Chapter 2 – Relationships between Categorical Variables

Introduction: An important field of exploration when analyzing data is the study of relationships between variables. A lot of thought has been put into determining which variables have relationships and the scope of that relationship. Is a person's diet related to having high blood pressure? Is the city a person lives in related to whether or not they have tuberculosis? Is being in a car accident related to texting while driving? These are all important questions that statisticians, data analysts and data scientists explore.

Relationships can be categorical ⇔ categorical, categorical ⇔ quantitative, and quantitative ⇔ quantitative. In this chapter, we will begin to explore the relationships between two categorical variables.

Remember, statistics is a deep well of mathematics and knowledge learned by years of study. There are much more advanced techniques for studying relationships, but we will be focusing on a basic introduction to the topic. You will find that a good understanding of this chapter will help tremendously when you go on to the more advanced techniques later on. For example, I find my students have many problems understanding the Chi-Squared distribution because they lack the foundational understanding of two-way tables and analyzing differences between categories.

Note on Terminology: *When studying relationships between variables you will hear different words used to describe the relationship. The most common are "relationship", "association", or "correlation". "Correlation" is often used for describe a relationship between two quantitative variables (quantitative ⇔ quantitative), while "relationship" and "association" are used for two categorical variables (categorical ⇔ categorical) or for a categorical - quantitative relationship study (categorical ⇔ quantitative).*

In this chapter, we will be using the terms "relationship" and "association".

Note on Causation: *One of the most famous statements in statistics is that "correlation is not causation". Proving that one thing causes another is a much more complex kind of study and involves controlling confounding variables and experimental design. The main thing to remember is that just because there is a relationship, that does not prove causation. There may be many other factors involved.*

Section 2A – Two-Way Tables with Technology

When studying relationships between categorical variables, we start with a two-way table. A two-way table is a summary of counts or frequencies for two categorical data sets. Let us look at the hospital data again from the last chapter.

Example 1

Patient ID#	Age	Gender	Blood Type	Rh Factor	Floor
1	23	M	A	-	SDS
2	68	M	O	+	ER
3	51	F	AB	+	Med/Surg
4	74	M	O	-	ICU
5	49	F	O	+	SDS
6	62	F	O	+	Med/Surg
7	35	M	A	+	SDS
8	46	F	O	+	Med/Surg
9	72	F	O	+	ER
10	61	M	B	+	SDS
11	43	F	A	-	Med/Surg
12	81	M	O	+	ICU
13	65	M	A	+	Med/Surg
14	59	F	O	-	SDS
15	44	F	B	+	ICU
16	26	M	O	+	ER
17	58	F	AB	-	ER
18	45	M	O	+	SDS
19	55	M	O	+	Med/Surg
20	71	M	A	+	ER

Suppose we want to analyze the relationship and proportions for a patient's gender and their blood type. Notice gender is one categorical data set with two variables (male and female). Blood type is another categorical data set with four variables (A, B, AB, and O). To make a two-way table, pick one of the variables to be the row and the column. I am going to pick gender to be my rows and blood type to be my columns. Since there are two options for the rows and four options for the columns, we will have a "2 by 4" table (2 rows and 4 columns, not counting totals).

	Type A	Type B	Type AB	Type O
Female				
Male				

Now we just need to count and fill out the table. It should be noted that no data analyst or statistician does this by hand. All use either excel or a statistics software. Remember we live in the age of "big data". No one wants to count variables in a data set with twenty thousand values, and that is not even "big".

Since we are introducing the topic, see if you can count the amount for each box. You can use tally marks if you wish. Where the "Female" row meets the "Type A" column we should put how many female patients had type A blood. (There was only one.) Where the "Male" row meets the "Type O" column we should put how many male patients had type O blood. (There was six.)

	Type A	Type B	Type AB	Type O
Female	1			
Male				6

See if you can find the rest of the counts (frequencies) for the table.

You should get the following table. There were twenty patients so the numbers in the two-way table should add up to twenty. This is called the "grand total". Also, notice there were no males with type AB blood, so we needed to put a zero in that cell.

	Type A	Type B	Type AB	Type O
Female	1	1	2	5
Male	4	1	0	6

Before we can analyze the relationship and proportions, we need to calculate all the row and column totals. This is automatically done with excel or statistics software programs. Notice the "grand total" is always in the bottom right corner of the table. Keep in mind that this is still considered a two-by-four table. Totals are not included in the size of a table.

	Type A	Type B	Type AB	Type O	Total
Female	1	1	2	5	9
Male	4	1	0	6	11
Total	5	2	2	11	Grand Total = 20

Notice a few things about this table. The row totals (9 and 11) add up to the grand total (20). Also the column totals (5, 2, 2, and 11) add up to the grand total. Be careful. The row totals plus the column totals does not add up to the grand total.

Creating a two way table with technology

In Statcato, copy and paste the data for the groups and the variable you want to compare. This will usually involve two columns of categorical data. Then go to the "Statistics" menu and click on "Multinomial Experiments". Now click on "Cross Tabulation and Chi-Square". Pick one column of data to be the row and the other column of data as the column. Uncheck the box that says, "Perform chi-squared test". That is a more advanced analysis. Also, do not click on anything under the "frequency (optional)" menu. Now push "OK".

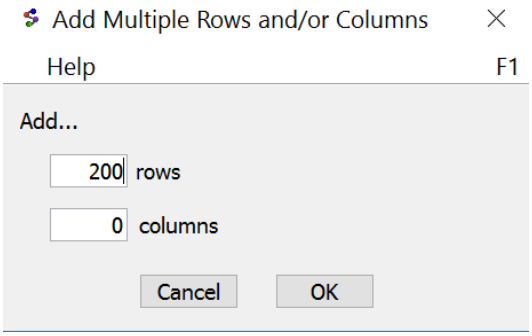
Statistics => Multinomial Experiments => Cross Tabulation => OK

Example 2

Suppose we want to make a two-way table describing gender and type of transportation to college using the math 075-survey data fall 2015. I started by opening the math 075 survey data 2015 in Excel. I then copy and pasted the gender column and the transportation column into Statcato.

Important Reminder: If your data set is over 300 entries, you will need to add some rows to Statcato. The math 075-survey data had close to 500 students, so we will need to add some rows to the spreadsheet in Statcato before copy and pasting from Excel. (I added 200 more rows to Statcato before I tried to copy and paste.)

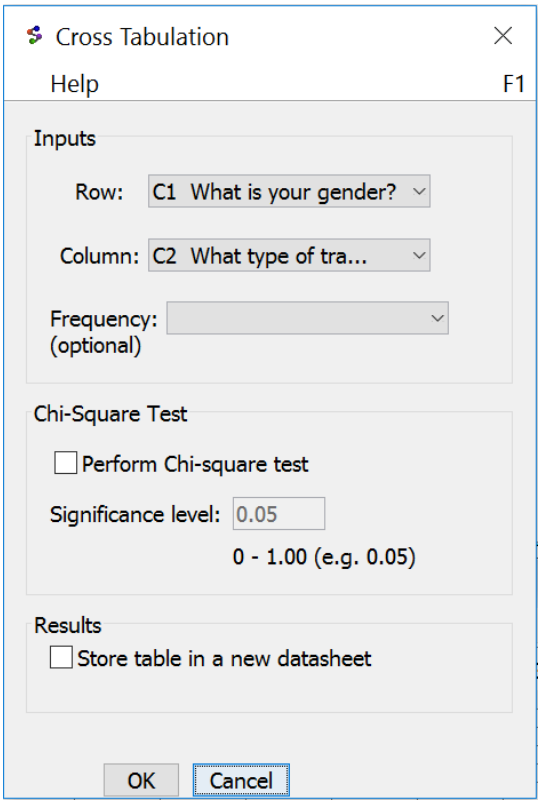
Edit => Add Multiple Rows/Columns => Put how many rows in box => OK



I then went to the statistics menu, multinomial experiments and cross tabulation. You can put either variable as the row or column. I put the gender as my row variable and the transportation as the column variable and pushed "OK".

Statistics => Multinomial Experiments => Cross Tabulation and Chi Square => Pick Row and Columns => OK

Note: Do not check the box that says, "Perform chi-square test". This is a more advanced analysis you may learn in future statistics classes. You also do not need to select the frequency (optional) box.



Cross Tabulation and Chi-Square

rows in C1 What is your gender?, columns in C2 What type of tra...

	Bicycle	Carpool	Drive alone	Dropped off by someone	Other	Public transportation	Skate	Walk	All
Female	1.0 (0.21%)	24.0 (4.99%)	197.0 (40.96%)	29.0 (6.03%)	1.0 (0.21%)	15.0 (3.12%)	0.0 (0%)	3.0 (0.62%)	270.0 (56.13%)
Male	1.0 (0.21%)	15.0 (3.12%)	162.0 (33.68%)	19.0 (3.95%)	3.0 (0.62%)	9.0 (1.87%)	1.0 (0.21%)	1.0 (0.21%)	211.0 (43.87%)
All	2.0 (0.42%)	39.0 (8.11%)	359.0 (74.64%)	48.0 (9.98%)	4.0 (0.83%)	24.0 (4.99%)	1.0 (0.21%)	4.0 (0.83%)	481.0 (100.00%)

Notice that the totals are found in the row and column that say "All". What size is this two-way table? Remember the totals are not included in the size of a two-way table, only the number of variables in each categorical data set. Since gender has two options (male or female) and

transportation had eight options (Bicycle, Carpool, Drive alone, Dropped off by someone, Other, Public transportation, Skate, or Walk), this is considered a "two by eight" table.

Problem Set Section 2A

Directions: Here is some data taken from the medical records department at a local hospital. The data includes age, gender, blood type (A, B, AB, O), Rhesus factor (Rh + or Rh -) and part of the hospital the patient was in (Medical/Surgical, Intensive Care Unit , Same Day Surgery, Emergency Room).

Patient ID#	Age	Gender	Blood Type	Rh Factor	Floor
1	23	M	A	-	SDS
2	68	M	O	+	ER
3	51	F	AB	+	Med/Surg
4	74	M	O	-	ICU
5	49	F	O	+	SDS
6	62	F	O	+	Med/Surg
7	35	M	A	+	SDS
8	46	F	O	+	Med/Surg
9	72	F	O	+	ER
10	61	M	B	+	SDS
11	43	F	A	-	Med/Surg
12	81	M	O	+	ICU
13	65	M	A	+	Med/Surg
14	59	F	O	-	SDS
15	44	F	B	+	ICU
16	26	M	O	+	ER
17	58	F	AB	-	ER
18	45	M	O	+	SDS
19	55	M	O	+	Med/Surg
20	71	M	A	+	ER

1. Create a two-way table that we could use to compare gender to the Rh factor. Include the grand total and all of the row and column totals. What is the size of the table (# rows by # columns)?

2. Create a two-way table that we could use to compare blood type to Rh factor. Include the grand total and all of the row and column totals. What is the size of the table (# rows by # columns)?

3. Create a two-way table that we could use to compare gender to the floor the patient went to. Include the grand total and all of the row and column totals. What is the size of the table (# rows by # columns)?

4. Create a two-way table that we could use to compare the Rh factor the floor the patient went to. Include the grand total and all of the row and column totals. What is the size of the table (# rows by # columns)?

5. Create a two-way table that we could use to compare the blood type to the floor the patient went to. Include the grand total and all of the row and column totals. What is the size of the table (# rows by # columns)?

Sometimes we can make a category out of quantitative data. For example. The patient's age are numbers that measure something and have units, so it is quantitative. However, we could separate the ages into ranges that could work like a category. The age ranges are arbitrary and you can make the ranges yourself. For example, we could use three categories: 40 or under, 41-59 years old, 60 years or older.

6. Create a two-way table that we could use to compare the gender to the patient's age (40 or under, 41-59, 60 or above). Include the grand total and all of the row and column totals. What is the size of the table (# rows by # columns)?

7. Create a two-way table that we could use to compare the part of the hospital to the patient's age (40 or under, 41-59, 60 or above). Include the grand total and all of the row and column totals. What is the size of the table (# rows by # columns)?

Directions: Use the fall 2015 math 075 survey data and Statcato to create a two-way table for the following variables. Do not forget to add additional rows to Statcato before copy and pasting so the entire data set fits.

Adding Rows in Statcato: *Edit => Add Multiple Rows/Columns => Put how many rows in box => OK*

Making a two-way table: *Statistics => Multinomial Experiments => Cross Tabulation and Chi Square => Pick Row and Columns => uncheck "perform chi-square" => OK*

Note: *Often when people fill out a survey or give data, they may not answer all the questions. To make a two-way table from two categorical data sets, we need to use those people that answered both questions. If a value is left blank, Statcato will say, "The number of row labels and the number of column labels need to be the same". This means that we need to go through and delete out anyone that did not answer both questions. This is often called "cleaning the data".*

8. Use the math 075-survey data fall 2015 to create a two-way table with campus (Valencia or Canyon Country) and gender (male or female). Include the grand total and all of the row and column totals. You may need to delete out people that did not answer both questions, i.e. clean the data. What is the size of the table (# rows by # columns)? Copy and paste the table into a word document or draw the two-way table on a piece of paper. You do not need to include the percentages given in Statcato, just the frequencies and totals.

9. Use the math 075 survey data fall 2015 to create a two-way table with contact lenses or glasses (yes or no) and hair color (brown, black, blond(e), red, other). Include the grand total and all of the row and column totals. You may need to delete out people that did not answer both questions, i.e. clean the data. What is the size of the table (# rows by # columns)? Copy and paste the table into a word document or draw the two-way table on a piece of paper. You do not need to include the percentages given in Statcato, just the frequencies and totals.

10. Use the math 075 survey data fall 2015 to create a two-way table with texting while driving (yes or no) and being in a car accident (yes or no). Include the grand total and all of the row and column totals. You may need to delete out people that did not answer both questions, i.e. clean the data. What is the size of the table (# rows by # columns)? Copy and paste the table into a word document or draw the two-way table on a piece of paper. You do not need to include the percentages given in Statcato, just the frequencies and totals.

11. Use the math 075 survey data fall 2015 to create a two-way table with smoking cigarettes (yes or no) and political party (democrat, republican, independent, other). Include the grand total and all of the row and column totals. You may need to delete out people that did not answer both questions, i.e. clean the data. What is the size of the table (# rows by # columns)? Copy and paste the table into a word document or draw the two-way table on a piece of paper. You do not need to include the percentages given in Statcato, just the frequencies and totals.

Section 2B – Using Bar Charts and Pie Charts to Summarize Two-Way Tables

Bar Charts and Pie Charts can be very useful to summarize two-way table information. Once we have created a two-way table, we may wish to compare the counts (frequencies) or percentages for the two categorical data sets.

The following two-way table was created using Statcato and the Fall 2015 Math 075 Survey data and describes the political party and whether or not a person smokes cigarettes.

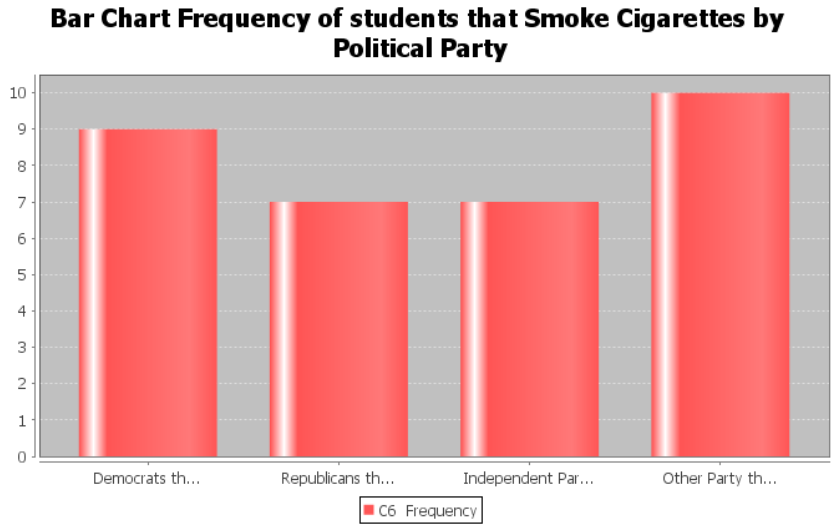
	Democratic	Independent	Other	Republican	All
No Cigarettes	184.0	82.0	87.0	93.0	446.0
Yes Smokes Cigarettes	9.0	7.0	10.0	7.0	33.0
All	193.0	89.0	97.0	100.0	479.0

There are many questions we can answer from this table and several different graphs we can use. Side by side bar charts are very useful for comparing variables.

	Frequency
Democrats that smoke	9
Republicans that smoke	7
Independent Party that smoke	7
Other Party that smoke	10

This information can be summarized nicely by a bar chart.

To Create a Bar Chart with Statcato: *Graph => Bar Chart => Select Column with counts (Add Series) => Select column with categories (names) => OK*

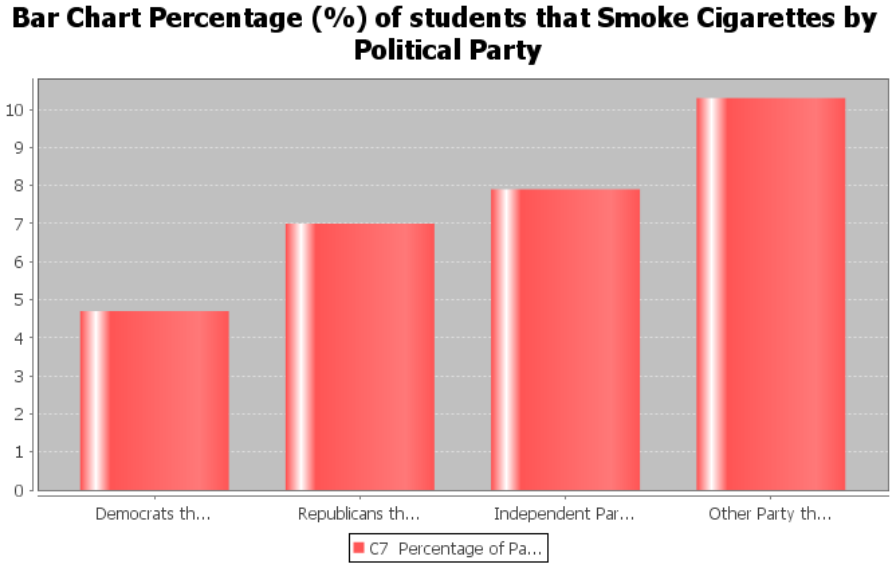


This graph can be misleading since we are not considering the totals for each of these parties. Dividing the number of people that smoke cigarettes by the total for each political party gives us the following:

	Frequency	Percentage of Party (%)
Democrats that smoke	9	9 / 193 x 100% ≈ 4.7%
Republicans that smoke	7	7 / 100 x 100% ≈ 7.0%
Independent Party that smoke	7	7 / 89 x 100% ≈ 7.9%
Other Party that smoke	10	10 / 97 x 100% ≈ 10.3%

When typing the percentages into Statcato, remember to not put the % symbol. We can label it as a percentage.

	Frequency	Percentage of Party (%)
Democrats that smoke	9	4.7
Republicans that smoke	7	7.0
Independent Party that smoke	7	7.9
Other Party that smoke	10	10.3



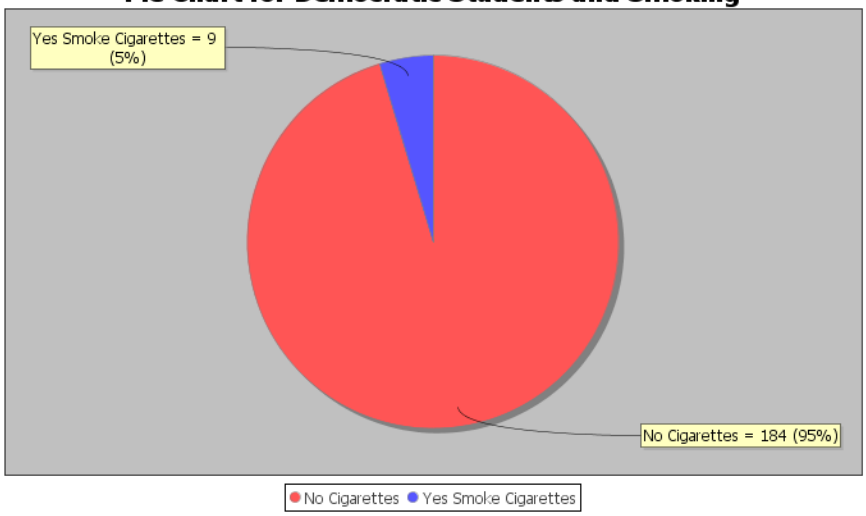
Notice that even though there were more democrats that smoke than republicans, the percentage of democrats that smoke was actually less than the percentage of republicans.

Multiple Pie Charts are another way to summarize two-way table data. We can make a pie chart for each political party showing both the smoking and non-smoking students. For each pie chart, we will use a single column where the counts (frequencies) are. Do not forget to push the "add series". Then use the cigarettes as the categories.

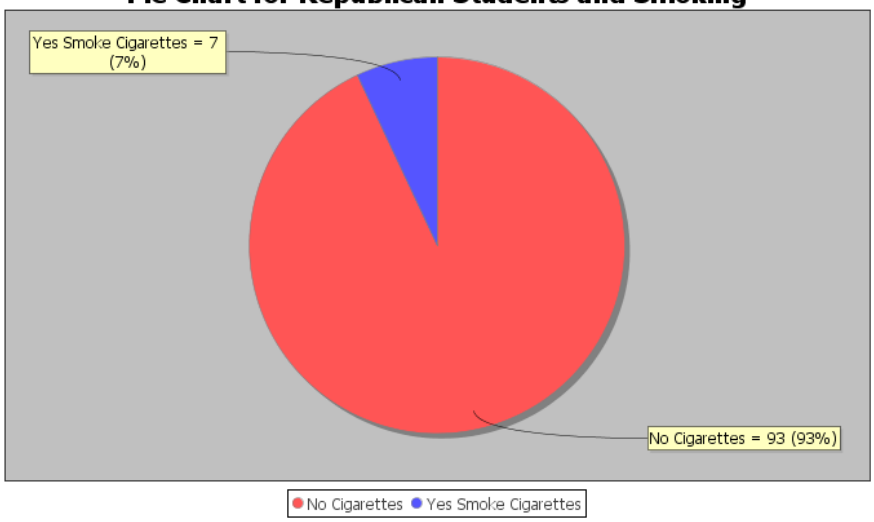
	Democratic	Independent	Other	Republican
No Cigarettes	184.0	82.0	87.0	93.0
Yes Smokes Cigarettes	9.0	7.0	10.0	7.0

To Create a Pie Chart with Statcato: *Graph => Pie Chart => Select Column with counts (Add Series) => Select column with categories (names) => OK*

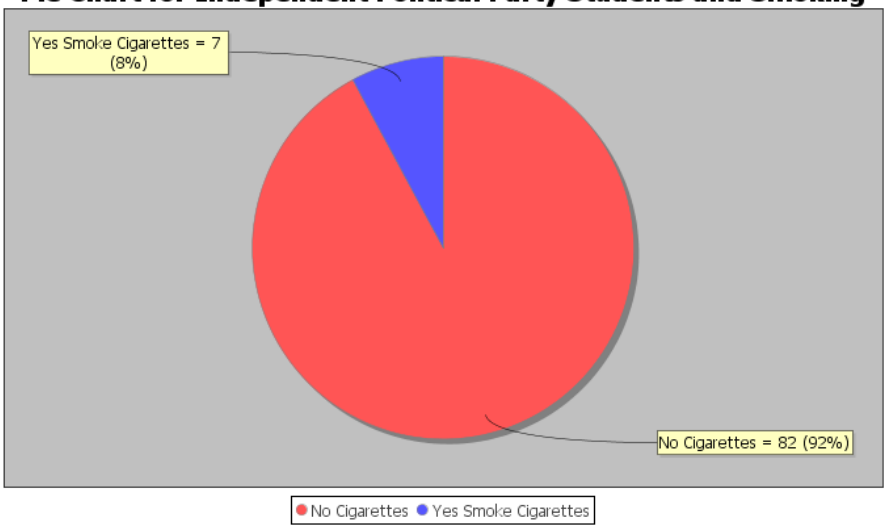
Pie Chart for Democratic Students and Smoking



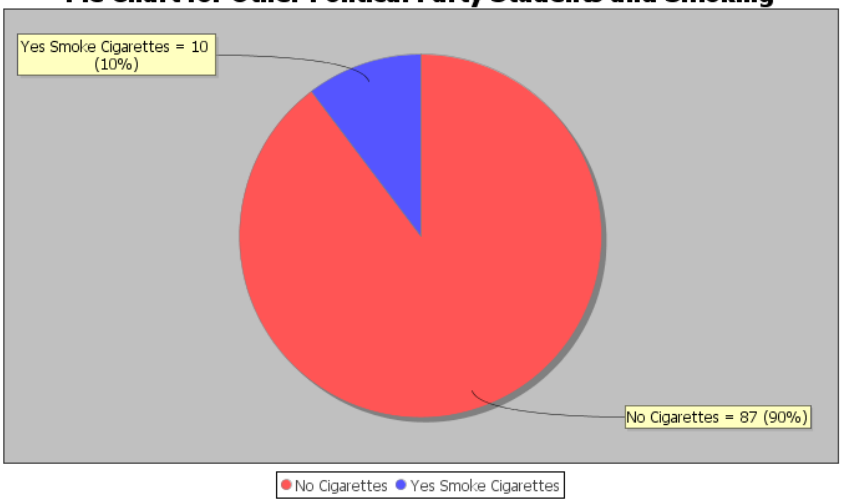
Pie Chart for Republican Students and Smoking



Pie Chart for Independent Political Party Students and Smoking



Pie Chart for Other Political Party Students and Smoking



We can see from these pie charts that the "other" political party has the highest percentage of smokers.

Problem Set Section 2B

Directions: Here is some data taken from the medical records department at a local hospital. The data includes age, gender, blood type (A, B, AB, O), Rhesus factor (Rh + or Rh -) and part of the hospital the patient was in (Medical/Surgical, Intensive Care Unit , Same Day Surgery, Emergency Room).

Patient ID#	Age	Gender	Blood Type	Rh Factor	Floor
1	23	M	A	-	SDS
2	68	M	O	+	ER
3	51	F	AB	+	Med/Surg
4	74	M	O	-	ICU
5	49	F	O	+	SDS
6	62	F	O	+	Med/Surg
7	35	M	A	+	SDS
8	46	F	O	+	Med/Surg
9	72	F	O	+	ER
10	61	M	B	+	SDS
11	43	F	A	-	Med/Surg
12	81	M	O	+	ICU
13	65	M	A	+	Med/Surg
14	59	F	O	-	SDS
15	44	F	B	+	ICU
16	26	M	O	+	ER
17	58	F	AB	-	ER
18	45	M	O	+	SDS
19	55	M	O	+	Med/Surg
20	71	M	A	+	ER

1. Use the two-way table for gender and Rh factor that you created in problem set 2A. Make a side-by-side bar plot for males and females showing the number of Rh+ patients. Make a side-by-side bar plot for males and females showing the number of Rh- patients. Make a pie chart for the female patients showing both Rh+ and Rh- . Make a pie chart for the male patients showing both Rh+ and Rh- .

2. Use the two-way table for blood type and Rh factor that you created in problem set 2A. Make a side by side bar plot comparing the number of Rh+ patients for each of the four blood types (A,B,AB, and O). Make a side by side bar plot comparing the number of Rh- patients for each of the four blood types (A, B, AB, and O). Make a pie chart for the type A patients showing both Rh+ and Rh- . Make a pie chart for the type B patients showing both Rh+ and Rh- . Make a pie chart for the type AB patients showing both Rh+ and Rh- . Make a pie chart for the type O patients showing both Rh+ and Rh- .

3. Use the two-way table that you created in problem set 2A describing gender and floor the patient went to. Make a side-by-side bar plot for females showing the number females that went to each floor. Make a side-by-side bar plot for males showing the number males that went to each floor. Make pie charts for the each floor giving the frequency and percentage of males and females.

4. Use the two-way table with campus (Valencia or Canyon Country) and gender (male or female) that you created in problem set 2A from the Math 075 Survey Data Fall 2015. Create a side-by-side bar chart with frequencies that gives the gender make up for the Valencia campus. Create a side-by-side bar chart with percentages that gives the gender make up for the Valencia campus. Create a side-by-side bar chart with frequencies that gives the gender make up for the Canyon Country campus. Create a side-by-side bar chart with percentages that gives the gender make up for the Canyon Country campus. Create a pie chart with frequencies and percentages for the female students only describing the two campuses. Create a pie chart with frequencies and percentages for the male students only describing the two campuses.

5. Use the two-way table with contact lenses or glasses (yes or no) and hair color (brown, black, blond(e), red, other) that you created in problem set 2A from the Math 075 Survey Data Fall 2015. Create a side-by-side bar chart describing the frequencies of hair color for students that wear contacts or glasses. Create a side-by-side bar chart describing the percentages of each hair color for students that wear contacts or glasses. Create a pie chart with frequencies and percentages describing glasses/contact information for just students with brown hair. Create a pie chart with frequencies and percentages describing glasses/contact information for just students with black hair. Create a pie chart with frequencies and percentages describing glasses/contact information for just students with blond hair. Create a pie chart with

frequencies and percentages describing glasses/contact information for just students with red hair.

6. Use the two-way table with texting while driving (yes or no) and being in a car accident (yes or no) that you created in problem set 2A from the Math 075 Survey Data Fall 2015. Create a side by side bar plot for only students that text and drive with the frequencies for car accidents and no car accidents. Create a side by side bar plot for only students that text and drive with the percentages for car accidents and no car accidents. Create a side by side bar plot for only students that do not text and drive with the frequencies for car accidents and no car accidents. Create a side by side bar plot for only students that do not text and drive with the percentages for car accidents and no car accidents. Create a pie chart for students that have been in a car accident describing the frequencies and percentages for students that do and do not text and drive. Create a pie chart for students that have not been in a car accident describing the frequencies and percentages for students that do and do not text and drive.

Section 2C – Marginal and Joint Percentages from Two-Way Tables

Analyzing two categorical data sets involves not only creating two-way tables, bar charts and pie charts, but also being able to find and analyze proportions and percentages.

Remember that a proportion is found by taking the amount (frequency) and dividing by the total.

$$\text{Proportion} = \frac{\text{Amount (Frequency)}}{\text{Total}}$$

To convert that proportion into a percentage, simply multiply the decimal proportion by 100%.

Basic Marginal Percentages

Let us start with looking at basic marginal proportions. These are proportions where the amount involves only a single variable and the total is everyone in the data (grand total).

Look at the following two-way table created from the Fall 2015 Math 075 Survey data. This table describes the relationship between smoking and political party.

	Democratic	Independent	Other	Republican	All
No Cigarettes	184.0	82.0	87.0	93.0	446.0
Yes Smokes Cigarettes	9.0	7.0	10.0	7.0	33.0
All	193.0	89.0	97.0	100.0	479.0

Remember, analyzing data involves asking questions and finding the answers to those questions.

For example. Here are a few questions that came to mind when I looked at this two-way table.

Example 1

What percentage of the students smoke cigarettes?

Notice we are looking at all of the students (not just democrats), so we should use the grand total as our total.

Proportion of students that smoke = Amount of Smokers / Grand Total = 33 / 479 = 0.0688935 ≈ 0.069

Percentage of students that smoke ≈ 0.069 x 100% = 6.9%

Notice the amount of smokers was found at the far right of the "Yes Smokes Cigarettes" row. This amount is found in the margin. This is why it is sometimes called a marginal proportion.

Example 2

What percentage of the students identified as other political party?

Notice we are looking at all of the students (not just smokers), so we should use the grand total again as our total.

Proportion of students that are other political party = Amount of other political party / Grand Total = 97 / 479 = 0.2025052 ≈ 0.203

Percentage of students that are other political party ≈ 0.203 x 100% = 20.3%

Notice the amount of students that identified as other political party was found at the bottom of the "other" column.

Joint Percentages

Sometimes we want to find a proportion or percentages where the amount (frequency) involves more than one variable. These are often called "joint proportions" or "joint percentages".

There are two types of joint proportions.

AND: This is when we want to know the proportion or percentage involving two things being true about a person or object.

OR: This is when we want to know the proportion of percentage involving either one variable or another variable being true about the person or object.

Let us look at the political party and cigarette data again.

Example 3

What percentage of all the students both smoked and were Republican?

Notice there are two variables involved, republican and smoking. The key though is that we want the proportion for both things being true about the person. We cannot look at only smokers and we cannot look at republicans. We need the amount of smoking republicans. This is a classic "AND" proportion since both things need to true about the student.

Notice also we are picking from all students, so our total should be the grand total again.

	Democratic	Independent	Other	Republican	All
No Cigarettes	184.0	82.0	87.0	93.0	446.0
Yes Smokes Cigarettes	9.0	7.0	10.0	7.0	33.0
All	193.0	89.0	97.0	100.0	479.0

Proportion of smoking republicans = amount of smoking republicans / grand total

= 7 / 479 ≈ 0.014613778 ≈ 0.015

Percentage of students that both smoke and are republican ≈ 0.015 x 100% ≈ 1.5%

Notice the amount of students that both smoke and are republican can be found in a single cell where the Republican column meets the Yes Smoking row.

Example 4

Suppose we only wanted to know the percentage of students that either smoke or are republican. (Not both)

This would be a classic "OR" joint proportion. The key is that we will now need to include everyone that smokes, as well as everyone that is republican. When calculating an "OR" joint proportion, you will need to do some adding to find the amount.

	Democratic	Independent	Other	Republican	All
No Cigarettes	184.0	82.0	87.0	93.0	446.0
Yes Smokes Cigarettes	9.0	7.0	10.0	7.0	33.0
All	193.0	89.0	97.0	100.0	479.0

Proportion of students that either smoke or are republican =
amount of students that either smoke or are republican / grand total

= (93 + 7 + 10 + 7 + 9) / 479 = 126 / 479 ≈ 0.263048 ≈ 0.263

Percentage of students that either smoke or are republican ≈ 0.263 x 100% ≈ 26.3%

Important Note: Notice that we did not use the row and column totals when calculating an "OR" joint proportion. If we added the total for smokers (33) plus the total for republicans (100), we would have gotten 133 as our amount. This would be wrong. The correct amount was 126. Adding the row and column totals gives you the wrong answer because we would have added the 7 smoking republicans twice. It is best not to use the row and column totals.

Problem Set Section 2C

Directions: The following two-way table was created from the Math 075 Survey Data Fall 2015 and describes the student's favorite social media and whether or not they have a tattoo. Use the table to find the given proportions and percentages. Write your answers as a fraction, decimal and as a percentage.

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}} \qquad \text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

To convert proportion into percentage, multiply by 100%.

	Facebook	Instagram	Other	Snapchat	Twitter	Total
No Tattoo	68	103	31	84	80	366
Yes Has a Tattoo	33	39	10	16	13	111
Total	101	142	41	100	93	Grand Total = 477

Basic Proportions

1. What percent of the students have a tattoo?

2. What proportion of the students prefer Snapchat?

3. What percent of the students do not have a tattoo?

4. What proportion of the students prefer Instagram?

Joint Proportions "AND"

5. What percent of the students both have a tattoo and prefer Facebook?

6. What proportion of the students do not have a tattoo and prefer Twitter?

7. What percent of the students do not have a tattoo and prefer Snapchat?

8. What proportion of the students both have a tattoo and prefer a different social media (Other)?

Joint Proportions "OR"

9. What percent of the students either have a tattoo or prefer Instagram?

10. What proportion of the students prefer either Twitter or Snapchat?

11. What percent of the students either do not have a tattoo or prefer Facebook?

12. What proportion of the students prefer either Facebook or Twitter?

Section 2D – Conditional Percentages and Categorical Relationships

When studying relationships between categorical variables, we focus on looking at conditional proportions. A conditional proportion is looking at a particular part of the two-way table and not everyone. So let us look at finding conditional proportions first and then we can see what these tell us about categorical relationships.

Conditional Percentages

A condition is having prior information about the variable. It usually involves the words "IF" or "Given" or "Out of".

Look at the following two-way table from the Fall 2015 Math 075 Survey data describing whether or not students have tattoo and what social media they prefer.

	Facebook	Instagram	Other	Snapchat	Twitter	Total
No Tattoo	68	103	31	84	80	366
Yes Has a Tattoo	33	39	10	16	13	111
Total	101	142	41	100	93	Grand Total = 477

Example 1

What percentage of the students have a tattoo. This has no condition since we are picking from all the students. Hence, we use the grand total.

Proportion of students with tattoo = 111 / 477 ≈ 0.233 ≈ 23.3%

What if I want to look only at the students that prefer Instagram? This is now a condition. I want to know the percentage of the Instagram students that have a tattoo.

The key to finding a conditional proportion is to circle the row or column that has your condition. In this case, I want to only use the row or column with the students that prefer Instagram. I always recommend circling or highlighting the row or column with your condition.

Note: To calculate a conditional proportion, circle the row or column with your condition. Then only use the amount and total in that row or column.

	Facebook	Instagram	Other	Snapchat	Twitter	Total
No Tattoo	68	103	31	84	80	366
Yes Has a Tattoo	33	39	10	16	13	111
Total	101	142	41	100	93	Grand Total = 477

Proportion of the Instagram students that have a tattoo =

Amount of Instagram students with a tattoo / total number of Instagram students

= 39 / 142 ≈ 0.275 ≈ 27.5%

Notice that we highlighted the row or column with the condition (Instagram) and only used numbers in that row or column. Also, notice that a condition can make a huge difference in the overall percentage. Only about 23.3% of all the students have a tattoo, but if we only look at the students that prefer Instagram, that percentage increased to 27.5%.

Conditional proportions may be worded in various ways. Look for the key words "If", "Given", or "Out of".

What percentage of the Instagram students have a tattoo?

If we only look at students that prefer Instagram, what percentage have a tattoo?

What percentage of students have a tattoo if we are given that the students prefer Instagram?

Example 2

Let us try another conditional proportion.

What proportion of the students with tattoos prefer Facebook?

Which variable is the condition? Look for the "if", "given" or "out of". In this problem, it says out of the students with tattoos.

Which variable are we finding the proportion for? Facebook

Remember: To calculate a conditional proportion, circle the row or column with your condition. Then only use the amount and total in that row or column.

	Facebook	Instagram	Other	Snapchat	Twitter	Total
No Tattoo	68	103	31	84	80	366
Yes Has a Tattoo	33	39	10	16	13	111
Total	101	142	41	100	93	Grand Total = 477

Proportion of students with tattoo that prefer Facebook =

Amount of Tattoo students that like Facebook / Total students with tattoo =

= 33 / 111 ≈ 0.297 ≈ 29.7%

Notice we only used the amount and total from the row or column with the condition.

Determining if Categorical Variables are Related

Determining relationships between variables can be complicated and there are more advanced techniques for determining the scope of the relationship. There is a guiding principle behind relationship studies though.

Relationship Principle

- **When values (conditional proportions) are significantly different, there is a relationship between the variables.**
- **When values (conditional proportions) are close, there is not a relationship between the variables.**

Example 3

Is liking Facebook related to having a Tattoo?

The key is to look at two proportions with the same variable, but with different conditions. For example, the percentage of tattoo students that like Facebook verses the percentage of no tattoo students that like Facebook.

Notice both are finding the percent that like Facebook, but the condition is different.

	Facebook	Instagram	Other	Snapchat	Twitter	Total
No Tattoo	68	103	31	84	80	366
Yes Has a Tattoo	33	39	10	16	13	111
Total	101	142	41	100	93	Grand Total = 477

Students with Tattoo: Percent that like Facebook = 33 / 111 ≈ 29.7%

Students without a Tattoo: Percent that like Facebook = 68 / 366 ≈ 18.6%

The closer the percentages, the less of a relationship there is. In a sense, if these percentages were very close (like 19% and 20%), then the condition does not matter. When that happens, we say there is no relationship or no association.

The farther apart the percentages, the more of a relationship there is. In this example, they do seem to be quite different. It appears that the students with tattoos that like Facebook is close to 10% higher than the students without a tattoo. We can also look at the percentage ratio.

Percentage Ratio = Higher Percentage / Lower Percentage = 29.7% / 18.6% ≈ 1.6

Remember if the ratio is close to 1, it is not very significant. This ratio is closer to 2, so indicating a significant difference.

Since the percentages are different, this indicates there is a relationship (or association) between liking Facebook and having a tattoo.

Does this mean that having a tattoo causes a person to like Facebook?

NO!!!

Remember Relationships (Associations, Correlations) do not prove Causation. There are many reasons behind a person liking Facebook other than having a tattoo!

Multiple Categorical Variable Study: Determining which variables are most related

The study of multiple variables is vital in statistics. Remember that a relationship does not prove cause and effect because there are usually many confounding variables involved. If we ever hope to understand cause and effect, we need to study these other variables and determine which ones have a strong relationship with our response variable and which do not.

How can we determine which categorical variables have the strongest relationship and which have very little relationship? There are many more advanced techniques for this including things like "P-value" and "Chi-squared", but again we are not at that level yet. Go back to the relationship principle and focus on the study of conditional proportions.

Relationship Principle

- **When values (conditional proportions) are significantly different, there is a relationship between the variables.**
- **When values (conditional proportions) are close, there is not a relationship between the variables.**

Example 4 – Multivariable Relationships

Let us look at the topic of car accidents. What variables might be related to a Math 075 student having a car accident? Which variables have the strongest or weakest relationship? We will use the Math 075 Survey Data to explore Campus Location, Gender, Texting while Driving, and Type of Transportation and see if we can identify which variables have the strongest relationship to having a car accident and which variables have the weakest relationship.

Cleaning the Data: I started by putting the data for car accident, campus, gender, texting while driving and transportation in a fresh excel spreadsheet. I went through and deleted any rows that had missing values for any of the five data sets. These people did not answer one of the questions. If you do not clean the data, Statcato may give an error message when you try to make the two-way table.

Let us see if location is related to a Math 075 student having a car accident. We can use technology to create a two-way table with Campus (location) and Car Accidents. The key is to compare conditional percentages that have the same variable for the percentage (car accidents) but a different condition (campus).

Start by adding some rows to Statcato. Go to the edit menu and then add rows or columns. This data set was close to 500 values so I added an additional 200 rows. To create a two-way table with Statcato from raw data, copy and paste the data into Statcato. Go to the statistics menu, then multinomial experiments, then cross tabulation and chi-square. Pick one of the variables to be the row and the other to be the column. You do not need a chi-squared test at this point.

	Canyon Country Campus	Valencia Campus	All
No Car Accident	149	236	385
Car Accident	41	52	93
All	190	288	478

Car accident percentage given the student went to Canyon Country = 41/190 x 100% ≈ 21.6%

Car accident percentage given the student went to Valencia = 52/288 x 100% ≈ 18.1%

Percentage ratio *(higher / lower)* = 21.6%/18.1% ≈ 1.2

	Female	Male	All
No Car Accident	216	169	385
Car Accident	52	41	93
All	268	210	478

Car accident percentage given the student was female = 52/268 x 100% ≈ 19.4%

Car accident percentage given the student was male = 41/210 x 100% ≈ 19.5%

Percentage ratio *(higher / lower)* = 19.5%/19.4% ≈ 1.005

	Bicycle	Carpool	Drive alone	Dropped off by someone	Other	Public transportation	Skate	Walk	All
No Car Accident	2	31	276	45	2	24	1	4	385
Car Accident	0	8	81	2	2	0	0	0	93
All	2	39	357	47	4	24	1	4	478

Note: The car accident question was asked if the student was driving, so we focused on the variables where the student could be driving.

Car accident percentage given the student carpools to school = 8/39 x 100% ≈ 20.5%

Car accident percentage given the student drives alone to school = 81/357 x 100% ≈ 22.7%

Percentage ratio *(higher / lower)* = 22.7%/20.5% ≈ 1.1

	Not Text and Drive	Does Text and Drive	All
No Car Accident	305	80	385
Car Accident	62	31	93
All	367	111	478

Car accident percentage given the student texts and drives = 31/111 x 100% ≈ 27.9%

Car accident percentage given the student does not text and drive = 62/367 x 100% ≈ 16.9%

Percentage ratio *(higher / lower)* = 27.9%/16.9% ≈ 1.65

Conclusion? When it comes to car accidents with the math 075 students, it seems that campus, gender and type of transportation are not very related. All three had conditional percentages that were very close with a ratio close to one. The most significant difference in the conditional percentages and the largest percentage ratio occurred with the texting while driving variable. Of the four explanatory variables we looked at, texting while driving had the strongest relationship with car accidents.

Side by Side Bar Charts

Conditional percentages in two-way tables can be summarized with a side-by-side bar chart (or split bar chart). It is a nice way of showing a lot of information on a single graph.

Look at the following two-way table for tattoos and favorite social media.

	Facebook	Instagram	Other	Snapchat	Twitter	Total
No Tattoo	68	103	31	84	80	366
Yes Has a Tattoo	33	39	10	16	13	111
Total	101	142	41	100	93	Grand Total = 477

A side-by-side bar chart can summarize all of this data. The graph can compare the percentage of Facebook users that have a tattoo and do not have a tattoo, but can also look at tattoos for Instagram, other, snapchat and twitter.

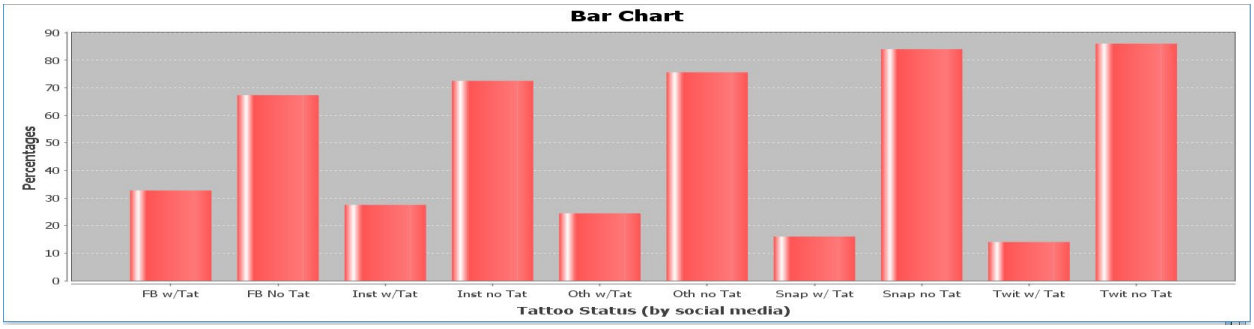
We will need to calculate these conditional percentages though and summarize them in a table.

	Percentage (%)
Facebook with Tattoo	33/101 ≈ 32.7%
Facebook No Tattoo	68/101 ≈ 67.3%
Instagram with Tattoo	39/142 ≈ 27.5%
Instagram No Tattoo	103/142 ≈ 72.5%
Other Social Media with Tattoo	10/41 ≈ 24.4%
Other Social Media No Tattoo	31/41 ≈ 75.6%
Snapchat with Tattoo	16/100 = 16%
Snapchat No Tattoo	84/100 = 84%
Twitter with Tattoo	13/93 ≈ 14.0%
Twitter No Tattoo	80/93 ≈ 86.0%

When plugging this information into Statcato, remember do not put the % symbol. The % symbol can be put in the label at the top. I also abbreviated the category names.

	Percentage (%)
FB w/Tat	32.7
FB No Tat	67.3
Inst w/Tat	27.5

Inst no Tat	72.5
Oth w/Tat	24.4
Oth no Tat	75.6
Snap w/ Tat	16
Snap no Tat	84
Twit w/ Tat	14
Twit no Tat	86



Multiple pie charts are also a nice way to compare conditional percentages. To make the pie chart for the tattoo information for each social media group, you will need to list the frequencies.

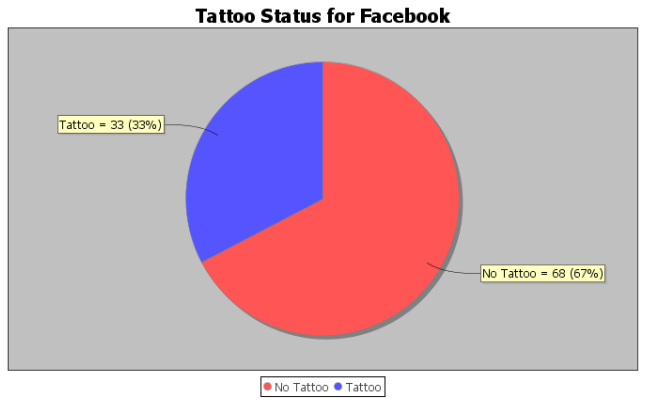
	Facebook Frequencies
Tattoo	33
No Tattoo	68

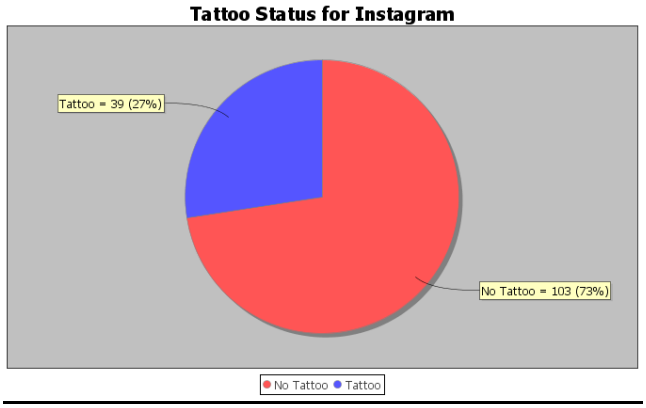
	Instagram Frequencies
Tattoo	39
No Tattoo	103

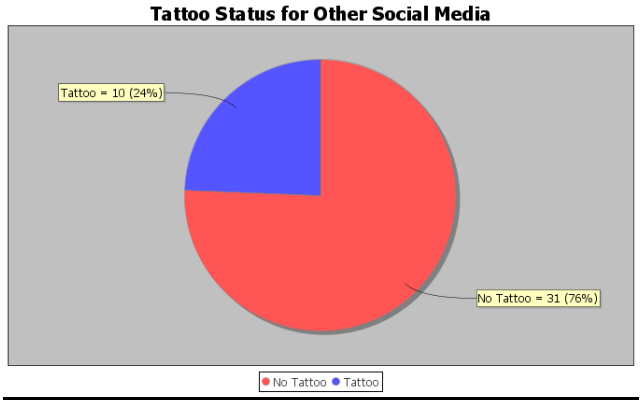
	Other Frequencies
Tattoo	10
No Tattoo	31

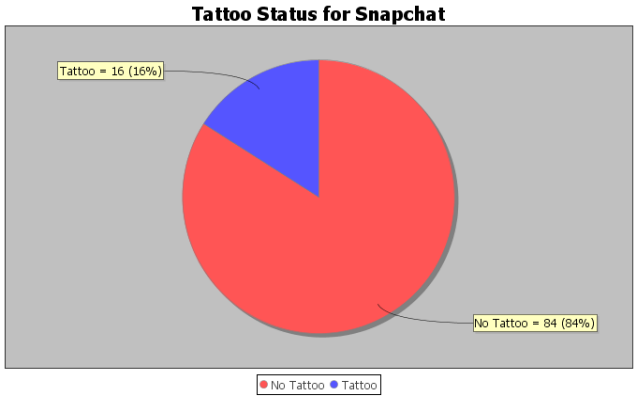
	Snapchat Frequencies
Tattoo	16
No Tattoo	84

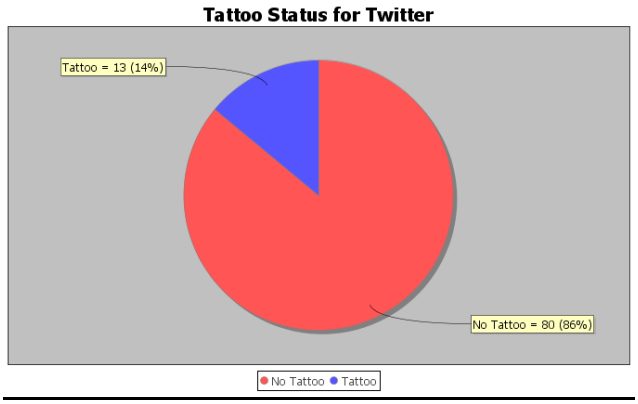
	Twitter Frequencies
Tattoo	13
No Tattoo	80











Problem Set Section 2D

The following two-way tables were created from the Math 075 Survey Data Fall 2015 and describes the student's favorite social media and whether or not they have a tattoo. Use the table to find the given conditional proportions and percentages. Write your answers as a fraction, decimal and as a percentage. Then determine whether you think having a tattoo is related to social media or not.

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}} \qquad \text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

	Facebook	Instagram	Other	Snapchat	Twitter	Total
No Tattoo	68	103	31	84	80	366
Yes Has a Tattoo	33	39	10	16	13	111
Total	101	142	41	100	93	Grand Total = 477

1. a) If we only look at the students with a tattoo, what percent of them prefer Twitter?

   b) If we only look at the students without a tattoo, what percent of them prefer Twitter?

   c) Look at your answers to parts (a) and (b). Do they appear close or significantly different?

   d) Do you think liking Twitter and having a tattoo are related? Explain your answer.

2. a) If we only look at the students that prefer Facebook, what proportion of them have a tattoo?

   b) If we only look at the students that prefer Snapchat, what proportion of them have a tattoo?

   c) Look at your answers to parts (a) and (b). Do they appear close or significantly different?

   d) Do you think social media and having a tattoo are related? Explain your answer.

The following two-way tables were created from the Math 075 Survey Data Fall 2015 and describes the student's gender and how they get to school. Use the table to find the given conditional proportions and percentages. Write your answers as a fraction, decimal and as a percentage. Then determine whether you think gender and transportation are related or not.

	Bicycle	Carpool	Drive alone	Dropped off by someone	Other	Public transportation	Skate	Walk	Total
Female	0	25	199	30	0	13	0	2	269
Male	2	14	159	18	4	11	1	2	211
Total	2	39	358	48	4	24	1	4	480

3. a) What proportion of the females carpool to school?

   b) What proportion of the males carpool to school?

   c) Look at your answers to parts (a) and (b). Do they appear close or significantly different?

   d) Do you think carpooling and gender are related or not? Explain your answer.

4. a) What percent of the females drive alone to school?

   b) What proportion of the males drive alone to school?

   c) Look at your answers to parts (a) and (b). Do they appear close or significantly different?

   d) Do you think carpooling and gender are related or not? Explain your answer.

5. a) What percent of the females are dropped off by someone?

   b) What proportion of the males are dropped off by someone?

   c) Look at your answers to parts (a) and (b). Do they appear close or significantly different?

   d) Do you think carpooling and gender are related or not? Explain your answer.

6. a) If we only look at the people that use public transportation, what proportion of them are female?

   b) If we only look at the people that are dropped off, what proportion of them are female?

   c) Look at your answers to parts (a) and (b). Do they appear close or significantly different?

   d) Do you think type of transportation and gender are related or not? Explain your answer.

7. Create a side-by-side (split) bar chart to summarize conditional percentages for the following two-way tables. For part (a), let the percentages be car accident or not and the conditions be the type of transportation.

a) Look at the car accident and transportation two-way table. Create a side-by-side (split) bar chart comparing the percentages for car accident or not for each type of transportation. Also, create multiple pie charts giving the frequency and percentage of car accidents and no accident for each transportation group.

	Bicycle	Carpool	Drive alone	Dropped off by someone	Other	Public transportation	Skate	Walk	All
No Car Accident	2	31	276	45	2	24	1	4	385
Car Accident	0	8	81	2	2	0	0	0	93
All	2	39	357	47	4	24	1	4	478

b) Look at the gender and transportation two-way table. Create a side-by-side (split) bar chart comparing the percentages for male and female for each type of transportation. Also, create multiple pie charts giving the frequency and percentage of female and male for each transportation group.

	Bicycle	Carpool	Drive alone	Dropped off by someone	Other	Public transportation	Skate	Walk	Total
Female	0	25	199	30	0	13	0	2	269
Male	2	14	159	18	4	11	1	2	211
Total	2	39	358	48	4	24	1	4	480

8. Let's look at the topic of smoking. What variables might be related to a Math 075 student smoking? Which variables have the strongest or weakest relationship? We will use the Math 075 Survey Data to explore Campus Location, Gender, Tattoo, Political Party, and Living with Parents and see if we can identify which variables have the strongest relationship to smoking and which variables have the weakest relationship.

a) Create a two-way table for smoking and campus. What percent of the Canyon Country campus students smoke? What percent of the Valencia campus students smoke?

b) Create a two-way table for smoking and gender. What percent of the female students smoke? What percent of the male students smoke?

c) Create a two-way table for smoking and tattoo. What percent of students with a tattoo smoke? What percent of students without a tattoo smoke?

d) Create a two-way table for smoking and political party. What percent of the republican students smoke? What percent of the democratic students smoke?

e) Create a two-way table for smoking and living with parents. What percent of the students that live with their parents smoke? What percent of students that do not live their parents smoke?

f) Examine the difference between the conditional percentages for your answers in a-f and rank the variables in order from the most significant difference to the least significant difference. You may want to look at the percentage ratio to help.

g) Remember the more different the conditional percentages in each variable the more the variable is related to smoking. Therefore, the ranking in "g" is also the order of strongest relationship to weakest relationship. Which variables (Campus Location, Gender, Tattoo, Political Party, and Living with Parents) had a stronger relationship with smoking?

h) Which variables (Campus Location, Gender, Tattoo, Political Party, and Living with Parents) had a weak or no relationship with smoking?

Chapter 2 Review

Here is a list of important ideas in this chapter.

- Be comfortable creating and analyzing two-way tables with technology from two categorical data sets
- Be able to create and analyze bar charts and pie charts to summarize two way table information
- Be able to find basic marginal proportions, joint proportions (AND / OR), and conditional proportions and be able to convert the proportions into percentages.
- Be able to look at relationships between categorical variables by looking at conditional proportions.
- Relationship Principle
  Values Significantly different => related
  Values Close => not related

Problem Set Chapter 2 Review

1. The following categorical data gives the gender (male or female) of people's pets and who takes care of the pet (caretaker). Create a two-way table from this data. Give the counts and the totals.

Pet Gender	Caretaker
F	Everyone
M	Everyone
F	Parents
F	Parents
M	Everyone
M	Parents
M	Everyone
M	Parents
M	Kids
M	Parents
M	Parents
M	Everyone
F	Everyone

	Kids	Parents	Everyone	Totals
Female Pet				
Male Pet				
Totals				Grand Total =

A total of 280 high school students were asked about their political affiliation. The following two-way table was created from the data. Use the table to answer the following question.

	Democrat	Republican	Other	Total
Freshmen	7	7	28	42
Sophomore	28	21	56	105
Junior	35	28	21	84
Senior	21	14	14	49
Total	91	70	119	280

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}} \qquad \text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

2. What proportion of the students identified with the "Other" political party? (Give your answer as a fraction, decimal proportion and as a percent.)

3. What percent of the students were in their senior year? (Give your answer as a fraction, decimal proportion and as a percent.)

4. What proportion of the students were both democrat and in their junior year? (Both must be true about person) (Give your answer as a fraction, decimal proportion and as a percent.)

5. What percent of the students were both republican and in their sophomore year? (Both must be true about person) (Give your answer as a fraction, decimal proportion and as a percent.)

6. What proportion of the students were either in their freshman year or in their senior year? (Either one can be true about person) (Give your answer as a fraction, decimal proportion and as a percent.)

7. What percent of the students were either democrat or in their senior year? (Either one can be true about person) (Give your answer as a fraction, decimal proportion and as a percent.)

A total of 280 High School Students were asked about their political affiliation. The following two-way table was created from the data. Use the table to answer the following question.

	Democrat	Republican	Other	Total
Freshmen	7	7	28	42
Sophomore	28	21	56	105
Junior	35	28	21	84
Senior	21	14	14	49
Total	91	70	119	280

$$\text{Proportion} = \frac{\text{Amount}}{\text{Total}} \qquad \text{Percentage} = \frac{\text{Amount}}{\text{Total}} \times 100\%$$

8. If we only look at the sophomores, what percent of them are democrat? (Give your answer as a fraction, decimal proportion and as a percent.)

9. If we only look at the seniors, what percent of them are democrat? (Give your answer as a fraction, decimal proportion and as a percent.)

10. Where the percentages in #8 and #9 close or significantly different?

11. Does the data suggest that grade level is related to being a democrat, or not related?

Project Chapter 2 - Two-Way Tables & Categorical Relationships

The class will be broken up into groups of three or four. Each group will pick a team name and two categorical variables from the Math 075 Survey Data Fall 2015 to study. Each group should have a different pair of variables to study.

Group#	Team Name	Categorical Variable A	Categorical Variable B
1		Political Party	Hair Color
2		Smoking	Political Party
3		Texting/Driving	Car Accidents
4		Smoking	Transportation
5		Gender	Political Party
6		Breakfast	Fixed Intelligence
7		Hair Color	Gender
8		Fixed Intelligence	Political Party
9		Tattoo	Gender
10		Political Party	Tattoo
11		Tattoo	Hair Color
12		Smoking	Tattoo

Poster Directions

- Use a statistics software to make a two-way table for the two chosen variables. Put the two-way table with totals on the front of your poster.
- Find two basic marginal percentage questions that involve only one variable. Remember to use the grand total. Put the questions on the front of your poster and the answers on the back of the poster.
- Find two "AND" joint percentage questions that involve two variables. Remember to use the grand total. Put the questions on the front of your poster and the answers on the back of the poster.
- Find two "OR" joint percentage questions that involve two variables. Remember to use the grand total. Put the questions on the front of your poster and the answers on the back of the poster.
- Find two conditional percentage questions that have the same variable for the percentage, but have a different condition. Remember to use the total for your condition. Put the questions on the front of your poster and the answers on the back of the poster.
- Put the following question on the front of your poster and the answer on the back of your poster: Are the two conditional percentages close or significantly different?

- Put the following question on the front of your poster and the answer on the back of your poster: Are the two variables involved in the conditional percentages related or not?
- Decorate your poster and make sure everyone in your group can present the poster.

Presentation Directions

Each group will put their poster up around the room. Chose one person from the group to present first. Everyone else in the class who is not presenting will find a poster that is not their own. Presenter asks the audience the questions and see if they can get the correct answers. Discuss the solutions and what you liked about the poster. Rotate presenters so that everyone has a chance to present. Each presentation should take about 5 minutes. Make sure audience rotates to new posters as well.