Overview

You have learned a bit about the structure and goals of the course and have heard from students who have engaged in their own research. Now you will begin to learn about the necessary steps in statistical inquiry, types of variables, and how to use code books to develop your own research question. Statistics is all about converting data into useful information.

## Lesson

Learn about the role of **populations** and **samples** in statistical inquiry. Understand the importance of choosing a representative sample from a population of interest. Consider the structure of a data set and the meaning and function of information provided within its rows and columns. Learn about how code books document the ways that the data are arranged within data sets, what the various numbers and letters mean, and any special instructions on how to use the data. Develop skills in reading code books and developing a research question. Click **HERE** for watch the video lesson.

## Choosing a Research Question:

Now it is time for you to formulate your own research question. You should spend ample time on this task as it will guide your work for the rest of the course.

---

**One of the simplest research questions that can be asked is whether two topics or constructs are associated.**

Is medical treatment seeking associated with income?

Is water fluorination associated with number of cavities during dentist visits?

Is humidity associated with caterpillar reproduction?

---

First, carefully read through the available code book(s). If more than one data set has been made available to you, select the one that you would like to work with.

Next, identify a specific topic of interest and begin to prepare a codebook of your own. Print individual pages from the larger codebook (i.e. items or questions that measure your selected topic).

During an additional review of the codebook, identify a second topic that you would like to explore in terms of its association with your original topic. Add the items or questions documenting this second topic to your personal codebook.

For example, after looking through the codebook for the NESARC study, I have decided that I am particularly interested in nicotine dependence. I am not sure which items I will use regarding nicotine dependence (e.g. symptoms or a diagnosis) so for now I will include all of the relevant items in my personal codebook.

While nicotine dependence is a good starting point, I need to determine what it is about nicotine that I am interested in. It strikes me that friends and acquaintances that I have known through the years that became hooked on cigarettes did so across very different periods of time. Some seemed to be dependent soon after their first few experiences with smoking and others after many years of generally irregular smoking behavior.

**I decide that I am most interested in exploring the association between level of smoking and nicotine dependence.**

I add to my codebook items reflecting smoking (e.g. smoking quantity and frequency).

Bring a **printed copy** of this personal code book to each class session.



# Literature Review

---

“Research is to see what everybody else has seen, and to think what nobody else has ever thought.”

- Albert Szent-Gyorgyi

## Overview

You have learned that a data set contains observations (i.e. the rows) and variables (i.e. the columns) and that to understand a data set you will first need to read through its code book. You are also now able to distinguish between categorical and quantitative variables. Next, we are going to focus on conducting a review of primary source journal articles in order to understand previous work conducted on your chosen research question.

To conduct a literature review, you will find and read published, primary source, peer-reviewed articles that have addressed your research question or questions that are similar to yours. This will allow you to better understand what is already known about the association or relationship that you will be studying.

## Lesson

Learn about the literature review process. Compare strengths and weaknesses of primary and secondary source scientific literature. Experience how the context and background information you find in your literature search can inform your research. Consider ways you may want to refine your topic or modify your question. Click **HERE** to watch the video lesson.

## Conducting a Literature Review:

Describe the association that you will examine and key words you found helpful during your literature search. List at least 5 of the most appropriate references that you have found and read. Describe findings and interesting themes that you have uncovered and include the specific research question that you hope to answer.

---

For example, given the association that I have decided to examine, I use such keywords as **nicotine dependence**, **tobacco dependence**, and **smoking**. After reading through several titles and abstracts, I notice that there has been relatively little attention in the research literature to the association between smoking exposure and dependence symptoms. I also expand a bit to include other substance use and dependence relationships that may provide relevant background.

### Example References:

Dierker, L., & Donny, E. (2008). The role of psychiatric disorders in the relationship between cigarette smoking and DSM-IV nicotine dependence among young adults. *Nicotine & Tobacco Research, 10*(3), 439-446. doi:10.1080/14622200801901898

Dierker, L., Hedeker, D., Rose, J., Selya, A., & Mermelstein, R. (2015). Early emerging nicotine dependence symptoms in adolescence predict daily smoking in young adulthood. *Drug and Alcohol Dependence, 151*, 267-271. doi:10.1016/j.drugalcdep.2015.03.009

Dierker, L., Donny, E., Tiffany, S., Colby, S. M., Perrine, N., Clayton, R. R., & Network, T. (2007). The association between cigarette smoking and DSM-IV nicotine dependence among first year college students. *Drug and Alcohol Dependence, 86*(2-3), 106-114. doi:10.1016/j.drugalcdep.2006.05.025

Selya, A. S., Dierker, L. C., Rose, J. S., Hedeker, D., Tan, X. M., Li, R. Z., & Mermelstein, R. J. (2013). Time-varying effects of smoking quantity and nicotine dependence on adolescent smoking regularity. *Drug and Alcohol Dependence, 128*(3), 230-237. doi:10.1016/j.drugalcdep.2012.08.026

Based on my reading of the above articles as well as others, I have noted a few common and interesting themes:

First, while it is true that smoking exposure is a necessary requirement for nicotine dependence, frequency and quantity of smoking are markedly imperfect indices for determining an individual's probability of exhibiting nicotine dependence symptoms (and the same seems to be true for other substances that cause dependence symptoms).

The association may differ based on ethnicity, age, and gender, although there is relatively little work on this.

One of the most potent risk factors consistently implicated in the etiology of smoking behavior and nicotine dependence is depression.

I have decided to further focus my question (is level of smoking associated with the likelihood of experiencing nicotine dependence?) by examining whether this association differs based on whether or not an individual has depression. I am wondering if at low levels of smoking nicotine dependence is more common among individuals with major depression than those without major depression.

I add relevant depression questions/items/variables to my personal codebook as well as several demographic variables (age, gender, ethnicity, etc.) and any other variables I may wish to consider.

---





# Working with Data

---

**“When you have mastered numbers, you will no longer be reading numbers, any more than you read words when reading books. You will be reading meanings.”**

**- W.E.B. Du Bois**

## Overview

You have learned about developing a research question based on existing data by using a code book to guide you. You refined your research question by conducting a literature review based on primary source journal articles. Now you are ready to use statistical software to work with the data.

**Exploratory data analysis** is the processes of converting raw data into a more useful form so that we can begin to discover important features and patterns in the data.

## Lesson

Learn to examine frequency distributions for each of the variables you have selected. Determine what values a variable takes and how often it takes those values. Write the code or take the steps required to generate frequency distributions using your statistical software program. As you engage with your data, learn how to consider whether or not you want to create a subset of the larger sample in order to answer your question. Click on a video lesson below based on the software that you will be using.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

## Calling in (importing) a data set

SAS      **libname** mydata "C:\folder name";  
          **data new; set** mydata.filename;

R        **load** ("filename-including-path.Rdata")  
          myData\_orig <- name-of-object-loaded-in-your-workspace

PYTHON    **import pandas**  
          **import numpy**  
          myData = **pandas.read\_csv**('nesarc\_pds.csv')

STATA     **use** "C:\folder name\file name", **clear**

SPSS      **GET FILE=**'C:\folder name\filename.sav'.

---

## Sorting data

SAS        **proc sort;** by unique\_id;

R        myData <- myData[**order**(myData\$unique\_id, **decreasing = FALSE**),]

PYTHON    myData = myData.**sort\_values**(by='unique\_id')

STATA     **sort** unique\_id

SPSS      **SORT CASES BY** unique\_id.

---

## Writing your First Program:

Submit your program and the corresponding results that display at least three of your variables as frequency tables. Write a few sentences that describe what you see in the frequency tables. Click **HERE** for an example.

## Displaying Frequency Tables

SAS        **proc freq; tables** VAR1 VAR2 VAR3;

R        **library(descr)**  
          **freq**(as.ordered(myData\$VAR1))  
          **freq**(as.ordered(myData\$VAR2))  
          **freq**(as.ordered(myData\$VAR3))

PYTHON    c1 = myData['VAR1'].**value\_counts**(sort=False, dropna=False)  
          **print**(c1)

STATA     **tab1** VAR1 VAR2 VAR3

SPSS      **FREQUENCIES VARIABLES=**VAR1 VAR2 VAR3.  
          **/ORDER=ANALYSIS.**

---





# Data Management

---

**“You can have a ridiculously enormous and complex data set, but if you have the right tools and methodology then it’s not a problem.”**

– Aaron Koblin

## Overview

You have learned the basic steps for using your statistical software program and have begun to examine each of your variables through the process of exploratory data analysis. By generating frequency tables, you were able to describe the pattern of responses for each variable. Now we will use these frequency distributions (along with your personal code book) to make additional decisions regarding the organization of each of your variables.

**Data management** is the process of organizing and recoding your data in order to help you answer your research question. The number and types of decisions you will need to make will depend on the variables that you have selected.

## Lesson

Learn what types of response categories need to be set to missing. Consider if you need to logically recode response options to assist in interpreting your results. See how to collapse a quantitative variable into categories or an already categorical variable into fewer categories. Create secondary variables combining two or more variables using logic statements. Learn how to create a quantitative variable from multiple categorical variables. Begin to think about how these different data management steps might help you best answer your research question. Click on a video lesson below.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

## Calling in (importing) a data set

### Basic operations

<b>SAS</b>	EQ or =	>= or GE	<= or LE	> or GT	< or LT	!= or NE
<b>R</b>	==	>=	<=	>	<	!=
<b>PYTHON</b>	==	>=	<=	>	<	!= or <>
<b>STATA</b>	==	>=	<=	>	<	!=
<b>SPSS</b>	EQ or =	>= or GE	<= or LE	> or GT	< or LT	NE

### 1. Need to identify missing data

Often, you must define the response categories that represent missing data. For example, if the number 9 is used to represent a missing value, you must either designate in your program that this value represents missingness or else you must recode the variable into a missing data character that your statistical software will recognize. If you do not, the 9 will be treated as a real/meaningful value and will be included in each of your analyses.

```
SAS      if VAR1=9 then VAR1=.
```

```
R      myData$VAR1[myData$VAR1 == 9 ] <- NA
```

```
PYTHON  myData['VAR1']= myData['VAR1'].replace(9, numpy.nan)
```

```
STATA   replace VAR1=. if VAR1==9
```

```
SPSS    RECODE VAR1 (9=SYSMIS).
```

### 2. Need to recode responses to "no" based on skip patterns

There are a number of skip patterns in some data sets. For example, if we ask someone whether or not they have ever used marijuana, and they say "no", it would not make sense to ask them more detailed questions about their marijuana

use (e.g. quantity, frequency, onset, impairment, etc.). When analyzing more detailed questions regarding marijuana (e.g. have you ever smoked marijuana daily for a month or more?), those individuals that never used the substance may show up as missing data (often designated "legitimate skip"). Since they have never used marijuana, we can assume that their answer is "no", they have never smoked marijuana daily. This would need to be explicitly recoded. Note that we commonly code a no as 0 and a yes as 1.

```
SAS      if VAR1=. then VAR1=0;
```

```
R      myData$VAR1[is.na(myData$VAR1)] <- 0
```

```
PYTHON  myData['VAR1'].fillna(0, inplace=True)
```

```
STATA   replace VAR1=. if VAR1==9
```

```
SPSS    RECODE VAR1 (SYSMIS=0).
```

### 3. Creating a new variable by recoding a string variable into numeric variable

It is important when preparing to run statistical analyses that all variables have response categories that are numeric rather than "string" or "character". All variables with string or character responses must therefore be recoded into numeric values. These numeric values are known as dummy codes in that they carry no direct numeric meaning.

```
SAS      if TREE='Maple' then TREE_N=1;
          else if TREE='Oak' then TREE_N=2;
```

```
R      Not necessary in R.
```

```
PYTHON  def TREE_N (row) :
          if row['TREE'] == 'Maple' :
              return 1
          if row['TREE'] == 'Oak' :
              return 2
```

```
myData['TREE_N'] = myData.apply(lambda row: TREE_N (row), axis = 1)
```

STATA     **generate** TREE\_N=.  
           **replace** TREE\_N=1 **if** TREE=="Maple"  
           **replace** TREE\_N=2 **if** TREE=="Oak"

**OR by using the encode command**

**encode** TREE, **gen**(TREE\_N)

SPSS     **RECODE** TREE ('Maple'=1) ('Oak'=2) **INTO** TREE\_N.

**or row['region'] row['region'] ==5 or row['region'] == 6 :**  
           **return 1**

**elif row['region'] == 4 or row['region'] == 7 or row['region'] == 7**  
**or row['region'] row['region']== 8 or row['region'] == 9 :**  
           **return 2**

myData['new\_region'] = myData.apply(lambda row: new\_region (row), axis=1)

STATA     **generate** new\_region = 2  
           **replace** new\_region=1 **if** region==1 | region==2 | region==3 | region==5 |  
           region==6

**OR by using the recode command**

**recode** region (1/3 5 6=2), **gen**(new\_region)

SPSS     **COMPUTE** new\_region=2.  
           **IF** (region=1 | region=2 | region=3 | region=5 | region=6) new\_region=1.

#### 4. Creating a new variable by collapsing response categories

If a variable has many response categories, it can be difficult to interpret the statistical analyses in which it is used. Alternatively, there may be too few subjects or observations identified by one or more response categories to allow for a successful analysis. In these cases, you would need to collapse across categories. For example, if you have the following categories for geographic region, you may want to collapse some of these categories:

**Region:** New England=1, Middle Atlantic=2, East North Central=3, West North Central=4, South Atlantic=5, East South Central=6, West South Central=7, Mountain=8, Pacific=9.

**New\_Region:** East=1, West=2.

SAS     **if** region=1 or region=2 or region=3 or region=5 or region=6 **then**  
           new\_region=1;  
           **else if** region=4 or region=7 or region=8 or region=9 **then**  
           new\_region=2;

R     myData\$new\_region[myData\$region == 1 | myData\$region == 2 |  
           myData\$region == 3 | myData\$region == 5 |  
           myData\$region == 6] <- 1  
       myData\$new\_region[myData\$region == 4 | myData\$region == 7 |  
           myData\$region == 8 | myData\$region == 9] <-2

PYTHON **def** new\_region (row) :  
           **if** row['region'] == 1 **or** row['region'] == 2 **or** row['region'] == 3

#### 5. Creating a new variable by aggregating across variables

In many cases, you will want to combine multiple variables into one. For example, while NESARC measures several individual anxiety disorders, I may be interested in anxiety more generally. In this case I would create a general anxiety variable in which those individuals who received a diagnosis of social phobia, generalized anxiety disorder, specific phobia, panic disorder, agoraphobia, or obsessive-compulsive disorder would be coded "yes" and those who were free from all of these diagnoses would be coded "no".

SAS     **if** socphob=1 **or** gad=1 **or** specphob=1 **or** panic=1 **or** agora=1 **or**  
           ocd=1 **then** anxiety=1; **else** anxiety=0;

R     myData\$anxiety <- rep(0, nrow(myData))  
       myData\$anxiety[myData\$socphob == 1 | myData\$gad==1 |  
           myData\$panic == 1 | myData\$agora==1 |  
           myData\$ocd == 1] <- 1  
       myData\$anxiety[is.na(myData\$socphob) & is.na(myData\$gad) &

```
is.na(myData$panic) & is.na(myData$agora) &
is.na(myData$ocd)] <- NA
```

```
PYTHON def anxiety (row) :
    if row['socphob'] == 1 or row['gad'] == 1 or row['panic'] == 1
    or row['agora'] == 1 or row['ocd'] == 1 :
        return 1
    else :
        return 0
myData['anxiety'] = data.apply(lambda row: anxiety (row), axis = 1)
```

```
STATA gen anxiety=1 if socphob==1 | gad==1 | specphob==1 | ///
panic==1 | agora==1 | ocd==1
replace anxiety=0 if anxiety==.
```

```
SPSS IF (socphob=1 | gad=1 | specphob=1 | panic=1 | agora=1 | ocd=1)
anxiety=1.
RECODE anxiety (SYSMIS=0).
```

---

## 6. Create a quantitative variable

If you are working with a number of items that represent a single construct, it may be useful to create a composite variable/score. For example, I want to use a list of nicotine dependence symptoms meant to address the presence or absence of nicotine dependence (e.g. tolerance, withdrawal, craving, etc.). Rather than using individual binary variables (i.e. nicotine dependence present or absent), I want to examine the construct as a dimensional scale (i.e. number of nicotine dependence symptoms). In this case, I would want to recode each symptom variable so that yes=1 and no=0 and then sum the items so that they represent one composite score (i.e. the number of nicotine dependence symptoms endorsed).

```
SAS nd_sum=sum (of nd_symptom1 nd_symptom2 nd_symptom3
nd_symptom4);
```

```
R myData$nd_sum <- myData$nd_symptom1 +myData$nd_symptom2 +
myData$nd_symptom3 + myData$nd_symptom4
```

```
PYTHON myData['nd_sum'] = myData['nd_symptom1'] +
myData['nd_symptom2'] + myData['nd_symptom3'] +
myData['nd_symptom4']
```

```
STATA egen nd_sum=rsum(nd_symptom1 nd_symptom2 nd_symptom3 ///
nd_symptom4)
```

```
SPSS COMPUTE nd_sum=sum(nd_symptom1 nd_symptom2
nd_symptom3 nd_symptom4).
```

---

## 7. Labeling variables

Given the often-cryptic names that variables are given, it can sometimes be useful to label them.

```
SAS label VAR1='label';
```

```
R For frequency tables:
library (Hmisc)
label(myData$VAR1) <- "label"
```

```
PYTHON Not needed in Python
```

```
STATA label variable VAR1 "label"
```

```
SPSS VARIABLE LABELS VAR1 'label'.
```

---

## 8. Renaming variables

Given the often-cryptic names that variables are given, it can also sometimes be useful to give a variable a new name (something that is easier for you to remember and recognize).

```
SAS rename VAR1=newvarname;
```

R `names(myData)[names(myData)== "VAR1"] <- "newvarname"`

PYTHON `myData = myData.rename(index=str, columns={"VAR1": "newvarname"})`

STATA `rename VAR1 newvarname`

SPSS `COMPUTE newvarname=VAR1.`

`myData['VAR1']= myData['VAR1']..astype('category')`  
`myData['VAR1']=`  
`myData['VAR1'].cat.rename_categories(["value0label",`  
`"value1label", "value2label", "value3label"])`

STATA `label define LABELname1 0 "value0label" 1 "value1label" 2 ///`  
`"value2label" 3 "value3label"`  
`label values VAR1 LABELname1`

SPSS `VALUE LABELS VAR1 0 'value0label' 1 'value1label' 2`  
`'value2label' 3 'value3label'.`

## 9. Labeling variable responses/values

Given that nominal and ordinal variables have, or are given numeric response values (i.e. dummy codes), it can be useful to label those values so that the labels are displayed in your output.

SAS **Set up format before the data step.**  
`proc format; value formatname 0="value0label" 1="value1label"`  
`2="value2label" 3="value3label";`

**Before the end of the data step, tell SAS which variables you would like to format with these values.**

`format VAR1 formatname.`

R **Because the function doesn't name the existing levels, make sure you have them all in the right order.**

`levels(myData$VAR1)`

`levels(myData$VAR1) <- c("value0label", "value1label",`  
`"value2label", "value3label")`

PYTHON **Because the function doesn't name the existing levels, make sure you have them all in the right order.**

## 10. Need to further subset the sample

When using large data sets, it is often necessary to subset the data so that you are including only those observations that can assist in answering your particular research question. In these cases, you may want to select your own sample from within the survey's sampling frame. For example, if you are interested in identifying demographic predictors of depression among Type II diabetes patients, you would plan to subset the data to subjects endorsing Type II Diabetes.

SAS **put in the data step before sorting the data**  
`if diabetes2=1;`

R `title_of_subsetted_data <- myData[myData$diabetes2 == 1, ]`

PYTHON `title_of_subsetted_data = myData[myData.diabetes2 == 1]`

STATA **put this after the command**  
`if diabetes2==1`

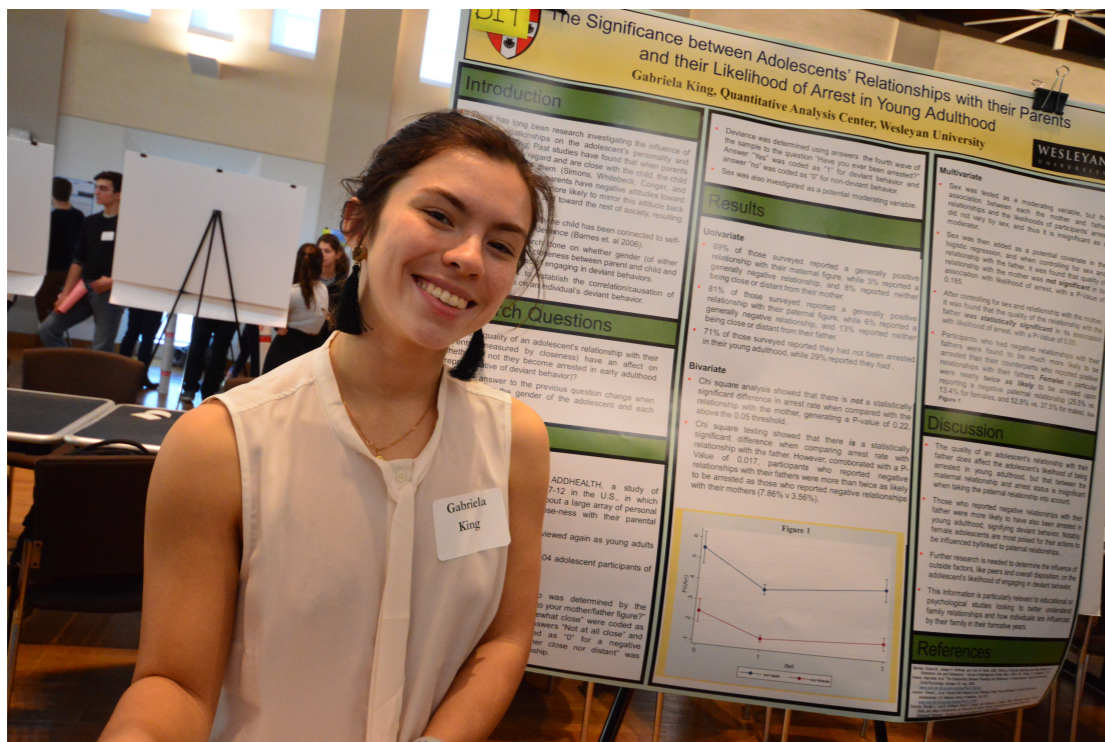
SPSS **must add as command option**  
`/SELECT=diabetes2 EQ 1`

## 11. Need to create groups that will be compared to one another

Often, you will need to create groups for the purpose of making comparisons. It is important to be certain that the groups that you would like to compare are of adequate size and number. For example, if you were interested in comparing complications of depression in parents who had lost a child through miscarriage vs. parents who had lost a child in the first year of life, it would be important to have large enough groups of each. It would not be appropriate to attempt to compare 5000 observations in the miscarriage group to only 5 observations in the first-year group.

## Data Management Assignment:

Submit the program that manages your data and the results that display your data managed variables as frequency tables. Describe the steps that you took in managing your data and write a few sentences that describe what you see in the frequency tables of your managed variables. Click [HERE](#) for an example.







# Univariate Graphing

---

“The purpose of visualization is insight,  
not pictures.”

—Ben Shneiderman

## Overview

You have chosen a research question, selected several variables, and implemented data management decisions that will help you answer your research question. Next, you will learn how to graph your newly managed variables one at a time (i.e. univariate) in order to better visualize their distributions.

**Univariate graphing** is the process of visualizing the distribution of one variable at a time.

## Lesson

There are a variety of conventional ways to visualize data - tables, histograms, bar graphs, etc. Learn how to visualize a frequency distribution as a graphical display. See how a bar chart that graphs one categorical variable contains bars representing each category on the x-axis and number or percent of observations on the y-axis. Create a histogram for quantitative variables that plots quantitative intervals on the x-axis and the number or percentage of observations on the y-axis. Understand how to describe a histogram's pattern with shape, center, and spread as well as deviations from those patterns (i.e. outliers). Learn how we use the numerical measures for the mean and median to describe the center of the distribution. See how we use measures of spread to describe the distribution's variability. Click on a video lesson below.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

## Univariate Graph (categorical variables)

SAS	<pre>proc sgplot; vbar CategVar/ stat=percent; title "Title Here";</pre>
R	<pre>library(ggplot2) ggplot(data=myData)+ geom_bar(aes(x=CategVar))+ ggtitle("Descriptive Title Here")</pre>
PYTHON	<pre>import seaborn seaborn.countplot(x="CategVar", data=myData) plt.xlabel("Label for CategVar") plt.title("Descriptive Title")</pre>
STATA	<pre>graph bar, over(CategVar)</pre>
SPSS	<b>Use graphical user interface (GUI)</b>

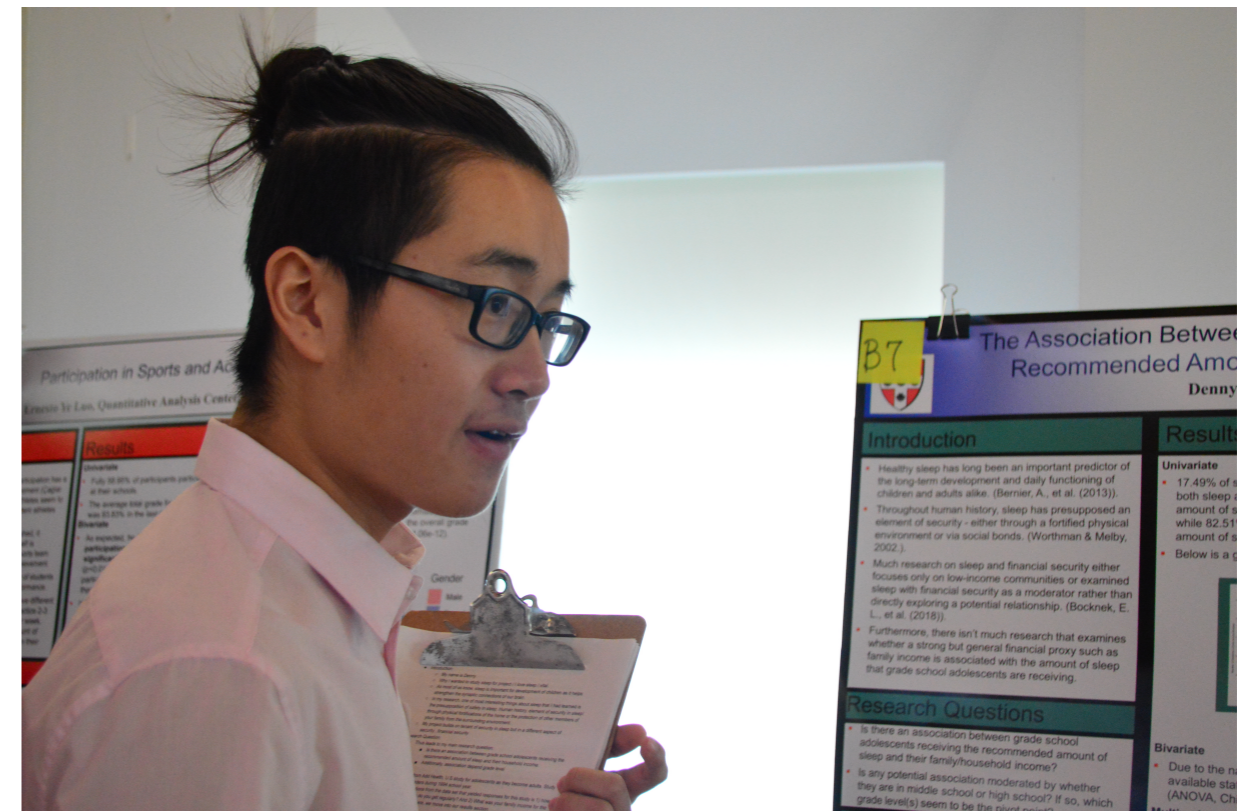
---

## Univariate Graph (quantitative variables)

SAS	<pre>proc sgplot; histogram QuantVar; title "Title Here";</pre>
R	<pre>ggplot(data=myData)+ geom_histogram(aes(x=QuantVar))+ ggtitle("Descriptive Title Here")</pre>
PYTHON	<pre>import seaborn seaborn.distplot(myData["QuantVar"].dropna(), kde=False) plt.xlabel("Label for QuantVar") plt.title("Descriptive Title Here")</pre>
STATA	<pre>histogram QuantVar</pre>
SPSS	<b>Use graphical user interface (GUI)</b>

## Making univariate graphs:

Submit univariate graphs for your two main constructs (i.e. data managed variables). Write a few sentences under each graph describing what it shows. For categorical variables, what response category shows the modal (i.e. most common) frequency and which shows the least common? For quantitative variables, describe the shape (i.e. symmetry/skewness and modality), center (i.e. midpoint) and spread (i.e. range) of the distribution, as well as any outliers. Some categorical variables are ordered (e.g. 1=strongly agree, 2=agree, 3=neither agree nor disagree, 4=disagree, 5=strongly disagree). For ordered variables what is the modal frequency and is there anything that can be said about the shape, center and spread of the distribution? Click [HERE](#) for an example.





# Bivariate Graphing

---

“The greatest value of a picture is when it forces us to notice what we never expected to see.”

—John Tukey

## Overview

So far, we have examined data obtained from one variable at a time (either categorical or quantitative) and learned how to describe the features and distribution of the variable using the appropriate visual displays and numerical measures. Now we will consider two variables simultaneously and explore the relationship between them using, as before, visual displays and numerical summaries.

When graphing your variables, it is important that each graph provides clear and accurate summaries of the data that do not mislead.

## Lesson

Understand why we impose a causal model on our research question despite the fact that causation cannot be directly evaluated based on observational data. Assign roles to each of your variables. Which will play the role of **explanatory variable** and which will play the role of **response variable**? Learn to use the graphing decision flow chart to determine, based on variable types, the appropriate graph for visualizing each relationship. Consider what types of bivariate graphs will help you graphically visualize your research question. Understand how to interpret bivariate graphs. Note that in some cases, the bivariate association may differ for population subgroups. Graphing the subgroups by adding a third variable will help to visually determine if population subgroup differences may exist. Click on a video lesson below.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

### Bivariate Graph (categorical explanatory, binary categorical response)

SAS `proc sgplot; vbar CategExplanatoryVar /  
response=CategResponseVar stat=mean;  
title "Title Here";`

R `ggplot(data=myData)+  
stat_summary(aes(x=CategExplanatoryVar, y=CategResponseVar),  
fun.y=mean, geom="bar")+  
ggtitle("Descriptive Title Here")`

PYTHON `seaborn.factorplot(x="CategExplanatoryVar",  
y="QuantResponseVar", data=myData, kind="bar", ci=None)  
plt.xlabel('Label for CategExplanatoryVar')  
plt.ylabel('Label for QuantResponseVar')  
plt.title('Descriptive Title Here')`

STATA `graph bar (mean) CategResponseVar, over(CategExplanatoryVar)`

SPSS [Use graphical user interface \(GUI\)](#)

---

### Bivariate Graph (categorical explanatory, quantitative response)

SAS `proc sgplot; vbar CategExplanatoryVar  
/ response=QuantResponseVar stat=mean;  
title "Title Here";`

R [##Below is code for bar graph](#)  
`ggplot(data=myData)+  
stat_summary(aes(x=CategExplanatoryVar,  
y=QuantResponseVar), fun.y=mean, geom="bar")`

[##Below is code for boxplots](#)  
`ggplot(data=myData)+  
geom_boxplot(aes(x=CategExplanatoryVar,  
y=QuantResponseVar))+ ggtitle("Descriptive Title Here")`

### [##Below is code for density plots](#)

`ggplot(data=myData)+  
geom_density(aes(x=QuantResponseVar,  
color=CategExplanatoryVar))+ggtitle("Descriptive Title Here")`

PYTHON `scat1 = seaborn.factorplot(x=CategExplanatoryVar,  
y=QuantResponseVar, data=myData, kind="bar", ci=None)  
print(scat1)`

STATA `graph box QuantResponseVar, over(CategExplanatoryVar)`

SPSS [Use graphical user interface \(GUI\)](#)

---

### Bivariate Graph (quantitative explanatory, quantitative response)

SAS `proc sgplot; scatter x=QuantExplanatoryVar  
y=QuantResponseVar;  
title "Title Here";`

R `ggplot(data=myData)+  
geom_point(aes(x=QuantExplanatoryVar, y=QuantResponseVar))+  
geom_smooth(aes(x=QuantExplanatoryVar, y=QuantResponseVar),  
method="lm")`

[## Note, adding 3rd variable by using the color argument](#)  
`ggplot(data=myData)+  
geom_point(aes(x=QuantExplanatoryVar, y=QuantResponseVar,  
color=CategThirdVar))+  
geom_smooth(aes(x=QuantExplanatoryVar, y=QuantResponseVar,  
color=CategThirdVar), method="lm")`

PYTHON `seaborn.regplot(x="QuantExplanatoryVar",  
y="QuantResponseVar",  
fit_reg=False, data=myData)  
plt.xlabel('Label for QuantExplanatoryVar')`

```
plt.ylabel('Label for QuantResponseVar')
plt.title('Descriptive Title Here')
```

STATA **twoway (scatter** QuantResponseVar QuantExplanatoryVar) ///  
**(lfit** QuantResponseVar QuantExplanatoryVar) // adds reg line

SPSS **GRAPH /scatterplot(bivar)=**QuantExplanatoryVar with  
QuantResponseVar. [or use graphical user interface \(GUI\)](#)

---

### Bivariate Graph with a Third Variable (Categorical Explanatory Variable, Categorical Response Variable, Categorical 3rd Variable):

SAS **proc sgplot; vbar** ExplVar /**response=**RespVar **group=**ThirdVar  
**groupdisplay=cluster stat=mean;**  
**title** "Title Here";

R `tab1 <- ftable(myData$CategResponseVar,`  
`myData$CategExplanatoryVar,`  
`myData$CategThirdVar)`  
`tab1`  
`tab1_colProp <- prop.table(tab1, 2)`  
`tab1_colProp`

PYTHON **seaborn.factorplot**(x="CategExplanatoryVar",  
y="CategResponseVar", hue="CategThirdVar", data=myData,  
kind="bar", ci=None)  
**plt.xlabel**('Label for CategExplanatoryVar')  
**plt.ylabel**('Label for CategResponseVar')  
**plt.title**('Descriptive Title Here')

STATA **ssc install catplot**  
**catplot** CategResponseVar CategExplanatoryVar, ///  
**percent**(CategExplanatoryVar over (CategThirdVar)

SPSS [Use graphical user interface \(GUI\)](#)

---

### Bivariate Graph with a Third Variable (Categorical Explanatory Variable, Quantitative Response Variable, Categorical 3rd Variable):

SAS **proc sgplot; vbar** ExplVar /**response=**RespVar **group=**ThirdVar  
**groupdisplay=cluster stat=mean;**  
**title** "Title Here";

R **fable**(by(myData\$QuantResponseVar,  
list(myData\$CategExplanatoryVar, myData\$CategThirdVar),  
mean, na.rm = TRUE))

PYTHON **seaborn.barplot**(x="CategExplanatoryVar", y="QuantResponseVar",  
hue="CategThirdVar", data=myData, ci=None)  
**plt.xlabel**('Label for CategExplanatoryVar')  
**plt.ylabel**('Label for QuantResponseVar')  
**plt.title**('Descriptive Title Here')

STATA **graph bar (mean)** QuantResponseVar, ///  
**over**(CategExplanatoryVar) **over**(CatThirdVar) ///  
**blabel**(bar) // displays mean value for each bar

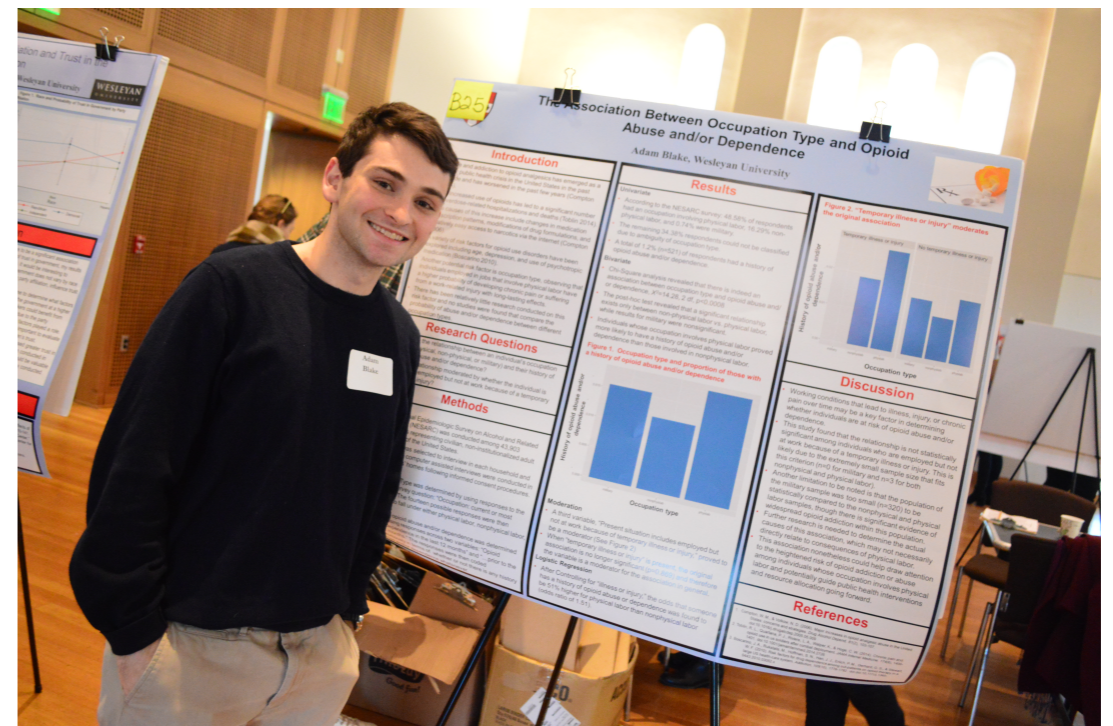
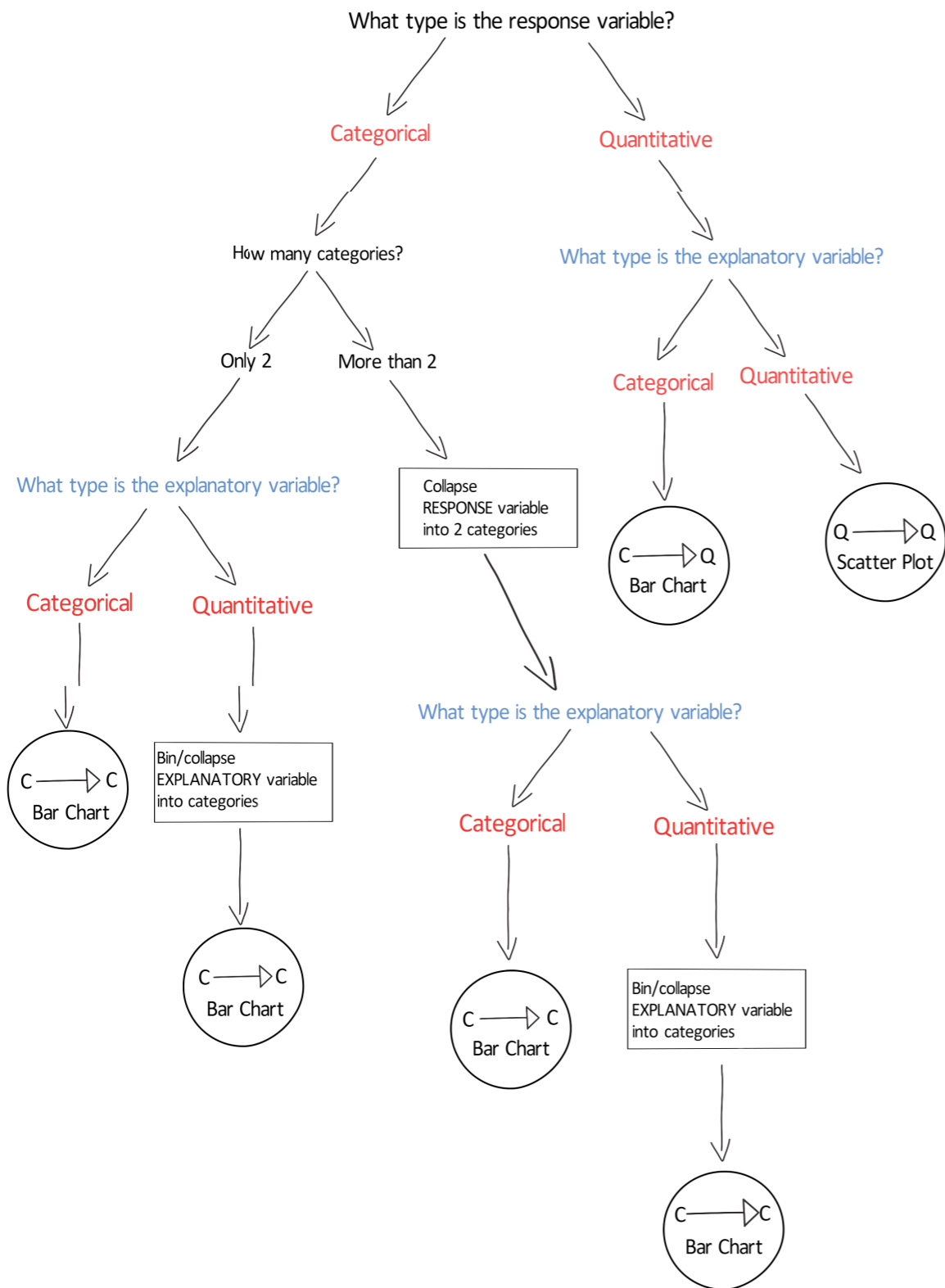
SPSS [Use graphical user interface \(GUI\)](#)

---

## Making Bivariate graphs:

Submit a graph showing the association between your explanatory variable and your response variable. Describe what the graph shows. Include an additional graph(s) of your bivariate relationship by a third variable. Describe any similarities or differences in the association between your explanatory and response variable for different levels of your third variable (i.e. different population subgroups). Click **HERE** for an example.

# Graphing Flow Chart:





# Hypothesis Testing

---

"Statistics may be defined as 'a body of methods for making wise decisions in the face of uncertainty'."

- W.A. Wallis

## Overview

Thus far, we have focused on descriptive statistics. Through our examination of frequency distributions, graphical representations of our variables, and calculations of center and spread, the goal has been to describe and summarize data. Now you will be introduced to inferential statistics. In addition to describing data, inferential statistics allow us to directly test our hypotheses by evaluating our research question (based on a sample) with the goal of generalizing the results to the larger population from which the sample was drawn.

**Hypothesis testing** is one of the most common applications of statistics to real life problems. It is used when we need to make decisions concerning populations on the basis of information on only a sample. A variety of statistical tests are used to arrive at these decisions (e.g. Analysis of Variance, Chi- Square Test of Independence, etc.).

## Lesson

Learn about the 4 steps involved in hypothesis testing: specifying the null ( $H_0$ ) and alternate ( $H_a$ ) hypotheses; choosing a sample; assessing the evidence; and drawing conclusions. Identify the null hypothesis ( $H_0$ ) and alternative hypotheses ( $H_a$ ) for your research question(s). See how to calculate how likely it is to get data like that observed when  $H_0$  is true. Understand when to reject the null hypothesis based on the calculated p-value. Compare your calculated p-value to the standard alpha level ( $\alpha = .05$  or 5%). Consider how this allows us to make inferences about the population based on our sample. Click **HERE** to watch the video lesson.

	(Population) Parameter	(Sample) Statistic
<b>Proportion</b>	$p$	$\hat{p}$
<b>Mean</b>	$\mu$	$\bar{x}$
<b>Standard Deviation</b>	$\sigma$	$s$

## Conducting a Hypothesis Test:

Write out your research question(s). Be sure to remove any words suggesting cause and effect (i.e. use the words associated, related, more likely, less likely, etc.). Also describe the **sample** you will focus on (e.g., representative sample of U.S. adolescents, representative sample of U.S. adults, individuals who have smoked in their lifetime, older adults between 60 and 95, etc.).

Which topic or construct is your **explanatory variable**? Which is your **response variable**? What is the variable name of your explanatory and response variables (i.e. your data management variables)?

Next, label your explanatory and response variables as **categorical or quantitative**. If the variable can be either categorical or quantitative (e.g. it is ordered with higher numbers meaning more and lower numbers meaning less) write “categorical” if it is your explanatory variable and “quantitative” if it is your response variable.

Finally, write the **null and alternate hypothesis** for your research question. Which **statistical test** will you need to run in order to answer your research question?

### Example:

Is smoking quantity associated with nicotine dependence? That is, do young adult smokers (my sample) who smoke more cigarettes have a higher likelihood of meeting criteria for nicotine dependence than those who smoke fewer cigarettes?

Explanatory variable: smoking quantity  
(NUMCIGMO\_EST - categorical)

Response variable: nicotine dependence  
(TAB12MDX – categorical)

Ho = There is no association between smoking quantity and nicotine dependence among young adult smokers.

Ha = There is an association between smoking quantity and nicotine dependence among young adult smokers.

C -> C: Chi-square Test of Independence







# Analysis of Variance

---

**Analysis of Variance (ANOVA) evaluates whether the means of two or more groups are statistically different from each other.**

## Overview

You have learned about inferential statistics, the steps in hypothesis testing and the role of probability in measuring or quantifying uncertainty as a way of drawing conclusions about a particular population from information available from only a sample. In the next three chapters, you will be introduced to specific inferential tools that allow us to use inferential principals to evaluate associations between different types of variables (i.e. those measured categorically and/or quantitatively).

Analysis of Variance (ANOVA) evaluates whether the means of two or more groups are statistically different from each other. This test is appropriate whenever you want to compare the means (quantitative response variable) of groups (categorical explanatory variable).

## Lesson

Learn about the ANOVA F test. Consider the null and alternative hypothesis. See how variation within the sample means and variation between the groups plays a role in our evaluation of statistical significance. Learn how to run an ANOVA test and interpret results within the context of real data. Understand the critical role of post hoc tests for interpreting ANOVA results when our explanatory variable includes more than two groups or levels. See how post hoc tests allow us to avoid increasing the likelihood of rejecting the null hypothesis when the null hypothesis should be accepted, a problem known as Type 1 error. Click on a video lesson below.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

## Analysis of Variance:

SAS `proc anova; class CategExplanatoryVar;  
model QuantResponseVar = CategExplanatoryVar;  
means CategExplanatoryVar;`

[\(with post hoc test\)](#)

`proc anova; class CategExplanatoryVar;  
model QuantResponseVar = CategExplanatoryVar;  
means CategExplanatoryVar /duncan;`

R `myAnovaResults <- aov(QuantResponseVar ~  
CategExplanatoryVar, data = myData)  
summary(myAnovaResults)`

[\(with post hoc test\)](#)

`myAnovaResults <- aov(QuantResponseVar ~  
CategExplanatoryVar, data = myData)  
TukeyHSD(myAnovaResults)`

PYTHON `import statsmodels.formula.api as smf  
import statsmodels.stats.multicomp as multi  
model1 = smf.ols(formula='QuantResponseVar ~  
C(CategExplanatoryVar)', data=myData)  
results1 = model1.fit()  
print (results1.summary())`

[\(with post hoc test\)](#)

`model1 = smf.ols(formula='QuantResponseVar ~  
C(CategExplanatoryVar)', data=myData)  
results1 = model1.fit()  
print (results1.summary())`

[continued...](#)

```
sub1 = myData[['QuantResponseVar',  
'CategExplanatoryVar']].dropna() mc1 =  
multi.MultiComparison(sub1['QuantResponseVar'],  
sub1['CategExplanatoryVar'])  
res1 = mc1.tukeyhsd()  
print(res1.summary())
```

STATA `oneway QuantResponseVar CategExplanatoryVar, tabulate  
  
\(with post hoc test\)`

`oneway QuantResponseVar CategExplanatoryVar, sidak`

SPSS `UNIANOVA QuantResponseVar BY CategExplanatoryVar.  
  
\(with post hoc test\)`

`UNIANOVA QuantResponseVar BY CategExplanatoryVar  
/POSTHOC=CategExplanatoryVar (TUKEY)  
/PRINT=ETASQ DESCRIPTIVE.`

## Conduct an ANOVA:

Run an ANOVA using a quantitative response variable and a categorical explanatory variable. State your research question, the null and alternate hypotheses and whether the ANOVA F value is statistically significant or not. Submit your output/results. Describe your results. You will need to analyze and interpret post hoc paired comparisons in instances where your original statistical test was significant, and you were examining more than two groups (i.e. more than two levels of a categorical, explanatory variable). Click **HERE** for an example.



# Chi-Square Test of Independence

---

**A Chi-Square Test of Independence evaluates whether the conditional proportions of two variables are statistically independent of one another.**

## Overview

The last statistical test that we studied (Analysis of Variance or ANOVA) involved examining the relationship between a categorical explanatory variable and a quantitative response variable. Next, we will consider inference in the context of relationships between two categorical variables, corresponding to case C->C.

The Chi-Square Test of Independence allows us to examine our observed data and to evaluate whether we have enough evidence to conclude, with a reasonable level of certainty ( $p < 0.05$ ), that two categorical variables are related.

## Lesson

Learn about the Chi-Square Test of Independence. Consider the null and alternative hypotheses when using this test. See how the fit of the observed data with the data that is expected plays a role in our evaluation of statistical significance. Learn how to run a Chi-Square Test of Independence and interpret results within the context of real data. Understand the critical role of post hoc tests for interpreting Chi-Square results when our explanatory variable includes more than two groups or levels. See how post hoc tests allow us to avoid increasing the likelihood of rejecting the null hypothesis when the null hypothesis should be accepted, a problem known as Type 1 error. Practice controlling for Type I error using the Bonferroni adjustment. Click on a video lesson below.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

## Chi-Square Test of Independence:

SAS `proc freq; tables CategResponseVar*CategExplanatoryVar / chisq;`

**(with post hoc test)**

`if (CategExplanatoryVar = 1) or (CategExplanatoryVar = 3);  
proc freq; tables CategResponseVar*CategExplanatoryVar / chisq;`

R `myChi <- chisq.test(myData$CategResponseVar,  
myData$CategExplanatoryVar)  
myChi  
myChi$observed # for actual, observed cell counts  
prop.table(myChi$observed, 2) # for column percentages  
prop.table(myChi$observed, 1) # for row percentages`

**(with post hoc test)**

`library(fifer)  
myChi <- chisq.test(myData$CategResponseVar,  
myData$CategExplanatoryVar) observed_table<- myChi$observed  
chisq.post.hoc(observed_table, popsInRows=FALSE,  
control="bonferroni"),1:2]`

PYTHON `import scipy.stats  
ct1=pandas.crosstab(myData['CategResponseVar'],  
myData['CategExplanatoryVar'])  
print ('chi-square value, p value, degrees of freedom,  
expected counts')  
  
cs1= scipy.stats.chi2_contingency(ct1)  
print (cs1)  
  
# column percentages  
colsum=ct1.sum(axis=0)  
colpct=ct1/colsum  
print(colpct)`

**(with post hoc test)**

`#for each Chi Sq pair data subset  
ct1=pandas.crosstab(myData['CategResponseVar'],  
myData['CategExplanatoryVar'])  
cs1= scipy.stats.chi2_contingency(ct1)  
print (cs1)`

STATA `tab CategResponseVar CategExplanatoryVar , chi2 row col`

**(with post hoc test)**

`tab CategResponseVar CategExplanatoryVar if ///  
CategExplanatoryVar==1 | CategExplanatoryVar==3, chi2`

SPSS `CROSSTABS/TABLES= CategResponseVar by  
CategExplanatoryVar /STATISTICS=CHISQ.`

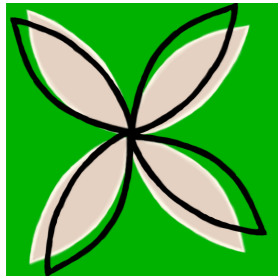
**(with post hoc test)**

**TEMPORARY.**

`SELECT IF CategExplanatoryVar=1 OR CategExplanatoryVar =3.  
CROSSTABS TABLES= CategResponseVar CategExplanatoryVar  
/STATISTICS=CHISQ.`

## Conduct a Chi-Square Test:

Run a Chi-Square Test of Independence using a categorical response variable and a categorical explanatory variable (your response variable must have only two levels). State your research question, the null and alternate hypotheses and whether the Chi-Square value is statistically significant or not. Submit your output/results. Describe your results. You will need to analyze and interpret post hoc paired comparisons in instances where your original statistical test was significant, and you were examining more than two groups (i.e. more than two levels of a categorical, explanatory variable). Click **HERE** for an example.



# Pearson Correlation

---

**A correlation coefficient assesses the degree of linear relationship between two variables. It can take values from +1 to -1.**

## Overview

The last statistical test that we studied (Chi-Square Test of Independence) involved the relationship between a categorical explanatory variable and a categorical response variable. Next, we will consider inference in the context of relationships between two quantitative variables, corresponding to case Q->Q.

The Pearson Correlation examines our observed data and evaluates whether we have enough evidence to conclude with a reasonable level of certainty ( $p < 0.05$ ) that there is a relationship between two quantitative variables. A correlation coefficient evaluates the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect, positive, linear relationship between the two variables. A correlation of -1 means there is a perfect, negative linear relationship between the two variables. In both cases, knowing the value of one variable, you can perfectly predict the value of the second.

## Lesson

Learn about the Pearson Correlation test. Consider the null and alternative hypothesis when using this test. See how the strength and direction of the relationship between the two variables plays a role in our evaluation of statistical significance. Learn how to run a Pearson Correlation test and interpret results within the context of real data. Click on a video lesson below.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

## Pearson Correlation:

SAS `proc corr; var QuantResponseVar QuantExplanatoryVar;`

R `cor.test(myData$QuantResponseVar,  
myData$QuantExplanatoryVar)`

PYTHON `import scipy  
sub1= myData[['QuantResponseVar', '  
QuantExplanatoryVar']].dropna()`

`print ('association between QuantExplanatoryVar and  
QuantResponseVar')`

`print (scipy.stats.pearsonr(sub1 ['QuantResponseVar'],  
sub1 ['QuantExplanatoryVar']))`

STATA `corr QuantResponseVar QuantExplanatoryVar`

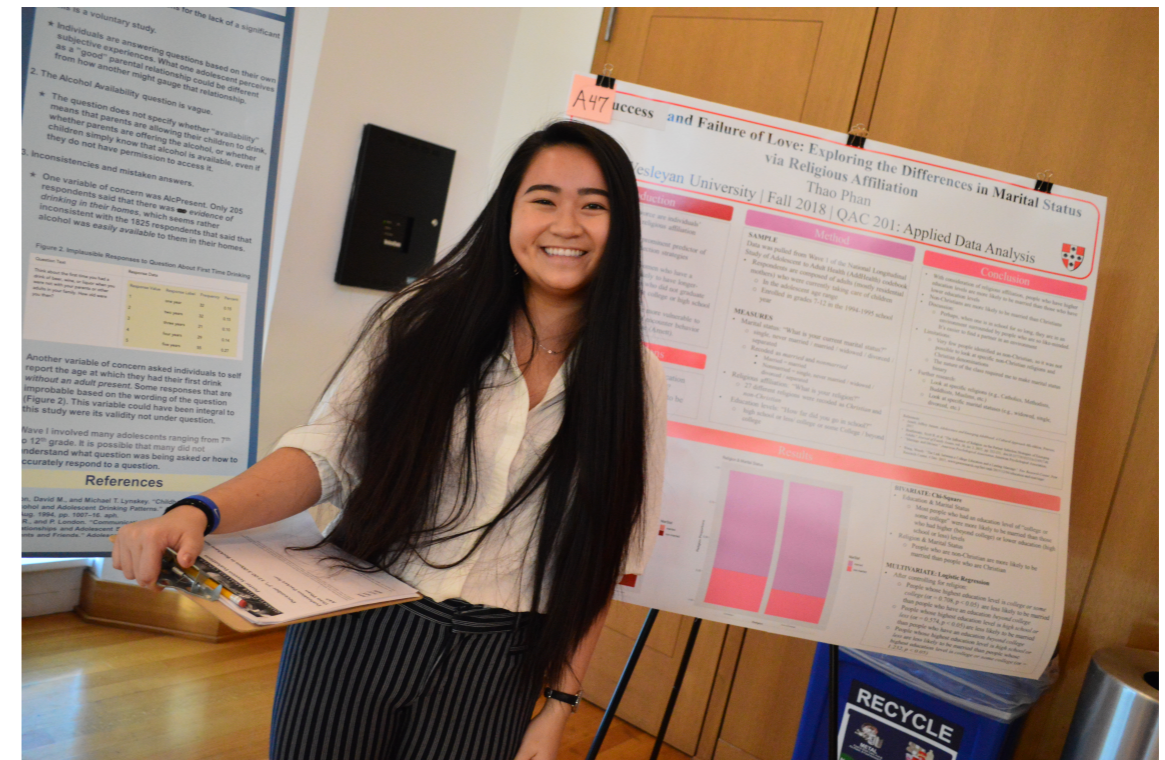
or

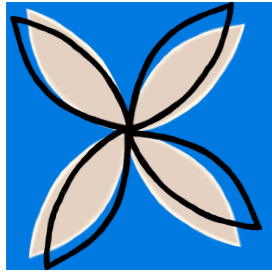
`pwcorr QuantResponseVar QuantExplanatoryVar, sig`

SPSS `CORRELATIONS/VARIABLES= QuantResponseVar  
QuantExplanatoryVar/STATISTICS DESCRIPTIVES.`

## Conduct a Pearson Correlation:

Run a Pearson correlation test using a quantitative response variable and a quantitative explanatory variable. State your research question, the null and alternate hypotheses and whether the Pearson Correlation Coefficient is statistically significant or not. Submit your output/results. Describe your results. Click **HERE** for an example.





# Moderation

---

“The world is one big data problem.”

- Andrew McAfee

## Overview

You have learned about specific inferential tools (i.e. ANOVA, Chi-Square Test of Independence, and Pearson Correlation) that allow us to use inferential principals to evaluate associations between variables (i.e. those measured both categorically and/or quantitatively). You will now learn to examine a bivariate relationship to determine if it depends on a third variable (i.e. the relationship may differ for population subgroups). This is known as moderation (or statistical interaction)

In statistics, moderation occurs when the relationship between two variables depends on a third variable. In this case, the third variable is referred to as the moderating variable or simply the moderator. The effect of a moderating variable is often characterized statistically as an interaction; that is, a third variable that affects the direction and/or strength of the relation between your explanatory and response variable.

## Lesson

Learn about moderation. Consider how we can test bivariate relationships that we examined with ANOVA, Chi-Square Test of Independence, and Pearson Correlation to determine if the relationship is different at different levels of a categorical third variable. Consider the null and alternative hypothesis when testing for moderation. Understand how examining moderation (i.e. statistical interaction) is a way to better understand your data. Click on a video lesson below.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

### Moderation for ANOVA:

```
SAS      proc sort; by CategThirdVar;  
        proc anova; class CategExplanatoryVar;  
        model QuantResponseVar=CategExplanatoryVar;  
        means CategExplanatoryVar; by CategThirdVar;  
  
R      by(myData, myData$CategThirdVar, function(x)  
      list(aov(QuantResponseVar ~ CategExplanatoryVar, data = x),  
      summary(aov( QuantResponseVar ~ CategExplanatoryVar, data =  
      x))))  
  
PYTHON #subset by categorical 3rd variable  
  
sub2=myData[(myData['CategThirdVar']=='Group 1')]  
sub3=myData[(myData['CategThirdVar']=='Group 2')]  
  
import statsmodels.api  
import statsmodels.formula.api as smf  
  
model2 = smf.ols(formula='QuantResponseVar ~  
C(CategExplanatoryVar)', data=sub2).fit()  
print (model2.summary())  
  
model3 = smf.ols(formula= formula='QuantResponseVar ~  
        C(CategExplanatoryVar)', data=sub3).fit()  
  
print (model3.summary())  
  
STATA  bys CategThirdVar: oneway QuantResponseVar  
        CategExplanatoryVar, tab  
  
SPSS   SORT CASES BY CategThirdVar.  
        SPLIT FILE LAYERED BY CategThirdVar.  
        ONEWAY QuantResponseVar BY CategExplanatoryVar  
        / STATISTICS DESCRIPTIVES  
        / POSTHOC = BONFERRONI ALPHA (0.05).  
        SPLIT FILE OFF.
```

### Moderation for Chi-Square Test of Independence:

```
SAS      proc sort; by CategThirdVar;  
        proc freq; tables CategResponseVar*CategExplanatoryVar/chisq;  
        by CategThirdVar;  
  
R      by(myData, myData$CategThirdVar, function(x)  
      list(chisq.test(x$CategResponseVar, x$CategExplanatoryVar),  
      chisq.test(x$CategResponseVar, x$CategExplanatoryVar$observed),  
      prop.table(chisq.test(x$CategResponseVar,  
      x$CategExplanatoryVar) $observed, 2))) # column %s  
  
PYTHON #subset by categorical 3rd variable  
  
sub2=myData[(myData['CategThirdVar']=='Group 1')]  
sub3=myData[(myData['CategThirdVar']=='Group 2')]  
  
ct2=pandas.crosstab(sub2['CategResponseVar'],  
sub2['CategExplanatoryVar'])  
  
print (ct2)  
  
ct3=pandas.crosstab(sub3['CategResponseVar'],  
sub3['CategExplanatoryVar'])  
  
print (ct3)  
  
STATA  bys CategThirdVar: tab CategResponseVar CategExplanatoryVar, ///  
        chi2 row  
  
SPSS   CROSSTABS/TABLES=CategResponseVar by CategExplanatoryVar  
        by CategThirdVar  
        / CELLS = COUNT ROW  
        / STATISTICS = CHISQ.
```



## Moderation for Pearson Correlation:

SAS      **proc sort; by** CategThirdVar;  
          **proc corr; var** QuantResponseVar QuantExplanatoryVar;  
          **by** CategThirdVar;

R         **by**(myData, myData\$CategThirdVar, **function(x)**  
          **cor.test(x\$QuantResponseVar, x\$QuantExplanatoryVar))**

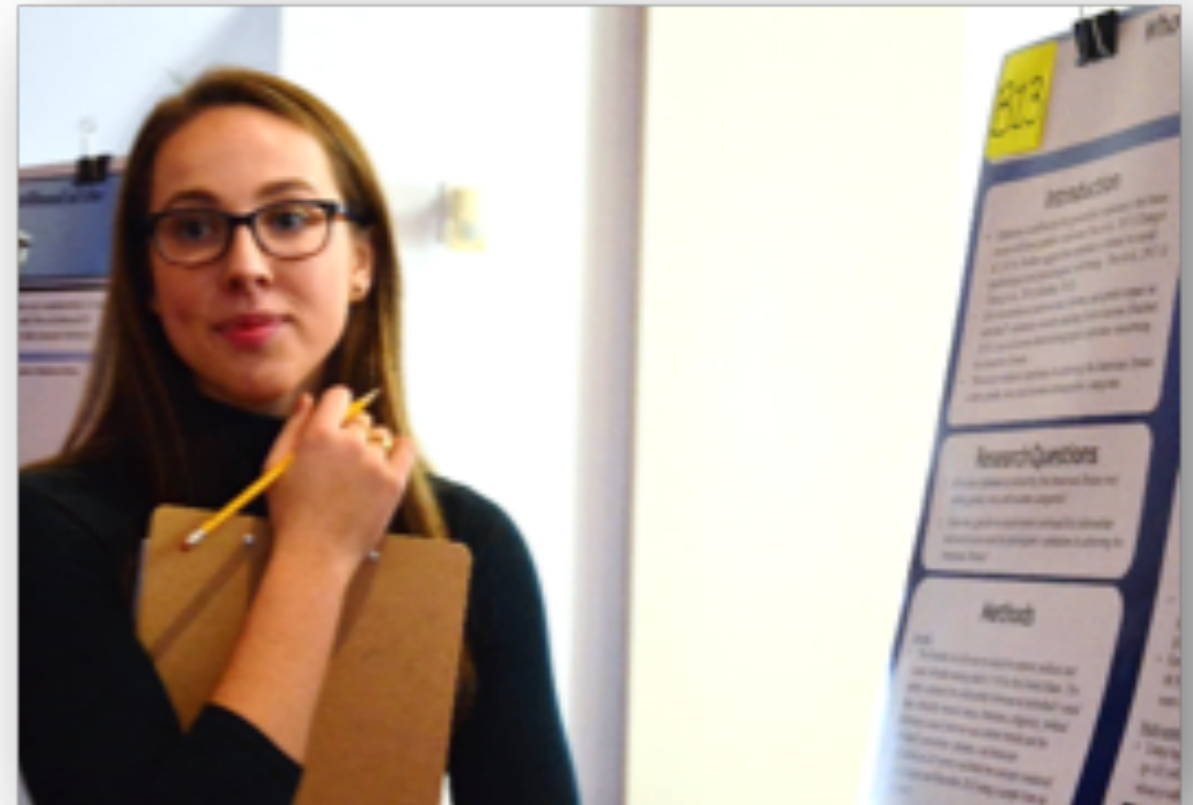
PYTHON   #subset by categorical 3rd variable  
          sub1=myData[(myData['CategThirdVar']== 1)]  
          sub2=myData[(myData['CategThirdVar']== 2)]  
          sub3=myData[(myData['CategThirdVar']== 3)]  
  
          **print (scipy.stats.pearsonr(sub1['QuantResponseVar'],**  
          sub1['QuantExplanatoryVar']))  
  
          **print (scipy.stats.pearsonr(sub2['QuantResponseVar'],**  
          sub2['QuantExplanatoryVar']))  
  
          **print (scipy.stats.pearsonr(sub3['QuantResponseVar'],**  
          sub3['QuantExplanatoryVar']))

STATA     **bys** CategThirdVar: **corr** QuantResponseVar QuantExplanatoryVar

SPSS     **SORT CASES BY** CategThirdVar.  
          **SPLIT FILE LAYERED BY** CategThirdVar.  
          **CORRELATIONS**  
          / **VARIABLES=** QuantResponseVar QuantExplanatoryVar  
          / **STATISTICS DESCRIPTIVES.**  
          **SPLIT FILE OFF.**

## Evaluating Moderation:

Test whether a categorical third variable moderates your association of interest (i.e. your research question) using ANOVA, Chi-Square Test of Independence, or Pearson Correlation. Submit the program along with corresponding output for the bivariate relationship and the moderation output. Describe any similarities or differences in the association between your explanatory and response variable for different levels of your third variable (i.e. different population subgroups). Click **HERE** for an example.





# The Question of Causation

---

“Statistics are no substitute for judgment.”

- Henry Clay

## Overview

You have tested for statistical associations and learned how to determine if a third variable moderates the strength and/or direction of a bivariate relationship. It is important to remember that your results should be considered in the context of observational data. Put simply, a significant association does not mean a causal relationship has been established. It is important not to use causal language in the presentation of your findings.

## Lesson

Learn about confounding variables (i.e. those variables that explain or account for a significant association between two variables). Consider the ethical reasons behind why some experiments cannot be conducted. Compare and contrast a prospective study and retrospective study in helping determine if a causal relationship exists between two variables. Understand that lurking variables make a relationship between two variables only appear strong. See how some relationships cannot be examined by cause and effect experiments, but with enough non-experimental evidence from many different sources we can make confident causal conclusions. Click **HERE** to watch the video lesson.



# Multivariate Models

---

"All models are wrong, but some are useful."

- George E. P. Box

## Overview

You have learned that a significant association or correlation does not mean causation. Next you will learn to determine statistically if a third variable is a confounder, or can account for a bivariate relationship you found to be significant using ANOVA, Chi-Square Test of Independence, or Pearson Correlation.

Using multivariate models allows us to statistically control for additional variables that may account for (i.e. confound) a significant bivariate association.

## Lesson

Learn about confounding and multivariate models. Understand that testing for likely confounding variables helps us to get slightly closer to establishing a cause and effect relationship when conducting an observational study. Determine if you need to use multiple regression because your response variable is quantitative or logistic regression because your response variable is categorical (with 2 levels). Consider the evidence for when a third variable is, or is not, a confounder. Click on a video lesson below.

**SAS**

**R**

**Python**

**Stata**

**SPSS**

### Multiple Regression (quantitative response variable):

SAS **Code binary variables as yes = 1 and no = 2**

```
proc glm; class CategExplanatoryVar CategThirdVar;
model QuantResponseVar = CategExplanatoryVar CategThirdVar
QuantThirdVar /solution;
```

R

```
my.lm <- lm(QuantResponseVar ~ ExplanatoryVar + ThirdVar1 +
ThirdVar2, data = myData)
summary(my.lm)
```

PYTHON

```
import statsmodels.api
import statsmodels.formula.api as smf

#note that categorical explanatory/third variables have to be entered
as C(CategVar)

lm1 = smf.ols('QuantResponseVar ~ ExplanatoryVar +
C(CategThirdVar1) + QuantThirdVar', data=myData).fit()

print (lm1.summary())
```

STATA

```
reg QuantResponseVar ExplanatoryVar ThirdVar1 ThirdVar2
```

SPSS

```
REGRESSION
/DEPENDENT QuantResponseVar
/METHOD ENTER ExplanatoryVar ThirdVar1 ThirdVar2.
```

## Conducting Multivariate Models:

Run a multiple regression model (quantitative response variable) or a logistic regression model (binary, categorical response variable). Submit the program that tests for confounding along with corresponding output. Describe in a few sentences what you found. Click [HERE](#) for an example.

### Logistic Regression (binary, categorical response variable):

SAS **Code your binary variables as yes = 1 and no = 2**

```
proc logistic; class CategExplanatoryVar CategThirdVar;
model BinaryResponseVar=CategExplanatoryVar CategThirdVar
QuantThirdVar;
```

R

```
my.logreg <- glm(BinaryResponseVar ~ ExplanatoryVar + ThirdVar1 +
ThirdVar2, data = myData, family = "binomial")
summary(my.logreg) # for p-values
exp(my.logreg$coefficients) # for odds ratios
exp(confint(my.logreg)) # for confidence intervals on the odds ratios
```

PYTHON

```
import statsmodels.api
import statsmodels.formula.api as smf

# logistic regression
lreg1 = smf.logit(formula = 'BinaryResponseVar ~ ExplanatoryVar +
ThirdVar1 + ThirdVar2', data = myData).fit()
print (lreg1.summary())

# odd ratios with 95% confidence intervals
params = lreg1.params
conf = lreg1.conf_int()
conf['OR'] = params
conf.columns = ['Lower CI', 'Upper CI', 'OR']
print (numpy.exp(conf))
```

STATA

```
logistic BinaryResponseVar ExplanatoryVar ThirdVar1 ThirdVar2
```

SPSS

```
LOGISTIC REGRESSION BinaryResponseVar with ExplanatoryVar
ThirdVar1 ThirdVar2.
```



# Poster Presentations

---

**“Numbers have an important story to tell. They rely on you to give them a clear and convincing voice.”**

**—Stephen Few**

## Overview

You have conducted a quantitative research project. Now you will learn how to present your results as a research poster and presentation. Posters offer the opportunity to engage with an audience and to meaningfully disseminate your research findings to others.

## Lesson

Learn keys to a successful poster presentation. Consider your audience and frame your research question and results in an understandable and interesting way. Understand you should be brief, use large font size, and incorporate graphics instead of text whenever possible. See how being clear and concise with a logical layout will ensure the viewing experience is intellectually and esthetically satisfying for your audience. Click **HERE** to watch the video lesson.

Here are links to poster templates that you can modify.

PowerPoint templates:      **Template 1; 3 Column; 40" x 36"**  
                                         **Template 2; 3 Column, 40" x 36"**  
                                         **Template 3; 3 Column, 40" x 36"**

## The Association between Smoking Quantity and Nicotine Dependence among Young Adults with and without Psychiatric Disorders

Lisa Dierker, Quantitative Analysis Center, Wesleyan University

### Introduction

- Psychiatric disorders are a potent group of risk factors consistently implicated in the development of nicotine dependence (Rahde et al., 2003; 2004).
- While the association has been well established in the literature, less is known about the ways in which psychiatric disorders may play a role in the emergence of nicotine dependence.
- Most research has focused on how psychiatric disorders and may be a risk factor for heavy smoking e.g. the self medication hypothesis (Khantzian, 1997).
- Alternately, however, psychiatric disorders may signal a greater sensitivity to nicotine dependence at low levels of smoking (i.e. individuals with psychiatric disorders may develop nicotine dependence symptoms at lower levels of smoking than those without psychiatric disorders).

### Research Questions

- Which psychiatric disorders are associated with nicotine dependence after controlling for comorbidity?
- Does the association between smoking quantity and nicotine dependence differ for individuals with and without a psychiatric disorders?

### Methods

**Sample**

- Young adults (age 18 to 25) who reported daily smoking in the past year (n=1320) were drawn from the first wave of the National Epidemiologic Study of Alcohol and Related Conditions (NESARC).
- NESARC is a nationally representative sample of non-institutionalized adults in the U.S.

**Measures**

- Lifetime psychiatric disorders were assessed using the NAAA, Alcohol Use Disorder and Associated Disabilities Interview Schedule – DSM-IV (AUDADIS-IV).
- The tobacco module includes questions on symptom criteria for DSM-IV nicotine dependence.
- Current smoking was evaluated through quantity ("On the days that you smoked in the last year, about how many cigarettes did you usually smoke?").

### Results

**Univariate**

- Fully 61% of daily young adult smokers met criteria for DSM-IV nicotine dependence in the past year.
- A total of 55% met criteria for one or more psychiatric disorders.

**Bivariate**

- Chi-Square analysis showed that daily, young adult smokers with a psychiatric disorder were significantly more likely to meet criteria for nicotine dependence (78.5%) than those without a psychiatric disorder (60.7%),  $\chi^2=152.3, 1 \text{ df}, p<.0001$ . The most common disorder among daily smokers was alcohol dependence (45% SE 1.9).
- As expected, the number of cigarettes smoked per day was significantly associated with DSM-IV nicotine dependence,  $OR=1.04 (1.03-1.06)$ .

**Multivariate**

- Major depression (MDD), specific phobia, alcohol dependence, and antisocial personality disorder (ASPD) were each associated with DSM-IV nicotine dependence after controlling for comorbidity.
- The interaction between number of cigarettes smoked per day four specific phobia, ASPD and MDD was not found to be significantly associated with the presence of nicotine dependence.
- At each level of use, the probability of nicotine dependence is significantly higher among those with the disorder than those without (Figure 1).
- In contrast, the interaction between alcohol dependence and number of cigarettes smoked per day was statistically significant when predicting nicotine dependence (Figure 2).

### Discussion

- Individuals with major depression, specific phobia and ASPD may be more sensitive to nicotine dependence across levels of smoking.
- Individuals with alcohol dependence are more sensitive to nicotine dependence at low levels of daily smoking, but not at the highest levels when compared to individuals without alcohol dependence.
- Notably, the present findings are based on cross-sectional data and do not reflect the smoking levels at which nicotine dependence emerges among those with and without psychiatric disorders.
- Further research is needed to determine whether sensitivity to nicotine dependence is based on physical and/or psychological differences related to psychiatric disorders.

### References

American Psychiatric Association. (2013). The self-administered questionnaire of alcohol use disorders in adolescence and young adulthood. *Journal of Abnormal Psychology, 122*, 20-21.

Rahde, M., Sorensen, E. W., Smith, E. A., Fong, L. W., & Cohen, S. M. (2004). Association between lifetime and current smoking, comorbidity with smoking, nicotine, nicotine dependence, and nicotine withdrawal. *Journal of Abnormal Psychology, 113*, 53-59.

Khantzian, D. J. (1997). The self-medication hypothesis of substance use disorders: A conceptualization. *Journal of Abnormal Psychology, 106*, 49-67.

**Figure 1. Smoking and Nicotine Dependence by Major Depression among Daily, Young Adult Smokers**

**Figure 2. Smoking and Nicotine Dependence by Alcohol Use Disorders among Daily, Young Adult Smokers**

## Complete a Research Poster

Use one of the poster templates to create a poster presenting your introduction, research questions, methods, results and discussion. Click [HERE](#) for an example.



# Acknowledgments



© 2019 Creative Commons, Attribution Non-Commercial - ShareAlike.

The development of this course was supported by grants #0942246, #1323084, and #1820766 from the National Science Foundation, Improving Undergraduate STEM Education (IUSE). See [www.passiondrivenstatistics.com](http://www.passiondrivenstatistics.com) for more information. Background statistical content is based on materials from the Carnegie Mellon Open Learning Initiative [www.oli.cmu.edu](http://www.oli.cmu.edu).